INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

RESOLUCIÓN AUTOMÁTICA DE LA ANÁFORA INDIRECTA EN EL ESPAÑOL

Tesis que presenta

Raúl Morales Carrasco

para obtener el grado de

Doctor en Ciencias de la Computación

Director:
Dr. Alexander Gelbukh

Codirector:
Dr. Ruslan Mitkov

México
2003

CONTENIDO

General

Detallado

Figuras

Tablas

RESUMEN

Esta tesis describe la investigación para desarrollar un método de resolución de la anáfora indirecta en el Español. Este método utiliza un diccionario de escenario ampliado con el contexto lingüístico para poder determinar la presencia de la anáfora indirecta.

Al iniciar la investigación se estableció la necesidad de descubrir los marcadores que identificaban la presencia de la anáfora indirecta en los textos escritos; conforme avanzaba el análisis fue necesario profundizar en la relación existente entre la sintaxis, la semántica y la pragmática cuando participan activamente en el discurso textual; con el fin de integrar el contexto lingüístico como parte del discurso comunicativo se requirió la identificación del rol de las expresiones referenciales, marcadas por los determinantes en las frases nominales; finalmente el descubrimiento de la fuerte interrelación entre la referencia, correferencia y la anáfora permitió desarrollar un modelo de escenario ampliado que identifica más claramente la presencia de la anáfora indirecta.

Se presentan los fundamentos y resultados que prueban la factibilidad de tres hipótesis: “la anáfora indirecta posee los mismos marcadores que la referencia y la correferencia en el discurso, la diferencia estriba en el tipo de relación, el método de inferencia y la información necesaria para resolverlas”; “la anáfora indirecta se presenta si, y sólo si, no existe correferencia”; y “la referencia se presenta si, y sólo si, no existen correferencia y anáfora indirecta”. La prueba manual de las hipótesis permitió modelar e implantar un sistema que detecta y resuelve la anáfora indirecta con un prototipo y diccionarios de sinónimos y escenarios construidos manualmente. Para trabajar con texto libre se limita a la anáfora indirecta nominal debido a que depende de la información del diccionario de escenarios y actualmente no existe un diccionario de relaciones verbal-nominal.

ABSTRACT

This thesis describes the research to develop a method for indirect anaphora resolution in Spanish. This method uses a scenario enriched with the linguistic context to determine indirect anaphora presence.

At the beginning off the research the necessity to discover the markers, that identified indirect anaphora presence, was outlined; according to the analysis advance it was necessary to deep in syntax, semantic and pragmatic relation when actively participate in textual discourse; in order to integrate the linguistic context as part of the communicative discourse identification of reference expressions, marked by determinants in nominal phrases, was required; finally, the discovery of the strong interrelation between reference, coreference and indirect anaphora let developed an enlarged scenario model that better clearly identified indirect anaphora presence.

Fundaments and results that evidence feasibility of three hypotheses are presented: “indirect anaphora has same markers than reference and coreference in discourse, the difference lie on the type of relation, inference method and information necessary to resolve them”; “indirect anaphora is present if, and only if, there is not coreference”; “reference is present if, and only if, there is not indirect anaphora”. Manual test of these hypotheses let model and implement a system that detects and resolve indirect anaphora with a prototype and dictionaries build explicitly for it. To work with free text indirect anaphora resolution is limited to nominal expressions because it depends of scenario dictionary, and actually there is not a verbal to nominal relations dictionary.

1 INTRODUCCIÓN

1.1 Antecedentes

Desde los primeros estudios del lenguaje, pero más en los últimos años, la resolución de la anáfora ha sido foco de investigación de filósofos, lingüistas, científicos del conocimiento e IA (**I**nteligencia **A**rtificial), de sicolingüistas y de lingüistas computacionales [Mitkov, 98a]. Su importancia radica, entre otras razones, en que la anáfora:

- es uno de los fenómenos más complejos dentro del lenguaje natural [Huang, 2000; Mitkov, 2001]
- es considerada uno de los problemas fundamentales de la lingüística y Chomsky se apoya en ella, para mantener la teoría de que la facultad del habla es innata [Chomsky, 1986]
- se ha demostrado que en ella interactúan factores sintácticos, semánticos y pragmáticos [Hirst, 1981; Huang, 2000]
- es necesaria en un amplio rango de tareas del PLN (**P**rocesamiento del **L**enguaje **N**atural) como interfaces en lenguaje natural, la comprensión del lenguaje, traducción automática, extracción de información y generación automática de resúmenes [Hirst 1981; Carter 1987; Fox 1987; Aone y McKee 1993; Cornish 1996; Fretheim y Gundel 1996; Hahn et al 1996; Kameyama 1997; Mitkov, 2001]

La anáfora indirecta establece un enlace asociativo entre una entidad lingüística (palabra o expresión) con alguna entidad implícita introducida previamente a través del texto en el discurso. En las últimas décadas ha recibido una especial atención dentro de diferentes disciplinas que la han tratado desde varias perspectivas. Así, dentro de la tradición lingüística, se encuentran Erkü y Gundel [1987], Huang [1994] y Matsui [1995]; dentro de la sicolingüística, están Clark y Haviland [1977], y Sanford y Garrod [1981]; dentro de la

Inteligencia Artificial, está Sidner [1983]; y finalmente dentro de la lingüística computacional están Murata y Nagao [1996] y Gelbukh y Sidorov [1999].

La ***anáfora indirecta,*** mencionada por primera vez en el trabajo de Chafe [1976], ***es uno de los casos más difíciles de relación anafórica*** [Erkü y Gundel, 1987; Kempson, 1982; Matsui, 1995; Huang, 1994; Murata y Nagao, 2000]. También conocida como: conexión referencial [Clark, 1977], anáfora asociativa [Hawkins, 1978], anáfora inferenciable [Prince, 1981], anáfora implícita u oculta [Sidorov y Gelbukh, 1999], conexión de referencia cruzada [Huang, 2000], ***ha sido un caso poco abordado*** en la lingüística computacional **a pesar de *su importancia para determinar la coherencia del texto*** [Mitkov, 2001]. Aumentar el conocimiento sobre ella, las condiciones que la determinan, sus mecanismos de procesamiento y dotar de ellos a la computadora para apoyar el PLN son las metas inmediatas a lograr en este trabajo.

1.2 Situación actual

Se podría sintetizar rápidamente con el comentario de que la mayoría de los problemas de la lingüística computacional se presentan en la resolución de la anáfora indirecta, debido a la necesidad de hacer explícita a la computadora toda la información y relaciones requeridas para "*entender*" el texto; con el fin de explicar lo anterior, se hará un rápido recorrido de las necesidades y los problemas encontrados en el PLN que de una u otra forma afectan la resolución de la anáfora y al final se abordará la anáfora indirecta en particular.

1.2.1 La anáfora

La mayoría del trabajo anterior en la resolución de la anáfora ha utilizado mucho conocimiento del dominio y lingüístico, además de requerir considerable captura manual, [Sidner, 1979; Carter, 1987; Carbonell y Brown, 1988; Rich y Luperfoy, 1988] lo que ha dificultado la representación y procesamiento. Sin embargo, la necesidad apremiante para desarrollar sistemas robustos y menos costosos ha llevado a los investigadores, a partir de 1990, a alejarse un poco del conocimiento lingüístico e intentar estrategias de solución que requieran menor conocimiento (Knowledge-poor) [Dagan e Itai, 1990; Kennedy y Bougarev,

1996; Baldwin, 1997; Mitkov, 1998a, 2000b y 2000c]. Se han utilizado además estrategias combinadas para la resolución de la anáfora en el Español [Palomar et al, 2001].

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.

El preprocesamiento del texto, o procesamiento previo a la aplicación del método de resolución de la anáfora a un texto, es un problema significativo ya que la exactitud es demasiada baja y como consecuencia el rendimiento de estos sistemas está lejos del ideal; la dependencia vital del sistema de resolución de la anáfora es tal que tendrá poco rendimiento, aunque el método sea muy bueno. En esta etapa, *los problemas principales* se encuentran en el análisis morfológico, etiquetado de partes de la oración, reconocimiento de entidades nominales, reconocimiento de pronombres, reconocimiento de palabras desconocidas, extracción de frases nominales, descomposición analítica (parsing), etc. Por ejemplo: la mejor exactitud encontrada para la descomposición analítica de textos sin restricción es alrededor del 87% [Collins, 1997]; el mejor rendimiento logrado con etiquetadores de entidades nominales da una exactitud del 96% cuando se prueba y utiliza en corpus con *noticias* sobre un dominio o tema específico [Mitkov, 2001].

Como resultado de las limitantes mencionadas, *la mayoría de los sistemas de resolución de la anáfora no operan de modo totalmente automático, y algunos métodos han sido simulados sólo manualmente*. Como ejemplos ilustrativos: la resolución propuesta por Hobbs no fue implantada en su versión original [Hobbs, 1976, 1978]; hay trabajos donde se corrigieron manualmente los resultados de las etapas de preprocesamiento (para poder utilizarlas en el algoritmo de resolución anafórica) [Lappin, 1994; Ferrandez et al, 1997;

Mitkov, 1998b]; finalmente hay trabajos donde se utilizaron corpus etiquetados manualmente sin etapa de preprocesamiento [Ge et al, 1998; Tetreault, 1999]. Reaccionando a la situación mostrada se han iniciado esfuerzos, a largo plazo, para lograr sistemas totalmente automatizados [Fukumoto et al, 2000; Tanev y Mitkov, 2000; Mitkov, 2001].

1.2.2 La anáfora indirecta

De los modelos de análisis de la anáfora indirecta, tres son los que más influencia han tenido en el área:

- El modelo de relevancia
- El modelo de tópico o focal
- El modelo de escenario

1.2.2.1 El modelo de relevancia

La teoría de relevancia supone que el mecanismo cognitivo central del ser humano es un dispositivo deductivo generador de inferencias que trabaja tratando de maximizar la relevancia con respecto a la comunicación; el principio de relevancia es el responsable de recuperar el contenido explícito o implícito de un enunciado. En otras palabras, al interpretar un enunciado **el receptor** estará siempre "*maximizando los efectos contextuales del enunciado*" y "*minimizando los esfuerzos de procesamiento del enunciado*" [Matsui, 1995]. Dentro de este modelo, Kempson [1988a] observa que la interpretación de la anáfora indirecta requiere un análisis semántico / pragmático, más que gramatical, y de la información asociada con premisas adicionales implícitas.

En el modelo de relevancia, la idea básica es que la interpretación de la anáfora indirecta se encuentra *suponiendo conexiones* que se apoyan en efectos contextuales apropiados pero *sin sujetar el lenguaje a esfuerzos injustificados* para obtener estos efectos. La validez del análisis de la anáfora con el modelo de relevancia depende crucialmente de cómo se aplica el principio, o más concretamente de cómo se pueden obtener y balancear tanto los "efectos contextuales" como los "esfuerzos de procesamiento". Desgraciadamente, en los trabajos consultados [Matsui, 1993, 1995; Kempson, 1988a, 1988b; Sperber y Wilson, 1995;

Levinson, 1989] no existe un mecanismo claramente satisfactorio para medir el balance costo-beneficio; hasta el momento, no parece que el principio de relevancia haya podido ser implantado confiablemente; y se ha reportado dificultad empírica al probarlo [Huang, 2000].

1.2.2.2 El modelo de tópico o focal

En el modelo de tópico, la idea básica es que la interpretación de la anáfora indirecta está determinada principalmente por el tema o tópico (aquello sobre lo que se está hablando) de las oraciones previas del discurso. Este enfoque está representado por los trabajos de Sidner [1983] y Erkü y Gundel [1987]. La interpretación de la anáfora indirecta se efectúa por un algoritmo que selecciona el foco del discurso con base en un conjunto ordenado de preferencias; además, las interpretaciones resultantes del algoritmo quedan sujetas a los requerimientos de consistencia con el conocimiento del mundo.

1.2.2.3 El modelo de escenario

En el modelo de escenario, la idea básica es que la interpretación de la anáfora indirecta se encuentra siempre referida a un dominio mental apropiado de referencia. Este enfoque está representado por el trabajo de Sanford y Garrod [Sanford y Garrod, 1981; Garrod y Sanford, 1994]. Apoyándose en nociones como: marcos [Minsky, 1975; Fillmore, 1982], esquemas [Rumelhart, 1980; Chafe, 1987] y de guiones [Schank y Abelson, 1977], denominaron a este dominio de referencia ***un escenario***. Un escenario, de acuerdo a Sanford y Garrod, puede ser activado desde tres dimensiones: ***actual***, porque se encuentra en el foco de atención del receptor y almacenado en la memoria primaria, ***o antigua***, si no es parte del foco de atención del receptor y se encuentra en la memoria secundaria; ***explícito***, se refiere a las entidades que han sido mencionadas directamente en el discurso, ***o implícito***, son entidades que no han sido explícitamente mencionadas pero que están relacionadas en forma relevante con algo mencionado en el discurso; ***de entidad***, representada por los individuos que son los principales protagonistas de una escena, ***o de rol***, representada por los papeles representados en los escenarios descritos en el discurso.

1.2.2.4 Comparación de modelos

Para observar las ventajas y desventajas de los tres modelos se puede intentar una comparación manual tomando como base el ejemplo de Erkü y Gundel [1987] y analizándolo desde cada enfoque.

(1) Juan entró a un **restaurante**. El *mesero* era italiano.

En el enfoque focal el *restaurante* es el foco del discurso y por lo tanto el antecedente del *mesero*. En el enfoque de escenario, el uso de *restaurante* invoca un escenario que contiene en forma implícita al menos un *mesero*. Finalmente, dentro del marco de relevancia, la suposición de conexión de que el *restaurante* donde Juan entró tiene al menos un *mesero* proviene de la extensión del contexto por el conocimiento enciclopédico (del sentido común); como consecuencia toda la interpretación es consistente con el principio de relevancia.

Tomando un ejemplo, un poco más complicado [Huang, 2000].

(2) Juan se detuvo por un café en un ***bar capuchino*** antes de comer en un **restaurante**. El *mesero* era italiano.

Este ejemplo contiene más de un antecedente posible (un **restaurante** o un **bar capuchino**) para la anáfora indirecta el *mesero*, donde el antecedente preferido sería un **bar capuchino**; de acuerdo al conocimiento común capuchino = bar italiano y como el mesero era italiano se puede inferir que era mesero del bar capuchino.

Esta interpretación sería correcta desde el modelo focal porque el algoritmo de Sidner tomaría un **bar capuchino** como el tópico o foco del discurso (por el orden de aparición en la primera oración). En contraste, en el análisis de escenario habría dos escenarios actuales activos, uno para un **bar capuchino** y otro para un **restaurante**, cada uno de ellos con posibilidad de tener *mesero*. Ya que no hay mecanismo para escoger entre ambos escenarios queda confuso como se puede derivar una interpretación correcta bajo este enfoque. Finalmente, dentro del marco de relevancia, asumiendo que un **bar capuchino** es más accesible (por ser de tipo italiano) que un **restaurante**, sería la conexión preferida para la interpretación correcta. Para finalizar las comparaciones, se analizan un par de ejemplos similares tomados de Huang [2000].

(3) Juan se detuvo por un café en un **bar capuchino** antes de visitar un **museo de instrumentos musicales**. El *mesero* era italiano.

(4) Juan se detuvo por un café en un **bar capuchino** antes de visitar **un museo de instrumentos musicales**. El *encargado* era italiano.

Intuitivamente, el primer ejemplo parece menos complejo que el segundo y el porqué se encuentra en el conjunto de factores que afectan la interpretación. Clark y Haviland [1977] han identificado: la *distancia* de la conexión (el número de suposiciones necesarias para la conexión), la *plausabilidad* de la conexión (el grado de veracidad de las suposiciones) y la *computabilidad* de la conexión (el grado de facilidad en el calculo de las suposiciones); otros factores pueden incluir la *accesibilidad* (facilidad de acceso) a los *antecedentes* y a las *suposiciones contextuales*, y la *coherencia general* del discurso [Huang, 2000; Matsui, 1995]. Bajo el enfoque focal o del tópico, el factor de *accesibilidad* a los antecedentes parece jugar un rol crucial para explicar porqué el primer ejemplo es menos complejo que el segundo: mientras el antecedente para *el mesero* en el primer ejemplo es el foco del discurso, el antecedente para *el encargado* en el segundo no lo es (puede existir *un encargado* o supervisor tanto en el museo como en el bar). Por otro lado, en el modelo de escenario se tendría que utilizar la noción de *accesibilidad* a las suposiciones contextuales para detectar las diferencias entre ambos ejemplos: el antecedente para *el mesero* en el primero se encuentra en el contexto (escenario) más accesible para el lector que el antecedente para *el encargado* en el segundo, considerando, por supuesto, que exista un mecanismo para decidir en cual de los escenarios actualmente activados es más accesible. Finalmente, en el análisis de relevancia la complejidad mayor del segundo ejemplo se puede atribuir al injustificado esfuerzo de proceso que debe realizar el lector para interpretar la anáfora indirecta *el encargado* en el discurso (como resultado de la mayor *accesibilidad* focal para *un bar capuchino* y el antecedente que se quiere, *un museo de instrumentos musicales*).

Se puede apreciar, que la validez del análisis de la anáfora indirecta, y de la anáfora en general, con el modelo de relevancia depende crucialmente en la aplicación del principio, o más concretamente de cómo pueden obtenerse y balancear tanto los “efectos contextuales” como los “esfuerzos de procesamiento”. Desgraciadamente, en todos los trabajos consultados [Matsui, 1993, 1995; Kempson, 1988a, 1988b; Sperber y Wilson, 1995; Levinson, 1989] no existe un mecanismo satisfactorio para medir el balance costo-beneficio. No parece, que el

principio de relevancia pueda ser implantado confiablemente y se ha reportado la dificultad empírica al probarlo [Huang, 2000].

De la comparación anterior, pueden apreciarse las razones por las cuales los trabajos encontrados para la resolución de la anáfora indirecta hayan optado por utilizar los dos primeros modelos, de escenario y tópico o focal.

1.2.3 Problemas pendientes de resolver

Aunque en los últimos diez años ha habido un considerable avance en el campo de resolución de la anáfora, existen aún considerables problemas sin resolver o que requieren ser atendidos para apoyar su resolución, y que representan los mayores retos para el desarrollo futuro.

Para empezar, no se identifica claramente un solo conjunto de factores (léxico, sintáctico, semántico y pragmático) en la resolución de la anáfora y si este conjunto de factores una vez agrupados estaría completo. En general los factores son divididos en *restricciones y preferencias* [Carbonell y Brown, 1988] pero otros autores arguyen que deberían considerarse como *escala de preferencias* mas o menos restrictiva llamándolas simplemente *factores* [Preuß et al, 1994], *síntomas* [Mitkov, 1995] o *indicadores* [Mitkov, 1998a].

Una vez definidos conviene ver: el impacto individual de cada factor y su secuencia o coordinación al actuar [Carter, 1990]; esclarecer la existencia o no de *dependencia* (o dependencia mutua) de los factores; verificar si son aplicables por igual a todas las lenguas o son específicos de cada lengua. Algunos autores apoyan la idea de que los factores tienen aplicabilidad general a todas las lenguas, pero que las lenguas difieren en la importancia relativa de los factores [Mitkov, 1997]; además se observa, que la diferencia se da por la evolución de las lenguas por lo que podemos hablar de lenguas donde predomina más la sintaxis que la pragmática y viceversa.

"Desde el punto de vista diacrónico, las lenguas parecen cambiar de ser más pragmáticas a más sintácticas; desde una perspectiva sincrónica, las diferentes lenguas están simplemente en diferentes etapas de este círculo evolutivo" [Huang, 2000].

La resolución de la anáfora indirecta requiere conocimiento implícito, previo o de "sentido común", aunado a un análisis pragmático; *lo que la gramática provee es meramente un conjunto de restricciones que el valor identificado de una expresión anafórica debe satisfacer* [Kempson, 1988a, 1988b]. De acuerdo a esto, la interpretación de los diferentes tipos de anáfora depende de los diferentes tipos de información disponibles al lector, por lo que la interpretación para establecer el valor de la anáfora referencial y la acotada por variable (bound-anaphor) se logra con la información que se ha *presentado previamente en el contexto lingüístico*; la interpretación de la deixis anafórica se logra con la información *presente en el contexto del discurso*; y la interpretación de la anáfora indirecta a través de información de *conocimiento implícito en el contexto del discurso*, asociado con premisas adicionales.

1.3 Definición del problema

En el PLN se han desarrollado **sistemas** que dependen de la frecuencia de palabras (las más usadas) en el texto **que sólo establecen referencias explícitas a entidades mencionadas en el texto**, pero ¿Qué pasa con las referencias a las mismas entidades a través de fenómenos como la anáfora? **Limitan la eficacia y eficiencia de los mismos.** Como ejemplos de sistemas desarrollados con este enfoque se tienen sistemas para: encontrar los temas principales en documentos [Guzmán-Arenas, 1999] y sistemas de búsqueda temática de documentos [Alexandrov et al, 2000]. El poder identificar y resolver las referencias anafóricas aumentaría su efectividad; en otras palabras, ***para interpretar por completo el texto es necesario salvar la barrera de las referencias anafóricas, sin esto se mantiene la limitante actual en el PLN***.

Aunque en los últimos años ha habido un considerable avance en el campo de resolución de la anáfora, existen aún discusiones de carácter teórico y práctico que frenan, por así decirlo, el avance en la comprensión del lenguaje [Krahmer y Piwek, 2000]. Para avanzar en la comprensión del lenguaje natural, por medio de la verificación de la coherencia textual **es necesario investigar para obtener una definición más precisa** de la anáfora en general, y **de la anáfora indirecta** en particular, **descubriendo y estableciendo sus características distintivas que permiten elaborar un mejor modelo para implantarlo en la computadora**.

En este trabajo se continuó con la investigación iniciada por Gelbukh y Sidorov [1999] para determinar:

- las condiciones de validez en la formación y los rasgos distintivos (marcadores) que permiten *detectar la existencia posible o no de la anáfora indirecta en un texto*
- cómo debe interpretarse la anáfora indirecta
- cómo debe seleccionarse el antecedente apropiado ante la existencia posible de múltiples anáforas y antecedentes

1.4 Objetivo

El objetivo general es aumentar el conocimiento sobre la anáfora indirecta, las condiciones que la determinan, sus mecanismos de procesamiento y dotar de ellos a la computadora para apoyar el PLN, como contribuciones de este trabajo a la investigación, que se conduce en el mundo a largo plazo, en dos tareas: comprender como aprende el ser humano el lenguaje y construir programas que permitan a la computadora entenderlo; ambas están íntimamente relacionadas.

El objetivo específico de este trabajo ha sido: **desarrollar los modelos, los métodos, los diccionarios y el software que resuelvan la anáfora indirecta en los textos en español.**

1.5 Justificación

La resolución de la anáfora es fundamental para el desarrollo de interfases de lenguaje natural, principalmente en Internet que se está desarrollando aceleradamente, para poder involucrar más gente a la utilización y beneficios de la computación, de las Bibliotecas Digitales, etc., donde se requieren aplicaciones para obtener información textual de imágenes e indexarla, donde la falta de nitidez obliga a utilizar reconocimiento aproximado para apoyar el OCR [Wu et al, 1997; Morales, 1999]; se necesitan nuevos métodos para investigar y recuperar datos de la creciente información textual disponible; y es cada día más apremiante entender el texto a un nivel lingüístico más profundo para cubrir estas demandas [Uchida et al, 1999; Gelbukh, 2000].

Uno de los problemas puntuales del procesamiento de lenguaje natural es la verificación de coherencia textual y es donde la resolución de la anáfora tiene una aportación muy importante porque permitirá la generación automática de resúmenes.

1.6 Limitaciones y delimitaciones

Este trabajo identifica las condiciones en las que se da la anáfora indirecta considerando al resto de las referencias como parte de los algoritmos que se desarrollan; se ha profundizado sólo lo necesario en los demás tipos de referencia para apoyar la consecución del objetivo principal.

Se utilizaron las herramientas disponibles en Internet (por ejemplo en el análisis de co-ocurrencias); además se logró obtener un corpus etiquetado verificado manualmente, Clic-TALP V3.0 de la Universidad Politécnica de Cataluña; el corpus etiquetado facilitó la programación, al depender menos del preprocesamiento, y permitió desarrollar el prototipo inicial; sin embargo, impide que pueda utilizarse con nuevos textos (se requiere que estén etiquetados). Para resolver este problema y poder procesar cualquier texto libre se utilizó el corpus Clic-TALP y entrenó el etiquetador TnT logrando salvar esta limitante. Actualmente el sistema en su versión DEMO procesa cualquier texto libre que contenga menos de 4800 unidades léxicas (o tokens) equivalente a 45 KB aprox. Los archivos en formato diferente (html, Word, ps, etc.) deben convertirse a texto puro en ambiente Windows para poder ser procesados.

Es importante mencionar que en este documento no se pretende desarrollar un tratamiento completo de todas las construcciones y fenómenos involucrados en la resolución de referencias y de la anáfora, sino sólo de aquellos que se juzgan más importantes para establecer claramente los conceptos en que se fundamenta la resolución de la anáfora indirecta.

1.7 Organización del documento

Habiendo presentado el panorama del trabajo en la introducción, el capítulo 2 presenta el marco teórico de referencia, o marco conceptual seleccionado y conformado *desde el punto de vista lingüístico*. Se inicia con la descripción de los niveles del lenguaje y el contexto del discurso; se continúa con la descripción del texto y sus propiedades finalizando con el análisis de las referencias del discurso tomando en cuenta la función de los determinantes así como su relación con la elipsis nominal y la anáfora. Después se describen la anáfora directa e indirecta y finalmente se presentan sus características e interrelación por medio de ejemplos comentados que permiten visualizar el modelo computacional a desarrollar para resolver la anáfora directa.

En el capítulo 3, apoyándose en el marco conceptual presentado en el capítulo 2, se describe el método de resolución computacional propuesto y el análisis del sistema requerido. Se desarrolla exponiendo primero como se propone detectar y resolver la anáfora indirecta apoyándose en las expresiones referenciales para finalizar describiendo los algoritmos desarrollados para lograrlo.

En el capítulo 4 se describe la implantación tomando en cuenta el corpus utilizado, la información adicional en diccionarios y desarrollo del prototipo inicial; posteriormente se presentan los criterios y adecuaciones necesarias, que se hicieron para que el sistema pueda trabajar con texto libre.

El capítulo 5 se comentan: las razones que apoyan el diseño experimental desde la selección del corpus más apropiado y del tamaño de la muestra adecuado; la evaluación experimental del método diseñado con un prototipo inicial; la evaluación experimental con archivos con texto libre considerando las diferentes condiciones.

El capítulo 6 presenta las conclusiones obtenidas de la investigación realizada, las aportaciones logradas y las tareas pendientes a resolver en el futuro. Finalmente se reportan las referencias citadas y los anexos que contienen información adicional que soporta el trabajo realizado y cuya consulta permite ampliar el panorama del mismo.

2 FUNDAMENTOS LINGÜÍSTICOS

En este capítulo se describen las consideraciones lingüísticas que sirven de base al estudio del lenguaje y de la anáfora indirecta. Primero se describen brevemente los niveles de estudio del lenguaje, para ubicar el estudio de la anáfora indirecta entre ellos, y el concepto general del contexto. En segundo lugar se presenta el texto y sus propiedades: la adecuación, la cohesión y la coherencia; estas propiedades deben encontrarse en cualquier texto, y se identifica su interrelación y los mecanismos en que se apoyan.

2.1 Los niveles del lenguaje

El lenguaje para su estudio, análisis y procesamiento ha sido dividido en niveles de acuerdo a la cantidad y estructura (u organización) de los datos. Tomando en cuenta que el origen de la representación textual es la representación del habla, tenemos:

1. Fonética – estudia los sonidos del habla desde el punto de vista de sus características físicas o articulatorias que influyen en la generación de voz con diferente volumen y tonos.
2. Fonología – estudia el conjunto de relaciones de los sonidos en la generación e interpretación del habla. Se considera la rama de la lingüística que estudia los elementos fónicos, atendiendo a su valor distintivo y funcional [DRAE].
3. Morfología – estudia la estructura y relaciones de las agrupaciones de símbolos para la formación de palabras. Se considera la parte de la gramática que se ocupa de la estructura de las palabras [DRAE].
4. Sintaxis – estudia la estructura y relaciones en las agrupaciones de palabras para la formación de frases y oraciones. Se considera la parte de la gramática que enseña, por

medio de un conjunto de reglas, a coordinar y unir las palabras para formar las oraciones y expresar conceptos [DRAE].

5. Semántica – estudia la interpretación del significado de las palabras tomando en cuenta el uso general y sus relaciones sintácticas en forma independiente del contexto del discurso [SIL].

6. Pragmática – estudia la interpretación y el significado de las palabras tomando en cuenta el contexto del discurso en el que se utilizan, incluyendo la intención supuesta del emisor y la del receptor [SIL]. En otras palabras, es el estudio de como las expresiones analizadas gramaticalmente interactúan en relación con el contexto de interpretación.

La interpretación de la correferencia y de la anáfora se apoya en los niveles 4 al 6 (sintáctico, semántico y pragmático). La interpretación de la anáfora indirecta en particular requiere apoyarse principalmente en el nivel 6 (pragmático).

2.2 El contexto del discurso

En general, se entiende por contexto del discurso *el conjunto de conocimientos y creencias compartidos por los interlocutores de un intercambio verbal y que son necesarios para* ***producir e interpretar*** *sus enunciados*.

En el estudio del contexto del discurso se reconocen tres componentes: el sociocultural, el situacional y el lingüístico.

El contexto sociocultural es la configuración de datos que proceden de condicionamientos sociales y culturales sobre el comportamiento verbal y su adecuación a diferentes circunstancias. Hay regulaciones sociales, por ejemplo, sobre cómo saludar o sobre qué tratamiento o registro lingüístico usar en cada tipo de situación.

El contexto situacional, es el conjunto de datos accesibles a los participantes de una conversación, que se encuentran en el entorno físico inmediato. Por ejemplo: para que el enunciado "*cierre la puerta, por favor*" tenga sentido, es necesario que haya ciertos requisitos o *presuposiciones que son parte de la situación* de habla: que haya una puerta en el lugar donde ocurre el diálogo, y que esté abierta.

El contexto lingüístico está formado por *el material lingüístico que precede y sigue a un enunciado*. En las actividades de lectura el contexto lingüístico es de gran importancia para inferir palabras o enunciados que no conocemos. Vale la pena puntualizar aquí una diferencia entre el texto oral y el escrito; en el texto escrito el contexto lingüístico *incluye paulatinamente* la información necesaria para construir el contexto situacional en su proceso de generación o interpretación; esto se debe a que, entre emisor y receptor, no es posible: la interacción de los sentidos con elementos del entorno común; ni la posibilidad de solicitar una corrección o una aclaración en el proceso de comunicación, como ocurre en una conversación oral.

El contexto del discurso, necesario para el proceso de la ***producción e interpretación*** del lenguaje, puede imaginarse como un témpano flotante (iceberg) donde lo único perceptible o superficial considerado como **explícito** es: el conjunto de símbolos agrupados para formar unidades léxicas; estas unidades léxicas agrupadas en estructuras formando frases y oraciones; y el agrupamiento de oraciones para formar párrafos, que dan forma y estructura a un documento de texto; como se muestra en la figura 1.

Figura 1 El contexto

Pero, ¿Qué hay "debajo del agua"? Debajo del nivel del agua (indicado con la flecha) existe la información **implícita** requerida para la interpretación completa, que depende de: el conjunto de reglas gramaticales; los vocabularios del emisor (hablante o escritor) y del receptor (oyente o lector) idealmente iguales; la habilidad del emisor para expresar sus ideas;

la habilidad del receptor para integrar la información del discurso; la información previa o el conocimiento del tema; el conocimiento enciclopédico o del sentido común; y el acto comunicativo o actividad lingüística. *Se percibe sólo el texto* y lo demás permanece oculto; ***el témpano completo constituye el contexto del discurso necesario para el proceso de generación o interpretación del lenguaje natural***.

De acuerdo a lo anterior, el contexto del discurso *está en función de la situación determinada que da origen o produce una actividad lingüística*. Si esta situación es bien conocida, por el emisor y el receptor, se requiere el uso pocas palabras y de muchas si se ignora totalmente; conforme al principio de minimización "*Entre menos se dice, mayor significado tiene (cada vocablo)*" [Levinson, 1987:68]. Hay así una relación inversa entre la comunicación y la situación, y esta relación *tiende a una constante informativa* [Cerdá, 1975] como se ilustra en la figura 2.

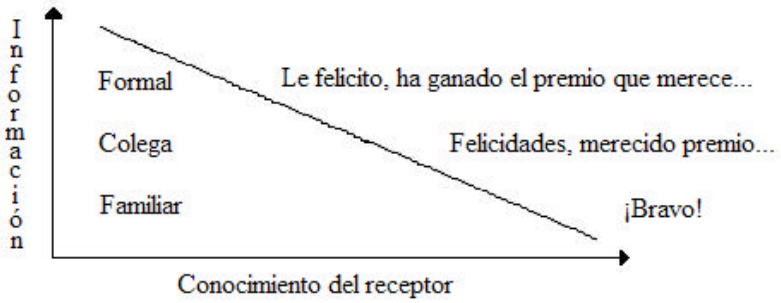


Figura 2 Principio de Minimización

Se puede observar que la situación, de felicitación por la entrega de un premio, depende de cuanta información haya sido intercambiada previamente por el emisor y el receptor (en conversaciones anteriores) y se supone que será mayor si existe una relación de confianza sobre las expectativas, esfuerzos por conseguir el premio y que regresa después de obtenerlo; así la relación empática y más completa con un familiar sólo requiere de un ¡bravo! ya que se conocen previamente los hechos. En cambio si la situación se presenta con el ejecutivo que entrega el premio, es necesario mencionar con mayor amplitud las razones de la felicitación. En algunos manuales de redacción [Schmelkes, 1993] se recomienda tomar en cuenta esta consideración cuando se escribe un documento, ya que es necesario ubicar la audiencia (posibles lectores), para dar la explicación más amplia y a la vez concisa.

2.3 El texto y sus propiedades

Aunque hoy en día no hay un gran consenso para definir el término texto, se considera que *es cualquier conjunto estructurado de enunciados que se producen en un proceso de comunicación.* Los textos pueden ser *orales o escritos*; literarios o no; para leer o escuchar, o para decir o escribir; largos o cortos; etc. Por lo tanto, se consideran textos los escritos de literatura, los diálogos y las conversaciones, las noticias, las pancartas publicitarias, etc.

Los enunciados que forman un texto están en función de lo que se quiere expresar: un enunciado puede contener información que amplíe, explique, corrija o contraste lo dicho anteriormente. Para que una manifestación verbal pueda ser considerada como texto deberá cumplir al menos las propiedades textuales de: adecuación, cohesión y coherencia. La adecuación va íntimamente ligada al contexto sociocultural del discurso; la cohesión a la dependencia gramatical entre las diferentes unidades que lo componen; y la coherencia al tipo de conocimiento que se requiere para interpretar el texto.

2.3.1 La adecuación

La adecuación es la propiedad que *considera el conocimiento y el dominio de la diversidad lingüística*. La lengua no es uniforme ni homogénea, sino que presenta variaciones según diversos factores: la geografía, la historia, el grupo social, la situación de comunicación, la interrelación entre los hablantes, el canal de comunicación, etc.

Dentro de un mismo dialecto, la lengua también ofrece registros muy diferentes: especializados, formales, coloquiales, etc. Por ejemplo *gastralgia, dolor de estómago, dolor de panza* o *dolor de barriga* pueden ser sinónimos en algunos contextos socioculturales, pero tienen valores sociolingüísticos diferentes: *gastralgia* está marcada formalmente y pertenece a un registro más culto y especializado como es la medicina (*gastro* = estómago y *algia* = dolor) que se usa en reportes de hospitales; *dolor de estómago* pertenece a un nivel de formalidad familiar o neutra, se utiliza en general y hasta en la publicidad comercial de radio y televisión; *dolor de panza* o *dolor de barriga* se considera coloquial si lo dice un niño en el ambiente familiar, o vulgar si lo expresa un adulto en público.

Un texto adecuado permite suponer que el autor supo escoger de entre todas las soluciones lingüísticas que da la lengua, la más apropiada para cada situación de comunicación. Para ello, es necesario utilizar el dialecto local o el estándar más general según los casos; y también es necesario dominar cada uno de los registros más habituales de la lengua: los medianamente formales, los coloquiales, los especializados más utilizados por el hablante, etc. Esto implica tener bastante conocimiento, aunque sea subconsciente, sobre la diversidad lingüística de la lengua: saber qué palabras son dialectismos locales, y que por lo tanto no serían entendidas fuera de su ámbito, y cuáles son generales; conocer la terminología específica de cada campo. En resumen, la adecuación *exige al usuario* de la lengua *sensibilidad sociolingüística para seleccionar el lenguaje apropiado en cada comunicación*.

2.3.2 La cohesión

La cohesión *es la propiedad que une el texto tomando en cuenta las articulaciones gramaticales*. Las oraciones que conforman un texto no son unidades aisladas e inconexas, puestas una al lado de otra, sino que están vinculadas o relacionadas con medios gramaticales diversos (puntuación, conjunciones, artículos, pronombres, sinónimos, entonación, etc.), de manera que conforman entre sí una imbricada red de conexiones lingüísticas, la cual hace posible su codificación y decodificación. En pocas palabras, la propiedad de la cohesión engloba cualquier mecanismo de carácter lingüístico o paralingüístico que sirva para relacionar las frases de un texto entre sí; es básicamente gramatical y afecta a la formulación superficial del mismo.

Para dotar de cohesión al texto se conocen diferentes mecanismos (o sistemas de conexión de oraciones) y figuras de construcción que consisten en reglas de concordancia de caso, género, número y persona o en alteraciones del orden "*normal*" de la frase entre las que se incluyen: la referencia, la deixis, la anáfora, las relaciones temporales (tiempos verbales), las relaciones semánticas entre palabras, los mecanismos paralingüísticos, la entonación, la puntuación, el hipérbaton, el pleonasmo, la elipsis y la silepsis.

Cabe mencionar que la entonación y la puntuación se consideran dos sistemas de cohesión paralelos con características y funciones particulares. *La entonación* es uno de los más importantes y expresivos mecanismos de cohesión que *sólo se da en la lengua oral*; en

contraste, *la puntuación es propia de la escritura* con las posibilidades de expresión limitadas a los signos gráficos.

La entonación indica si una oración termina o no, si se ha acabado de hablar, o si se trata de una interrogación, una admiración o una afirmación, etc.; tiene también otras funciones y capacidades expresivas que van mucho más allá de la cohesión: indica la actitud del hablante (seria, irónica, dubitativa, reflexiva, etc.), el énfasis que se pone en determinados elementos del texto: una palabra, una frase, etc. Si bien es cierto que determinadas formas de entonación se marcan en el escrito con signos gráficos apropiados (**? ,! , -**), otros muchos usos de la puntuación (oposiciones, enumeraciones, cambios de orden, etc.) tienen una explicación únicamente sintáctica, sin correlación tonal.

2.3.3 La coherencia

La coherencia es la propiedad que *indica cuál es la información pertinente que se ha de comunicar y cómo se ha de hacer* (en qué orden, con qué grado de precisión o detalle, con qué estructura, etc.). En pocas palabras, la coherencia es la propiedad que *se encarga de la cantidad, la calidad y la estructuración de la información*; es básicamente semántica y afecta a la organización profunda del significado del texto.

Entre el conjunto de medios que existen para conseguir la coherencia textual se tienen:

1. **Las presuposiciones.-** Se trata de la información que el emisor del texto supone que conoce el receptor. Es esencial para que un texto sea coherente, para el receptor, que el emisor haya "acertado" en sus presuposiciones.
2. **Las implicaciones.-** Se trata de las informaciones adicionales contenidas en un enunciado. Un enunciado del tipo "*cierra la puerta*" contiene, al menos, tres implicaciones: hay una puerta, la puerta está abierta y el receptor está en condiciones de cerrarla.
3. **El conocimiento del mundo.-** La coherencia de un texto depende también del conocimiento general que se tenga del mundo. Por ejemplo, un enunciado del tipo "*Los cuervos están de luto*" contradice el conocimiento general "normal" de la realidad porque los cuervos son considerados aves de carroña (comen cadáveres) y no lamentan la muerte de un ser viviente.

4. **El marco referencial** - Se trata del tipo de texto, su finalidad y la situación comunicativa en la que se produce. Dependiendo del marco, un determinado enunciado puede ser coherente, aunque choque con el conocimiento general "normal" del mundo. Por ejemplo, el enunciado considerado anteriormente, "*Los cuervos están de luto*", sería coherente en un texto que trate de películas mexicanas ya que es el título de una de ellas.
5. **Tema y Rema** - El **tema** es el asunto o materia del discurso; es aquello de lo que se habla o escribe y a lo que se deben subordinar todos y cada uno de los enunciados del texto; es lo que el emisor supone "*conocido*" por el receptor y sirve de base para recibir lo "*desconocido*" o nueva información que se denomina **rema** o comentario. El equilibrio entre lo que ya se sabe y lo desconocido asegura la comprensión y el interés de la comunicación y sólo cuando esta correlación tema-rema se ajusta adecuadamente la comunicación tiene éxito. Además, el tema y el rema van cambiando a medida que el receptor decodifica el texto, porque lo que es desconocido ($rema_1$) pasa a ser conocido (o parte del tema) y sirve de puente para recibir los nuevos datos ($rema_2$, $rema_3$, ... $rema_n$). Este proceso se conoce como tematización y es la base de la generación o interpretación progresiva de la información en el texto.

Pero ¿Qué es la coherencia? En los diccionarios se le define como "*la conexión o unión de una cosa con otra*" [DRAE, 1995] y como "*conexión o enlace lógico de una cosa con otra*" [ESPASA, 2001]. Aplicada al discurso se puede definir como "*la conexión, continuidad o coordinación, que se observa entre los componentes del discurso*". Se han identificado cinco tipos de coherencia, que no son independientes entre sí y se encuentran íntimamente interrelacionadas: **temporal**, la que identifica la continuidad de cuando los hechos están ocurriendo; **localidad**, que identifica el lugar donde ocurren los eventos; **causal**, que consiste en el porqué (las razones) los hechos ocurren; **estructural**, que tiene que ver con la forma en que se describen los hechos en el discurso [Gernsbacher, 1997]; y **referencial**, la que identifica a quién o qué se está discutiendo.

La coherencia referencial, íntimamente ligada a la anáfora indirecta, se observa como un proceso incremental de procesamiento de información transmitida del emisor (escritor o hablante) al receptor (lector u oyente) tanto en lenguaje escrito como en el oral [Gernsbacher,

1997]. Este proceso tiene que ver con señales que el emisor coloca explícitamente en el texto o discurso; se requiere pues, el reconocimiento de estas señales por el receptor para interpretar la coherencia de la nueva información con la previamente recibida. En el texto se pueden identificar señales que en forma explícita hacen referencia a entidades mencionadas previamente en el discurso; por ejemplo en la anáfora los pronombres él, ella, etc.

(5) **María** terminó su noviazgo con **Juan**. *Él* se molestó mucho con *ella*.

Otras señales no están explícitamente en el texto y llegar a identificar las entidades a que hacen referencia requiere un conocimiento del proceso; por ejemplo el fenómeno de elipsis (u omisión) del pronombre, "implícito" en la conjugación del verbo del siguiente ejemplo.

(6) Ayer [*tu*] jugaste un buen partido de fútbol.

Algunas señales de coherencia están aún más ocultas e identificar las entidades a que hacen referencia requiere un proceso de inferencia. Para interpretar estas señales el receptor debe apoyarse en su conocimiento previo adquirido del mundo real (de los eventos, hechos y relaciones). Ejemplo:

(7) Juan N. fue **asaltado** ayer. *El ladrón* continúa prófugo.

En la nota periodística anterior se observa que la coherencia requiere conocimiento previo de que en un asalto (evento) participan la víctima del asalto (Juan) y el ladrón o asaltante.

La estrategia de solución intentada para resolver la coherencia textual, es resolver cada tipo de coherencia uno por uno, por lo que este trabajo se enfoca sólo a la ***coherencia referencial*** considerando que la anáfora indirecta es un tipo de referencia que relaciona entidades en el discurso.

3 TRABAJO RELACIONADO

Se encontraron pocos trabajos realizados, desde el punto de vista de lingüística computacional, sobre la anáfora indirecta; de los encontrados, se consideran cuatro representativos: dos dedicados al Japonés [Murata, 1996, 2000] y dos al Inglés [Gelbukh y Sidorov, 1999; Muñoz et al, 2000], ninguno al Español. Una es la tesis "Resolución de la anáfora en oraciones del japonés usando expresiones superficiales y ejemplos" [Murata, 1996] sobre la resolución de la anáfora en general con el capítulo 4 dedicado a la anáfora indirecta en particular y un artículo [Gelbukh y Sidorov, 1999] donde se propone un método de resolución de la anáfora indirecta.

En la tesis de Murata, se propone un método, basado en el modelo del tópico o focal, para resolver la anáfora indirecta en el Japonés utilizando las relaciones existentes entre dos verbos, almacenadas en un diccionario de marcos basado en casos típicos. Primero toma todos los posibles antecedentes del tópico o foco de las oraciones precedentes; en segundo lugar, pondera dichos antecedentes de acuerdo a su plausibilidad; y por ultimo, determina el antecedente requerido combinando la ponderación de los antecedentes, el peso de la similaridad semántica de cada relación almacenada en el diccionario y el peso relativo de la distancia entre la anáfora y su posible antecedente. Obtuvo una precisión de 68% y una recuperación de 63% en las oraciones de prueba comprobando que el uso de las relaciones es útil.

En el artículo, de Gelbukh y Sidorov [1999], el método detecta la anáfora indirecta expresada con los marcadores más frecuentes, en el Inglés e identificados por ellos, un artículo definido o un demostrativo y aplican el modelo de escenario basado en diccionario.

Se utiliza para descubrir relaciones anafóricas entre palabras en diferentes oraciones "*entre una palabra y una entidad implícitamente introducida en el texto previo*"; dicha entidad no tiene una representación superficial en el texto sino en el escenario prototípico de la palabra antecedente. Utilizan un diccionario donde cada "entrada" de palabra está relacionada con las

palabras que pueden participar potencialmente con la situación expresada por la “entrada”. Establecen, en el ámbito sintáctico, dos condiciones que hacen posible la presencia de la anáfora indirecta como condiciones necesarias (pero no suficientes); una vez detectada la anáfora potencial se buscan los posibles candidatos para antecedentes con base en la distancia lineal y estructural; se determina el grado de satisfacción por conteo hasta lograr un nivel de satisfacción preestablecido. Si se logra, significa que existe la relación anafórica indirecta de otra forma se supone inexistente.

El algoritmo de Gelbukh y Sidorov [1999a y 1999b] no prueba las palabras dentro de la misma frase simple y trabaja de la siguiente manera. Para la resolución de la anáfora indirecta toma en cuenta los dos problemas fundamentales: descubrir la presencia de la anáfora indirecta y resolver la ambigüedad de la relación anafórica.

El acercamiento al problema se hace en el orden opuesto; se intenta resolver la relación anafórica, y si se tiene éxito, se considera que el elemento de esta relación se encuentra en el discurso. El algoritmo para detectar la anáfora trabaja como sigue:

- Se considera cada palabra del texto.
- Si la palabra está precedida por un artículo determinado o un pronombre demostrativo es una anáfora potencial y el algoritmo intenta encontrar un antecedente plausible para él, buscando los posibles antecedentes candidatos con base en la distancia lineal y estructural de la anáfora potencial.
- Para cada antecedente potencial, se prueban las condiciones de referencia implícita y grado de compatibilidad. El grado de satisfacción, en lugar de una respuesta binaria de sí o no, se determina como una probabilidad; así, se combinan (multiplican) las probabilidades para las condiciones y la distancia, y se utiliza un valor límite para decidir cual pareja de palabras pasa la prueba, lo que significa que se encontró una relación anafórica o no dependiendo del resultado.
- El algoritmo se detiene cuando encuentra el fin de archivo (no hay más palabras por revisar).

4 RESOLUCIÓN CON ESCENARIO AMPLIADO

El método desarrollado utiliza el modelo de escenario ampliado con el contexto lingüístico para poder determinar la presencia de la anáfora indirecta. La resolución completa de la anáfora indirecta se puede dividir en dos problemas fundamentales a solucionar:

- detectar o descubrir la presencia de la anáfora indirecta.
- resolver la ambigüedad de la relación anafórica ante la presencia de varios antecedentes posibles.

El primer problema, la detección, requiere apoyarse en algún tipo de marcador del lenguaje que permita identificar la presencia de la anáfora indirecta. En esta investigación se observó que los marcadores de la anáfora indirecta son los determinantes en la frase nominal (la convierten en una expresión referencial) sin embargo, estos marcadores también pueden servir para marcar una correferencia o una nueva referencia. ¿Cómo discriminarlos? La estrategia a seguir fue tomar en cuenta *la secuencia u orden de resolución:* correferencia directa, correferencia indirecta, anáfora indirecta, referencia directa y referencia indirecta. Se buscó solucionar el primer problema en la resolución de la anáfora indirecta "*detectar la presencia de la anáfora indirecta*", al verificar que no existe correferencia directa, por comparación de cadenas, ni correferencia indirecta de expresiones nominales, por medio del conocimiento disponible en un diccionario de sinónimos.

Para el segundo problema "*resolver la ambigüedad de la relación anafórica*" ha sido necesario: obtener un diccionario del conocimiento común en donde cada entrada registre el conjunto de relaciones útiles para resolver la anáfora indirecta; después determinar la distancia "normal" (en número de oraciones) o rango de búsqueda posible de antecedentes para la anáfora indirecta detectada. Finalmente utilizando el principio de relevancia "el menor esfuerzo de procesamiento" asignar el antecedente "más cercano" que permita resolverla. El

método utilizado hace uso del principio de relevancia en el diseño de los algoritmos y del modelo de escenario para seleccionar la información útil en la resolución.

En este capítulo se describe el desarrollo del modelo de resolución de la anáfora indirecta en dos partes: primero el modelo lingüístico que permite comprender con suficiente detalle los elementos involucrados en el fenómeno y después el modelo computacional que analizado y diseñado para su implementación; en ambos casos se presentan ejemplos ilustrativos.

4.1 Modelo lingüístico

En el modelo lingüístico se relacionan las teorías para explicar el funcionamiento del fenómeno dentro del procesamiento del texto. Primero se describen las referencias iniciando con los determinantes que funcionan como marcadores comunes de las diferentes referencias presentes en el discurso; en segundo lugar, la elipsis nominal que afecta la función de los determinantes al hacerlos funcionar como pronombres extrínsecos; en tercer lugar, los conceptos de expresión referencial y entidad como soporte para explicar el de referencia y sus variantes: correferencia directa e indirecta y el de anáfora indirecta. Después se analiza el proceso de resolución de referencias, desde el punto de vista del receptor, tomando en cuenta los conceptos anteriores y el de entidad como soporte. Es hasta este momento en que se explica el concepto de anáfora directa e indirecta resumiendo en el proceso las diferencias y similitudes de las referencias y las relaciones necesarias para modelar el proceso de resolución. Finalmente se describe, aprovechando ejemplos específicos, la interrelación de los fenómenos de referencia, correferencia y anáfora para poder identificar claramente la presencia de la anáfora indirecta y los requerimientos para lograr su resolución.

4.2 Las referencias en el discurso

La anáfora indirecta, como uno de los elementos que apoya la resolución de la coherencia referencial, hace necesario profundizar en el tema de las referencias en el discurso, entender sus componentes y relaciones. En el discurso se hace referencia a entidades (normalmente referenciados por nombres) y a acciones realizadas por estas entidades (verbos)

cuyas relaciones se establecen por medio de expresiones referenciales. Por estas razones es necesario explicar las funciones de los determinantes, y como estas funciones son afectadas por el fenómeno de elipsis, en la formación de las expresiones referenciales. Después, se procede a explicar el proceso de resolución de referencias antes de abordar la anáfora directa e indirecta; finalmente se muestra su interrelación y requerimientos de resolución por medio de ejemplos específicos comentados.

4.2.1 La función de los determinantes

Determinante, también conocido como determinativo actualizador, es un término que denomina a la unidad léxica que precede un nombre en una frase nominal para especificar su referencia, incluyendo la cantidad del nombre. Por ejemplo:

(8) *Todos esos* carros están en venta.

Todos determina la cantidad y *esos* es un demostrativo que nos indica el lugar relativo al emisor donde se encuentran los carros.

En general, los determinantes dan origen a las expresiones o descripciones definidas (definite descriptions) donde se utilizan principalmente los artículos determinados (el, la, lo, las, los), los demostrativos (aquel, aquella, aquellas, aquellos, esa, esas, ese, esos, esta, estas, este, estos, tal, tales, semejante, semejantes), los posesivos (cuya, cuyas, cuyo, cuyos, mi, mis, nuestra, nuestras, nuestro, nuestros, su, sus, vuestra, vuestras, vuestro, vuestros, tu, tus, etc.) y los cuantificadores (todo, algún, cada). Con cualquier otro determinante se consideran expresiones indefinidas.

Los diversos determinantes de un texto establecen varios tipos de relaciones entre las expresiones: desconocido ➔ conocido, emisor ➔ receptor, cercano ➔ lejano, etc. En el ejemplo:

(9) *Un* caballero llegó al parque y encontró *un* zorro y *un* conejo. *Ese* conejo dijo *al* caballero que *aquel* zorro era amigo suyo...

Se oponen: *un* caballero ➔ *al* caballero (desconocido ➔ conocido), *Ese* conejo ➔ *aquel* zorro (cercano ➔ lejano).

En el estudio de la anáfora indirecta se ha intentado tener una taxonomía de descripciones definidas relacionadas con la coherencia textual. La razón principal es la identificación de marcadores que identifiquen la relación de correferencia directa o anáfora indirecta.

> "*Una de las principales distinciones que puede hacerse en el uso de descripciones definidas es entre la anáfora directa e indirecta, donde* TdirectaU *puede significar que la interpretación de la frase nominal sólo involucra la recuperación del referente del discurso introducido por una frase nominal correferente en el contexto lingüístico, y* TindirectaU *implicaría que la interpretación involucra algún tipo de procesamiento adicional*" [Fraurud, 96]

Se ha observado que las condiciones sugeridas, uso del artículo definido y de demostrativos como marcadores de la presencia de anáfora indirecta [Sidorov y Gelbukh, 1999], sólo se cumplen en algunos casos habiendo encontrado otras alternativas, que también marcan la presencia plausible de la anáfora indirecta, por ejemplo:

(10) Juan$_1$ compró *una* camioneta$_2$ usada. [Él$_1$] Tuvo que cambiar *dos* neumáticos$_3$, *algún* fanal$_4$ y *una* bocina$_5$ para usarla$_2$ como a él$_1$ le$_1$ gusta.

En el ejemplo (10) se observa que "*una* bocina$_5$" (artículo indefinido), "*dos* neumáticos$_3$" (cardinal) y "*algún* fanal$_4$" (determinante indefinido) marcan con la misma relación *parte_de* la anáfora indirecta con la camioneta$_2$ previamente mencionada.

En otros casos el artículo o el demostrativo no marca a la presencia de la anáfora indirecta sino una correferencia o una referencia a una nueva entidad del discurso. En el ejemplo (10) se observa que "*una* camioneta$_2$" (artículo indefinido) hace referencia a una nueva entidad en el discurso; en el ejemplo (9) tenemos que los demostrativos *Ese* y *aquel* funcionan estableciendo una deixis sobre dos entidades ya mencionadas conejo y zorro, haciendo una correferencia. Otros ejemplos apoyan la consideración anterior:

(11) Juan$_1$ está vendado porque ayer *el* perro$_2$ de *esa* casa$_3$ lo$_1$ mordió.

(12) Juan$_1$ chocó ayer. *Su$_1$* hija$_2$ *la* abogada$_2$ le$_1$ ayudó a tramitar *la* multa$_3$ en *el* departamento$_4$ de tránsito.

En el ejemplo (11) se observa que *el* (artículo) y *esa* (demostrativo), precediendo a los nombres perro$_2$ y casa$_3$, introducen nuevas entidades que dan información adicional sobre la causa del vendaje de Juan.

En el ejemplo (12) se observa que *su* (pronombre demostrativo) introduce una nueva entidad en el discurso (la hija de Juan); *la* (artículo) en su primera ocurrencia "*la* abogada$_2$" da nueva información adicional especificativa sobre cual de la hijas de Juan se está hablando; *el* (artículo) precediendo al "departamento$_4$ de tránsito" introduce una nueva entidad que da información adicional al evento del pago de una multa. Se observa que *la* (artículo) en su segunda ocurrencia "*la* multa$_3$" sí marca la presencia plausible de anáfora indirecta considerando, de acuerdo al conocimiento del sentido común, que un choque provoca una multa de tránsito al conductor; comprobar que esta hipótesis es verdadera es tarea del proceso de resolución de la anáfora indirecta tratado más adelante.

4.2.2 La elipsis nominal

Se conoce como elipsis a la figura de construcción, como mecanismo de economía lingüística, *para la omisión de un elemento del enunciado porque puede ser inferido o sobreentendido en el contexto del discurso*; no aparece ninguna entidad lingüística que deba ser vinculada con un antecedente, simplemente se deja un vacío, marcando entre corchetes la información omitida, por ejemplo:

(13) Juan *dibujó* una casa y Pedro [*dibujó*] una oveja.

(14) Juan toca el piano; María [*toca*] la guitarra.

La elipsis recibe su nombre completo en función del elemento omitido. Así en los ejemplos anteriores se le conoce como elipsis *verbal* porque es el *verbo* el elemento omitido. En este trabajo la importancia se concentra en la elipsis nominal, porque altera el uso normal del determinante como se muestra en los ejemplos:

(15) Juana compró **una** *lavadora nueva*, pero María compró **una** [*lavadora*] de segunda mano.

(16) El *compositor favorito* de Juan es Bach, pero **el** [*compositor favorito*] de José es Handel.

El determinante, **una** y **el** en los ejemplos, cumple la función determinativa sobre un nombre o una frase nominal, en la primera aparición, pero la elipsis permite omitirlos en la segunda aparición; en el segundo caso el determinante funciona como si fuera un pronombre extrínseco, afectando la función original del determinante en la oración.

4.2.3 Las expresiones referenciales

Una parte importante del algoritmo diseñado es la detección y marcado de las expresiones referenciales porque dependen del funcionamiento de las unidades léxicas de tipo determinante, que pueden representar funciones diferentes dentro de las oraciones, además de ser afectados por el fenómeno de elipsis; como primer ejemplo se presentan los determinantes *la y una*:

(17) Juan$_1$ baña a *la* niña$_2$ y José$_3$ *la$_2$* seca con la toalla$_4$.

(18) Juan$_1$ compró *una* paleta$_2$ de dulce. María$_3$ también compró *una$_4$*.

En el ejemplo (17), la primera aparición de *la* cumple la función de determinante (artículo definido) pero en la segunda funciona como pronombre en caso acusativo de la 3ª persona del singular. En el ejemplo (18) la primera aparición de *una* cumple la función de determinante (artículo indefinido) pero en la segunda funciona como pronombre extrínseco debido al fenómeno de elipsis; cabe hacer notar que, como en el ejemplo (39), la expresión *una$_4$* se refiere al mismo concepto de paleta pero a diferente objeto del mundo real, como se explicará en el apartado de la resolución de referencias 4.2.4.

Esta observación hizo necesario identificar las unidades léxicas que pueden funcionar como determinantes y sus funciones adicionales de mayor uso logrando elaborar una tabla de más de 100 que se presenta en el anexo A. Se presentan también, algunos ejemplos adicionales que agrupan estos fenómenos *indicando entre paréntesis la función* que cumple la palabra en *cursiva*.

(19) Ni él mismo sabía a ciencia *cierta* lo que pasaba. (adjetivo)

(20) Se expresaba con *cierta* dificultad al hablar ... (det. indef.)

(21) El techo tiene *algunas* manchas de humedad. (det. indef.)

(22) Caminé por las calles solitarias. En *algunas* había faroles... (pron. indef.)

(23) Juan le entregó una carta a Luis y *otra* a Sofía. (pron. indef.)

(24) Lo alcanzaremos en la *otra* calle. (det. indef.)

(25) ¿Corremos a la casa? Si, a la *una*, a las dos y a las tres. (sustantivo)

(26) Juan le regaló en su cumpleaños *una* pulsera. (det. cardinal)

(27) Es suficiente acudir de cada tres veces, *una*. (pron. indef.)

(28) Juan le guardaba desde entonces *una* gran fidelidad. (det. indef.)

Es importante observar en el ejemplo (25) el uso del determinante **la** que convierte en sustantivo a otro determinante **una** en la expresión "a **la** *una*" porque en Español "*se sustantiva todo aquello a lo que puede anteponerse un determinante*"; a continuación se presentan ejemplos de este fenómeno indicando entre paréntesis la categoría original de la palabra en *cursiva* que está siendo sustantivada y subrayando el determinante.

(29) Todos los colores me gustan pero el *rojo* es mi favorito. (adjetivo)

(30) Este *e*l es un artículo y no un pronombre ... (det. def.)

(31) Tu *reír* me fascina. (verbo inf.)

(32) El *ayer* ya no existe. (adverbio)

(33) Sobra esa *de* en tu oración de la tarea. (preposición)

(34) ¿Porqué pone tanto *pero* a este trabajo? (conjunción)

(35) La *frenada* del autobús fue muy rápida. (verbo part.)

De acuerdo a lo anterior, en el proceso de lectura (de izquierda a derecha) al encontrar un determinante se tienen tres casos posibles:

1) puede existir un nombre en forma inmediata o con modificadores adicionales interpuestos, considerado el caso más “normal”
2) puede existir una palabra que está cumpliendo la función del nombre (está siendo sustantivada) y que “normalmente” pertenece a una categoría diferente
3) puede no existir un nombre porque el determinante está cumpliendo otra función en la oración

Estas consideraciones son tomadas en cuenta en el algoritmo de detección y marcado de expresiones referenciales nominales. Mención especial merecen las preposiciones contraídas *al* (a el) y *del* (de el) que funcionan como determinantes y deben ser consideradas así por el algoritmo mencionado.

4.2.4 La referencia

Ferdinand de Saussure, en sus análisis teóricos, utilizaba siempre el término signo lingüístico bajo la siguiente definición: *asociación de un concepto a una imagen acústica específica*. De esté modo, el mismo concepto (denominado “significado”) suele tener en los diferentes lenguajes, diferentes imágenes acústicas (“significantes”). Tomando el *concepto* como un objeto mental y la imagen acústica como la palabra oral (en el lenguaje escrito como la expresión lingüística) es posible plantear el fenómeno de referencias en el discurso.

El discurso se da en el acto comunicativo, a través del texto, cuando el emisor introduce y discute sobre entidades (individuos, objetos y eventos, acciones, estados, relaciones o atributos) concretos o abstractos. Aunque existen varios tipos de expresiones referenciales, como las pro-formas locativa o temporal, este trabajo se limitará a las expresiones referenciales nominales.

Considerando la *entidad* como el concepto asociado con una expresión lingüística; parafraseando a Saussure “la asociación entre el significado y el significante”. Se puede definir la **expresión referencial** como *la estructura lingüística (expresión) permitida al emisor para introducir, o volver a mencionar, las entidades en el discurso*.

Entre las definiciones guía para la lingüística computacional en este dominio se tienen las siguientes [Trask, 1993]:

> Referencia .- el fenómeno por el cual una frase nominal en una oración o expresión particular es *asociada con alguna entidad* en el mundo conceptual o real, su referente.

Correferencia .- *la relación* que se obtiene entre dos frases nominales que se interpretan refiriéndose a una misma entidad extralingüística. En las representaciones lingüísticas, la correferencia es convencionalmente *denotada por coindexado*, por ejemplo:

(36) *María*$_1$, dijo que *ella*$_1$ vendría.

La tarea del receptor es llevar a cabo el proceso de *resolución de referencias*, que se logra con la secuencia de pasos siguiente:

1. identificar con la mayor precisión posible la entidad a la que el emisor está haciendo referencia (referido) con la expresión referencial
2. determinar si ésta entidad ya ha sido mencionada (referida) previamente en el contexto del discurso o es nueva
3. si ya ha sido mencionada (correferencia) creará el enlace entre expresiones referenciales en el contexto lingüístico; la *primera* mencionada ya posee el enlace a la entidad en el contexto del discurso
4. si es nueva debe de integrarla como parte del contexto lingüístico creando el enlace de la expresión referencial a la entidad en el contexto del discurso

Todo el proceso debe apoyarse en las características morfológicas y sintácticas del texto, que involucran la utilización de determinantes, y donde se encuentran íntimamente relacionados los conceptos de referencia, correferencia y anáfora indirecta, como se muestra en la figura 3.

En la descripción del proceso de resolución de referencias, representado en la figura 3, puede observarse que se considera a la *referencia* (1) como el enlace entre la expresión referencial y la entidad en el contexto del discurso; y la *correferencia directa* (2) como el enlace entre expresiones referenciales en el contexto lingüístico. En otras palabras, la correferencia directa es interna al contexto lingüístico (endofórica) y se apoya en la referencia,

que es externa al contexto lingüístico (exafórica), para mantener el enlace con la entidad en el contexto del discurso. También se observa que la *correferencia indirecta* (3) y la *anáfora indirecta* (4) requieren un proceso adicional de inferencia utilizando el conocimiento del sentido común para determinar su referencia o relación.

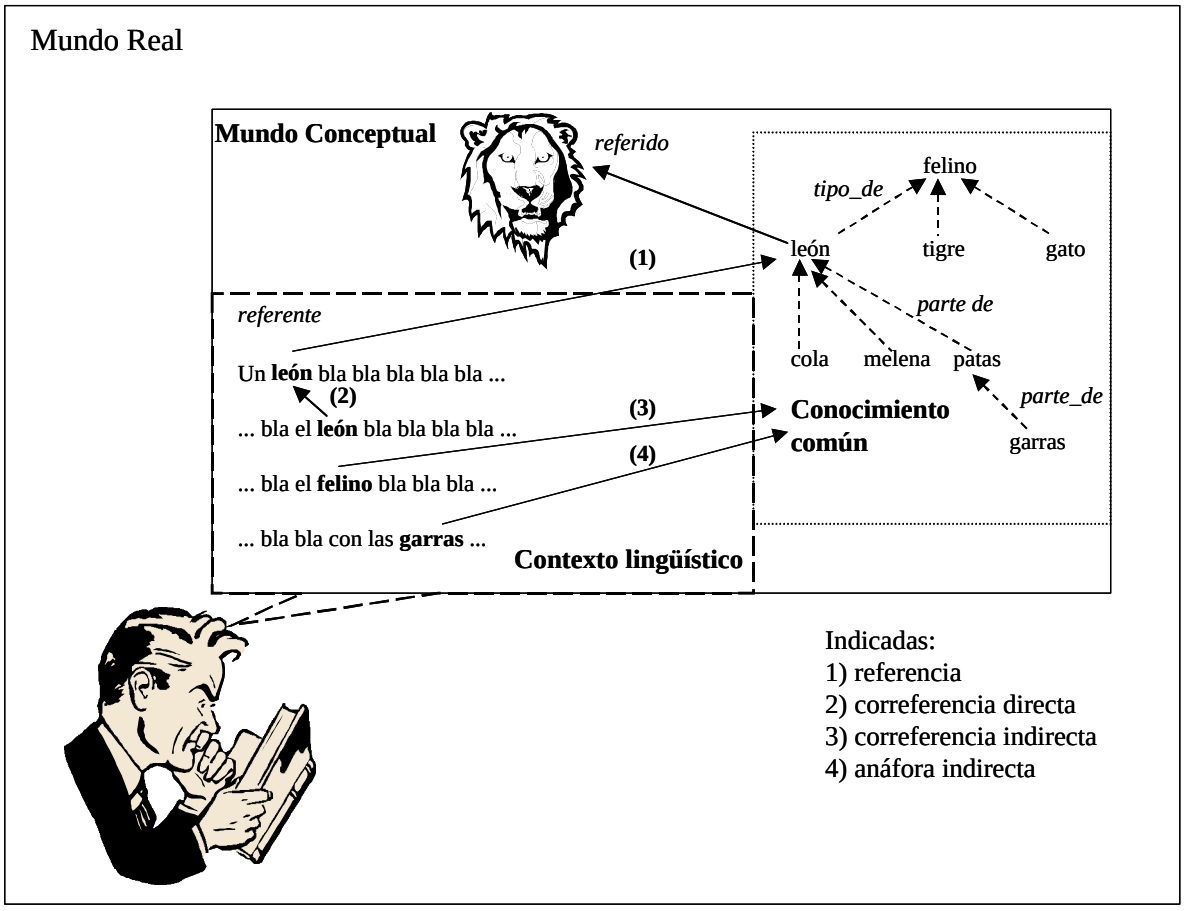


Figura 3 Proceso de resolución de referencias

Estas consideraciones son congruentes con la definición de correferencia de Trask y la del Glosario Lingüístico de Instituto Lingüístico de Verano [SIL] "*La correferencia es la referencia en una expresión al mismo referente en otra expresión*". Lo anterior se muestra con el siguiente ejemplo en el que ambos *tu* tienen el mismo referente:

(37) *Tu*$_1$ dijiste que [*tu*$_1$] vendrías.

Con esta visión de Saussure, la resolución de referencias permite compartir al emisor y receptor (a través del texto) los mismos conceptos por medio de la referencia y correferencia

directa. Compartir los mismos conceptos logra para ambos *el mismo sentido del texto* pero no implica que correspondan a las mismas entidades del mundo real.

Frege introdujo, en 1892, la distinción entre *sentido* y *referencia* (Alemán: *sinn* y *bedeutung*) aclarando que *el sentido de una expresión* se relaciona con las propiedades de su representación mental (o concepto) y *la referencia* se relaciona con el objeto o conjunto de objetos que la expresión denota. La hipótesis del triángulo, ver figura 4, dispone en cada uno de los tres vértices, la expresión (o signo lingüístico), el concepto (u objeto mental) y el objeto de la realidad limitada de los hechos, siguiendo el orden normal desde el habla hasta la realidad.

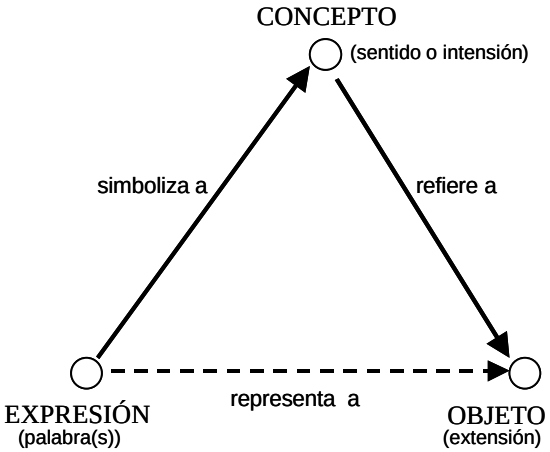


Figura 4 Triángulo referencial de Frege

En la figura 4 puede apreciarse que el uso efectivo del lenguaje nunca incide directamente sobre esta realidad, sino que llega a ella a través de un concepto más o menos convencional [Cerdá, 69]. Por ejemplo, en la noticia:

(38) El *presidente*$_1$ de la compañía XYZ$_2$ renunció la semana$_3$ pasada para ocupar un puesto$_4$ de gobierno$_5$, comentó el *presidente*$_6$ de la compañía [XYZ$_2$].

El **sentido** de la expresión referencial *presidente* es el mismo "*la persona responsable de dirigir una organización*", sin embargo, debido al intervalo de tiempo (desde la semana$_3$ pasada hasta el momento actual de la noticia) mostrado en el texto, se hace referencia en cada caso a objetos (personas) diferentes del mundo real; si el receptor no posee la información adicional requerida, su interpretación del texto se ve limitada a inferir que se trata de dos objetos diferentes sin poder establecer el enlace con el conocimiento común de la realidad.

Para poder llevar a cabo el proceso de resolución, en el ejemplo planteado, se requiere información adicional que permita resolver las referencias a los dos objetos del mundo real; esta información debe estar incluida en el contexto lingüístico de la noticia o en el conocimiento común del receptor. Lo anterior se muestra a continuación con la nota del ejemplo (38) modificada, incluyendo el nombre propio de cada *presidente*, representada en la figura 5.

(39) El *presidente*$_1$ de la compañía XYZ$_2$, *Juan P.*$_1$, renunció la semana$_3$ pasada para ocupar un puesto$_4$ de gobierno$_5$, comentó el *presidente*$_6$ de la compañía [XYZ$_2$], *José N*$_6$.

Si la información está incluida en la noticia (en el texto) el proceso de resolución de referencias se mantiene dentro del contexto lingüístico del discurso y la referencia se sigue considerando *referencia directa*, si es la primera mención de la entidad en el discurso, o *correferencia directa*, si ha sido mencionada anteriormente.

Si el receptor posee la información en su conocimiento común debe realizar un proceso adicional de inferencia para la resolución de referencias, entre el concepto y la entidad del mundo real, a la referencia obtenida con este proceso adicional de inferencia le denomino *referencia indirecta*, si es la primera mención del objeto en el discurso, o *correferencia indirecta*, si ha sido mencionado anteriormente.

(40) El *presidente*$_1$ habló ayer en el senado, les dijo que...

Si esta nota periodística la encontramos en un periódico mexicano de fecha actual el proceso de inferencia obtiene como referido al *presidente* Vicente Fox; si la nota fuese del año 1999, el proceso de inferencia obtiene al *presidente* Ernesto Cedillo; etc. Si el lector no tiene el conocimiento de la historia de México la referencia le permite relacionar el concepto pero estará incapacitado para conocer el objeto del mundo real; en otras palabras, ***el proceso de inferencia requiere el conocimiento para resolver la referencia***.

En la *correferencia* el referente y el referido hacen referencia a la misma entidad del discurso y al mismo objeto del mundo real; este fenómeno puede darse en forma *directa* (explícita) por identidad de cadenas léxicas (el referido es “idéntico” al referente) por ejemplo:

(41) Juan$_1$ discutió un *libro$_2$* interesante en su clase$_3$. Después, fuimos a tomar un café y discutí el *libro$_2$* con él$_1$.

(42) Un *carro$_1$* amarillo se estacionó frente a la casa$_2$. El *carro$_1$* permaneció allí casi una hora$_3$.

Puede observarse, en los ejemplos (41) y (42), que el concepto simbolizado por libro2 y carro1 hacen referencia a el mismo objeto del mundo real; diferente a como ocurre en los ejemplos (38) y (39), representados en la figura 5, donde el mismo concepto presidente se refiere a diferentes objetos Juan P. y José N.

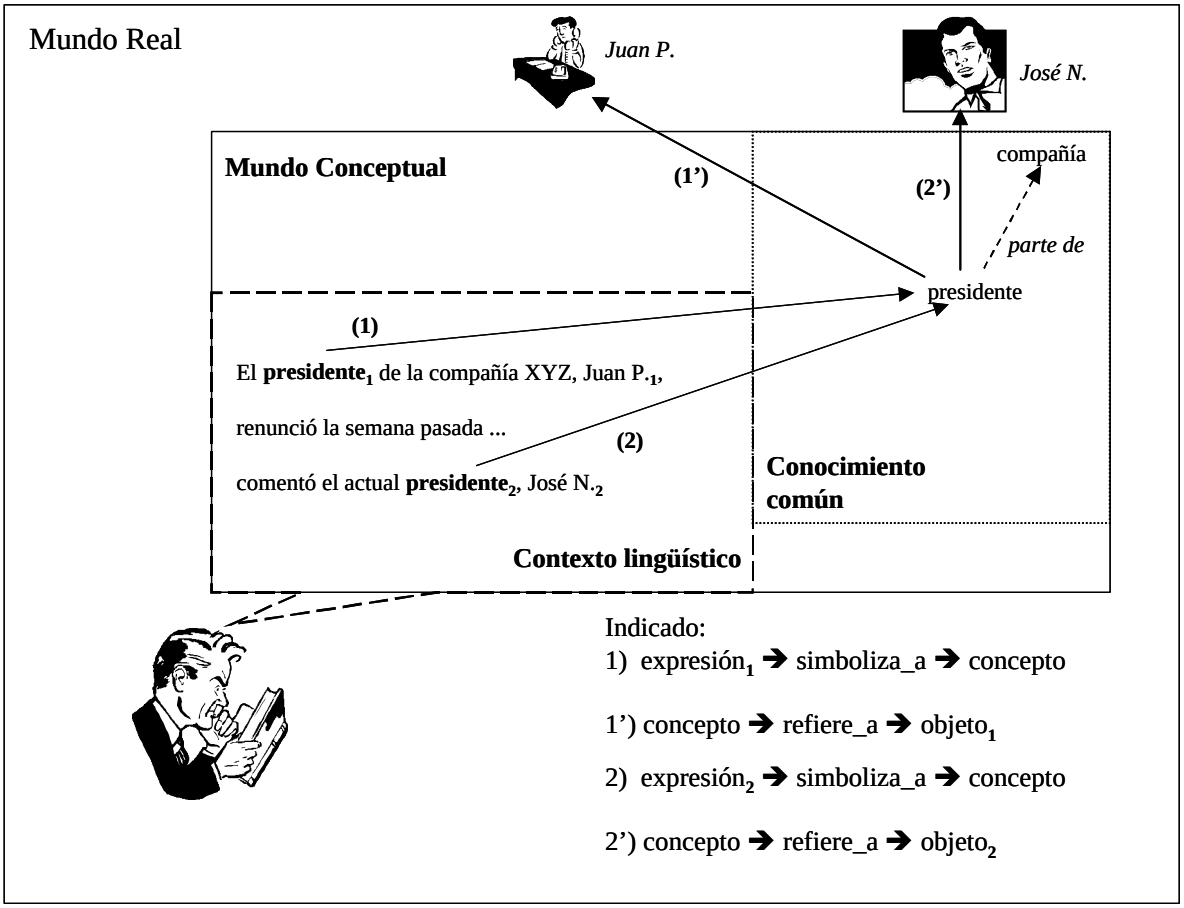


Figura 5 Ejemplo (39) de acuerdo a Frege

La *correferencia* se puede dar también en forma *indirecta* (implícita), a través de sinonimia por **hiperonimia o generalización** (el referente es más general que el antecedente), **hiponimia o especificación** (el referente es más específico que el antecedente) o

redescripción (se utilizan expresiones diferentes referidas a la misma entidad del discurso); lo anterior, se ilustra con ejemplos específicos a cada caso que se muestran a continuación.

(43) Hiperonimia o generalización

a) Un *carro*$_1$ se volcó en la carretera$_2$ anoche. El *vehículo*$_1$ venía a alta velocidad$_3$.

b) Un *autobús*$_1$... El *vehículo*$_1$...

c) Una *motocicleta*$_1$... El *vehículo*$_1$...

(44) Hiponimia o especificación

a) Juan$_1$ chocó ayer su *carro*$_2$ nuevo. El *sedán*$_2$ quedó desecho.

b) Juan$_1$ chocó ayer su *carro*$_2$ nuevo. El *chevy*$_2$...

c) Juan$_1$ chocó ayer su *volkswagwen*$_2$ nuevo. El *jetta*$_2$ ...

(45) Redescripción

a) Juan$_1$ chocó ayer su *carro*$_2$ nuevo. El *coche*$_2$ quedó desecho.

b) Un vocero$_1$ de *John Mayor*$_2$ dijo esta mañana$_3$ que el *primer ministro*$_2$ visitaría Moscú$_4$ la siguiente semana$_5$.

c) Pienso ahora en *María_Elena_Moyano*$_1$, la *Madre_Coraje*$_1$, y recuerdo aquel almuerzo$_2$ conmovedor, aquel poblado$_3$ espeluznante. El liderazgo$_4$ de *Maria_Elena*$_1$ nació de aquella miseria$_5$ y de una increíble voluntad$_6$ de superación.

Es importante tomar en cuenta que el análisis de las expresiones referenciales permite observar diferentes clases de entidades, en una escala que va de lo específico a lo genérico, disponibles al receptor [Fraurud, 1996]. Por ejemplo (se repite el ejemplo (39) para comodidad del lector):

El *presidente*$_1$ de la compañía XYZ$_2$, *Juan P.*$_1$, renunció la semana$_3$ pasada para ocupar un puesto$_4$ de gobierno$_5$, comentó el *presidente*$_6$ de la compañía [XYZ2], *José N*$_6$.

Las expresiones *Juan P.*$_{1}$ y *José N.*$_{6}$ son de la clase que Fraurud denomina *individual*, o entidades que tiene existencia por derecho propio (independiente de otras entidades), y que son directamente identificables por medio de nombres propios o designadores rígidos. La entidad de la clase *individual* es la más específica ya que señala un ente u objeto entre todos los demás. Cuando se interpreta una expresión de esta clase, que hace referencia a un individuo, la pregunta relevante es ¿Quién? o ¿Cuál?

La expresión *presidente* es de la clase, que Fraurud denomina *funcional*, de entidad que se concibe *sólo con relación a otra entidad*, que Fraurud denomina "ancla" (anchor), y que es identificable sólo a través ella; un ejemplo típico de esta clase es la relación *parte_de* como se puede observar en "la nariz [*de una persona*]". Cuando se interpreta una expresión de esta clase, que hace referencia a una entidad *funcional*, la pregunta relevante es ¿De quién? o ¿ De qué? La información de la entidad "ancla" debe ser introducida la primera vez que se utiliza la expresión referencial o estar en el contexto del discurso; en el ejemplo (39) "*de la compañía* XYZ_2" es una expresión definida que introduce la entidad *la compañía* XYZ_2 la primera vez que se menciona y la segunda vez es elidida pero puede ser recuperada.

La expresión *un puesto*$_{4}$ es de la clase más genérica, que Fraurud denomina *ejemplar* (instance), y se concibe como una entidad perteneciente a alguna categoría. Cuando se interpreta una expresión de esta clase, que hace referencia a una entidad *ejemplar*, la pregunta relevante es ¿Qué?; la categoría puede ser identificada por la información en el texto *de gobierno*; la expresión indefinida, en este caso, no requiere la identificación de un objeto específico del mundo real.

El mecanismo de inferencia, en el proceso de resolución de referencias, tiene como objetivo encontrar el referente dentro de la memoria *episódica* que *almacena la información de hechos sobre cosas y eventos de la clase individual*. En otras palabras, el mecanismo de inferencia para la resolución de referencias se basa en la *extensión* de la expresión referencial, hasta llegar como última instancia al nombre propio. Aquí se considera la extensión de la expresión referencial (ver figura 4) como *el conjunto de objetos a las cuales se aplica esta expresión* [Sowa, 1984].

Resumiendo, la referencia es el mecanismo de alusión a un objeto, a través de una entidad mencionada en el texto; es *referencia directa* si hace alusión por primera vez a un

objeto del mundo real; y es *referencia indirecta* si para aludir por primera vez al objeto se requiere un proceso de inferencia basado en relaciones de sinonímia; es *correferencia directa* si hace alusión a un objeto mencionado previamente; es *correferencia indirecta* si para aludir a un objeto mencionado previamente se requiere un proceso de inferencia que tome en cuenta las relaciones de sinonímia.

4.2.5 La anáfora

La anáfora deriva su nombre del griego ana = "otra vez" y phero = "traer o acarrear"; es considerada como una figura retórica de ***repetición*** de una expresión (palabra o grupo de palabras) al principio de enunciados, oraciones o líneas. El diccionario la define como "tipo de deixis que desempeñan ciertas palabras para ***asumir el significado*** de una parte del discurso ya emitida" [DRAE].

En lingüística se ha ampliado el término y se considera como "una ***relación*** entre dos expresiones lingüísticas donde la interpretación **de una** (llamada *anáfora*) está en alguna forma ***determinada por la interpretación de la otra*** (llamada *antecedente*)" [Huang, 2000]. Estás tres definiciones permiten observar una evolución, similar al concepto de referencia hasta llegar al triángulo de Frege, que va de la repetición de expresión hasta la repetición del sentido o significado; además de la dependencia del sentido de la anáfora con respecto al sentido del antecedente. Con estas observaciones se puede definir la anáfora como un mecanismo para hacer referencia de una entidad *anáfora* (o referente) a una entidad *antecedente* (o referido) que ya ha sido mencionada en el texto ***donde el antecedente provee la información necesaria para la correcta interpretación de la expresión anafórica***. La anáfora es considerada una de las principales formas de cohesión y consiste en la repetición de referencias a una misma entidad en oraciones sucesivas. Por ejemplo:

(46) *Juan*$_1$ no está de acuerdo. *Él*$_1$ cree que debe hacerse fuera y ya [*él*$_1$] ha empezado a sacar instrumentos$_2$ a la terraza$_3$.

Se puede observar que *Juan*$_1$, *Él*$_1$ y la elipsis de sujeto *[él*$_1$*]* en *ha empezado* se refieren a la misma entidad y hacen referencia al mismo objeto (persona) del mundo real; por lo tanto, en el ejemplo (46) se dan los fenómenos de anáfora y correferencia. Si no dispusiéramos de mecanismos para evitar la repetición del nombre *Juan*, el texto llegaría a ser reiterativo;

asimismo, si no existieran las referencias necesarias para esta entidad en el lugar adecuado, las frases serían incompletas y el texto no podría entenderse; el principal mecanismo de que disponemos para evitar estas repeticiones, es el uso de pronombres. Hay dos clases de pronombres: intrínsecos (o gramaticales) y extrínsecos.

Los pronombres **intrínsecos** son palabras que **siempre son pronombres** sin importar el contexto en que aparezcan. A esta clase pertenecen: los pronombres personales; los pronombres relativos: que, cual, quien y cuanto (con sus variantes de género y número); los pronombres interrogativos y exclamativos: quién, quiénes, cuál, cuáles; los pronombres demostrativos neutros: esto, eso, aquello; y los pronombres indefinidos: algo, nada, alguien, nadie, quienquiera, quienesquiera.

Los pronombres **extrínsecos** son palabras que sólo en determinados contextos desempeñan la función de pronombres. En la mayoría de los casos son determinativos que al omitir el sustantivo en el contexto (elipsis nominal) actúan como pronombres, como se observó en los ejemplos presentados al analizar las expresiones referenciales (4.2.3). Además el fenómeno de elipsis, al suprimir un elemento conocido que aparece muy cerca del original en el texto y que el receptor puede reconstruir (sujetos, complementos, etc.), logra funcionar como anáfora, sin la utilización del pronombre o alguna otra unidad léxica como se mostró en el ejemplo (46) con la elipsis de sujeto *[él$_1$]* en “ha empezado”. En otros casos, adverbios como: *allí*, allá, aquí, etc.; pueden actuar como sustitutos en algunos contextos determinados como se muestra en el ejemplo (47).

(47) Tus amigos$_1$ se fueron al bar *Universal$_2$*. Los$_1$ encontrarás a todos$_1$ *allí$_2$*.

Es importante, observar que en el fenómeno de la anáfora aparece una expresión que debe ser vinculada con otra previamente mencionada por medio de algún tipo de relación; si **ambas expresiones *se encuentran explícitamente*** en la oración **y la relación es preestablecida** por la gramática del lenguaje se le conoce como anáfora ***directa***.

(48) Juana$_1$ baña al *bebé$_2$* y María$_3$ *lo$_2$* seca con la toalla$_4$.

En el ejemplo (48) la anáfora *lo*$_2$ y el antecedente *bebé*$_2$ se encuentran explícitamente en la oración; además la relación del pronombre *lo*$_2$ está preestablecida en la gramática del Español como referencia a una entidad masculina de 3ª persona del singular.

La anáfora, como se ha visto, tiene un orden de aparición en la oración **antecedente ➔** *anáfora* (mecanismo de referencia hacia atrás), cuando el orden se invierte se le denomina *catáfora* porque el antecedente precede a la catáfora de la cual depende su interpretación (mecanismo de referencia hacia adelante). Ejemplos:

(49) Ya estaban *todos* allí esperándote: **Pepe, María, Juan y Pedro**.

(50) Si necesitas *una*, hay **toallas** en el ropero.

En los ejemplos *todos* y *una* se consideran catáforas. “Pepe, María, Juan y Pedro” así como “toallas” son expresiones que proveen la información necesaria para la correcta interpretación de las catáforas en el texto.

Se dice que la anáfora es **indirecta** cuando la anáfora, el antecedente o ambos se encuentran implícitos y se requiere información adicional del sentido común para resolverla por medio de un proceso de inferencia que identifique la relación entre ambas entidades. Por ejemplo:

(51) Juan estuvo *comiendo*. La *mesa* está sucia...

(52) Juan caminaba en *la sala de conciertos*. El *piano* era del siglo XIX.

En el ejemplo (51) la anáfora *mesa* tiene relación con el acto de *comer* (en general se utiliza una mesa para comer = la mesa donde Juan estuvo *comiendo*). En el ejemplo (52) la anáfora *piano* tiene relación con la *sala de conciertos* (el piano es un instrumento musical que se utiliza en los conciertos = el piano que Juan vio mientras caminaba en la sala de conciertos). Es importante observar que en el ejemplo (51) la anáfora es una entidad nominal y el antecedente es una entidad verbal, mientras en el ejemplo (52) la anáfora y el antecedente son entidades nominales. Las relaciones nominal ➔ verbal y nominal ➔ nominal sugieren la necesidad de información diferente para poder inferir el tipo de relación existente; estas consideraciones afectan el método de resolución de la anáfora.

El mecanismo de inferencia para la resolución de anáfora se basa en el *sentido* de la expresión referencial (ver figura 4). Aquí se considera el sentido de la expresión referencial como *la parte del significado que se obtiene de principios generales aplicados a esta expresión* [Sowa, 1984]. El mecanismo de inferencia en el proceso de resolución de la anáfora, a diferencia del caso de la referencia, tiene como objetivo encontrar si existe *una relación* entre la anáfora y el antecedente dentro de la memoria *semántica* que almacena la información de principios universales, atributos y relaciones de la clase ***Funcional*** [Sowa, 1984; Fraurud 96].

Resumiendo, la anáfora es considerada una de las principales formas de cohesión y consiste en la repetición de referencias a una misma entidad (concepto) en oraciones sucesivas dentro del texto; es *anáfora directa* si hace referencia a una entidad por medio de una relación preestablecida en el lenguaje; y es *anáfora indirecta* si hace referencia a una entidad por medio de una relación que puede identificarse, a través de un proceso de inferencia, tomando en cuenta el conocimiento del sentido común.

4.2.6 Interacción entre referencia y anáfora

Después de analizar los fenómenos de referencia, correferencia y anáfora en forma individual, en este apartado, se hará un análisis conjunto; se busca puntualizar las diferencias y similitudes apreciando en los ejemplos que la misma expresión puede ser: correferente y anáfora; sólo anáfora sin ser correferente; y finalmente sólo referente a nueva información si no es correferente ni es anáfora. Se busca responder a dos preguntas que surgen al considerar cada expresión ¿hay algún marcador específico para cada caso? ¿qué relación existe entre estos fenómenos?

Para responder a la primera pregunta se presenta el concepto y objeto "*libro*" en los siguientes ejemplos:

(53) El viernes$_1$ pasado, José$_2$ presentó el *libro$_3$* de Juan$_4$ "Volcanes activos del mundo"$_3$. Después cené con Juan$_4$ y comentó que le$_4$ había llevado tres años$_5$ escribir el *libro$_3$*.

(54) María$_1$ compró un *ejemplar$_2$* de "Aprenda fotografía en 21 días"$_2$ y se molestó porque el *libro$_2$* tenía imágenes$_3$ borrosas.

(55) Pedro$_1$ tiene algunos problemas$_2$ con Matemáticas$_3$. Él dice que el *libro$_4$* no es el [*libro$_4$*] adecuado.

El análisis del ejemplo (53) se observa que *libro$_3$* en su primera aparición **es una referencia directa** a un objeto del mundo real; *libro$_3$* en su segunda aparición **es una correferencia directa** por "identidad de cadenas léxicas"al mismo objeto a través del mismo concepto.

En el ejemplo (54) *libro$_2$* **es una correferencia indirecta** a *ejemplar$_2$* debido a la relación de sinonimia entre los conceptos. En el ejemplo (55) *libro$_4$* en su primera aparición **es una anáfora indirecta** a Matemáticas$_3$ ("normalmente" se utiliza un libro, es parte del curso, para estudiar cada asignatura o materia, en este caso Matemáticas$_3$); en el mismo ejemplo (55) por el fenómeno de elipsis en [*libro$_4$*] el determinante ***el*** funciona como pronombre extrínseco para hacer una anáfora directa con *libro$_4$*.

En todos los casos *libro* está precedido por el determinante *el*; la respuesta a la pregunta y conclusión del análisis es "***no hay un marcador específico para cada caso porque el mismo determinante (***artículo ***el) puede utilizarse para todos los casos***".

Es importante puntualizar que en la anáfora directa *la información para resolver la anáfora está explícita en la oración* (la parte visible, recordando el iceberg en la figura 1) y su relación está establecida por la gramática del lenguaje.

Por otro lado, en la anáfora indirecta *la información y relación para resolverla están implícitas* en el antecedente, en la anáfora o en ambos; lo anterior permite identificar tres casos posibles [Gelbukh y Sidorov, 1999], como se muestra en la tabla 1. El orden numérico, en la tabla, se establece tomando en cuenta la frecuencia y complejidad; por ejemplo, el caso "Indirecta 1" es más frecuente y menos complejo de resolver mientras que el caso "Indirecta 3" es el menos frecuente y más complejo de resolver (requiere información implícita de la anáfora y el antecedente); en cada ejemplo de anáfora indirecta se comentará el caso en que está incluida.

		Inf. del Antecedente	
		explícita	implícita
Inf. de la Anáfora	explícita	**Directa**	**Indirecta 1**
	implícita	**Indirecta 2**	**Indirecta 3**

Tabla 1 Casos de anáfora

Para responder a la segunda pregunta, ¿qué relación existe entre estos fenómenos?, se analizaron los fenómenos que se muestran en la columna **Descripción** de la tabla 2 y su relación con los diferentes tipos de expresión. En los ejemplos que se muestran a continuación se marca con **negrita** el referente o antecedente y con *cursiva* la correferencia o anáfora.

Descripción	Tipo	Expresión	Relación
referencia	directa	nombre propio	definida
		apelativo	definida
		det + nombre común	definida
			indefinida
	indirecta	det + nombre común	sinonimia
correferencia	directa	nombre propio	definida
		apelativo	definida
		det + nombre común	definida
	indirecta	det + nombre común	sinonimia
anáfora	directa	pronombre intrínseco	definida
		pronombre extrínseco	funcional
			elipsis
	indirecta	det + nombre común	holonimia
			meronimia
			rol

Tabla 2 Relaciones en expresiones nominales

4.2.6.1 Ejemplos de referencia directa

(56) **Hugo Sánchez**$_1$ pateó el balón$_2$ con fuerza$_3$ anotando gol$_4$.

(57) Los **Pumas**$_1$ lograron así el campeonato$_2$ del futbol$_3$ mexicano.

Los nombres propios son como etiquetas, identificadoras de seres y objetos que señalan *un determinado ser entre los demás de su clase*. En otras palabras, a través de los conceptos de clase (persona y equipo de fútbol) se puede hacer referencia a los seres y objetos señalados por el nombre propio. Son ejemplos de nombres propios: los nombres de pila, los apellidos, los apodos o apelativos. etc. **Hugo Sánchez**$_1$ y **Pumas**$_1$ son referencias directas (mencionan por primera vez) al jugador y al equipo de fútbol mexicanos, por medio de un nombre propio y un apelativo respectivamente.

(58) El {***perro de Juan*****}**$_1$ mordió a un **niño$_2$**.

En el ejemplo (58) “perro” es nombre común y por si mismo no identifica al objeto del mundo real sólo al concepto, se requiere pues el artículo definido “el” y la especificación del propietario “de Juan” para poder hacer la referencia directa (primera mención). Es el {**perro** de Juan}$_1$, y no otro, el que mordió a un niño$_2$, del cual no conocemos su identidad. Nótese que la expresión nominal está en función de la situación porque si Juan tuviese más de un perro para definir o individualizar al perro es necesario encontrar alguna característica que lo determine o distinga de los demás perros de Juan; por ejemplo el color (negro, blanco, café), la raza (bulldog, doberman, cazador), etc. En otras palabras, la expresión nominal es tan amplia (en número de palabras) como lo requiera el proceso de comunicación, lo que se muestra con el ejemplo modificado.

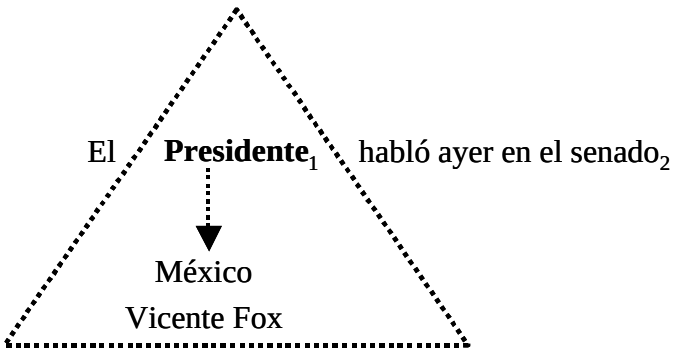
(59) El {*perro negro de Juan*}$_1$ mordió a un **niño$_2$**.

A través del concepto de clase niño se puede hacer referencia a los seres y objetos señalados por el nombre común. El artículo indefinido “un” se usa para hacer referencia directa (primera mención) a un objeto del mundo real **niño$_2$**; no conocemos su identidad pero sabemos que existe porque es el objeto mordido por el {perro negro de Juan}$_1$.

4.2.6.2 Ejemplo de referencia indirecta

Observando las relaciones de la tabla 2 se aprecia que los tipos “indirecta” requieren un proceso de inferencia apoyado en el conocimiento común para resolverlas. Para presentar mejor los ejemplos es necesario introducir dos símbolos: un triángulo y una flecha. El triángulo representa la información explícita, por medio de una palabra (de la oración) en su vértice superior (la parte visible, recordando el iceberg en la figura 1) y la información implícita, por medio de un conjunto de palabras que representan el conocimiento común para la resolución; por supuesto, se presentan sólo las palabras necesarias para ilustrar el ejemplo. La flecha sirve de guía o enlace en el proceso de resolución; a mayor número de flechas (enlaces) el proceso de resolución representado es más complejo; en este caso, las flechas se presentarán enumeradas para indicar la secuencia del proceso.

(60)



En el ejemplo (60), **Presidente**$_1$ y senado$_2$ son conceptos aplicables a países que tienen un sistema de gobierno republicano, similar al de México. Si está noticia aparece en un periódico mexicano, el receptor, haciendo uso del contexto lingüístico y por un proceso de inferencia, puede acceder al conocimiento común del país “México” y obtener que el “presidente actual” es Vicente Fox; identifica de este modo que el emisor hace **referencia indirecta** a Vicente Fox como Presidente actual. Es importante notar en el ejemplo que el sentido o significado comunicativo de la oración es completo a nivel conceptual; *para identificar el objeto del mundo real el receptor necesita llevar a cabo el proceso de inferencia.*

4.2.6.3 Ejemplos de correferencia directa

(61) Pienso ahora en **María_Elena_Moyano**$_1$ y recuerdo aquel almuerzo$_2$ conmovedor, aquel poblado$_3$ espeluznante. *Maria_Elena$_1$* tenía un liderazgo$_4$ que nació de aquella miseria$_5$ y de una increíble voluntad$_6$ de superación.

En el ejemplo (61) **María_Elena_Moyano**$_1$ es una referencia directa a una mujer o persona del género femenino (primera vez que se menciona) y *Maria_Elena*$_1$ es una correferencia directa, a través del nombre propio, porque es una mención posterior a la primera; normalmente el nombre propio se utiliza completo sólo una vez en el texto y en las demás menciones se utiliza únicamente parte del mismo (nombre o apellidos).

(62) **Mario_Moreno**$_1$ murió ayer en la tarde. *Cantinflas$_1$* será recordado como el más grande cómico$_2$ del cine$_3$ mexicano.

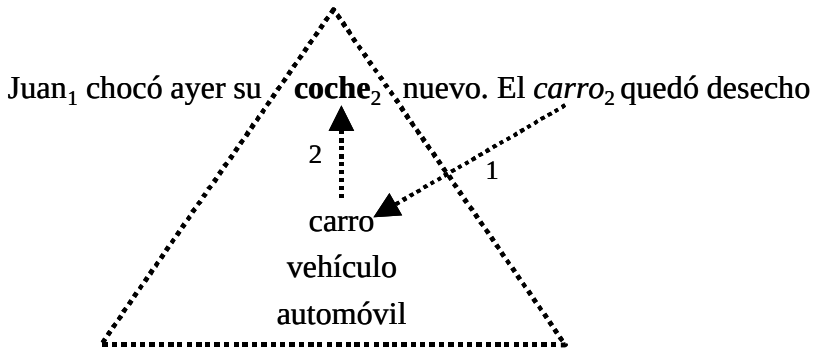
En el ejemplo (62) *Cantinflas*$_1$ es una **correferencia directa**, por medio de el apodo o apelativo, porque así se le conoce al actor cómico mexicano **Mario_Moreno**$_1$; es importante recordar que el apelativo es un tipo de nombre propio que se adjudica para individualizar; en otras palabras, se puede hablar de nombres propios sinónimos que se utilizan para referirse a un mismo ser u objeto del mundo real.

(63) Un **carro**$_1$ se estacionó frente a la casa$_2$. *El carro$_1$* estuvo allí casi una hora$_3$.

En el ejemplo (63) la **correferencia** es **directa** porque las cadenas léxicas son idénticas, *carro*$_1$, y en ambos casos hace referencia al mismo objeto del mundo real.

4.2.6.4 Ejemplo de correferencia indirecta

(64)



En el ejemplo (64) la **correferencia** es **indirecta** porque las unidades léxicas son diferentes, **coche**$_2$ y *carro*$_2$, y en ambas hacen referencia al mismo objeto del mundo real; para poder establecer el enlace de la referencia es necesario un proceso de inferencia que, a través de sus conceptos, permita "conocer" que *carro* y *coche* son sinónimos (el mismo o similar concepto). El proceso se ilustra con dos flechas numeradas que indican la secuencia: primero ir al conocimiento "implícito" del antecedente para encontrar la unidad léxica que señala al mismo concepto y segundo identificar que señala al mismo objeto del mundo real.

En los ejemplos se ha mostrado que para determinar el objeto referenciado es necesario hacerlo a través del concepto (ver la figura 4) principalmente en el caso de correferencia indirecta. Lo anterior permite concluir que el fenómeno de correferencia es intrínsecamente anafórico porque necesita comparar la similitud de conceptos para determinar al referente; en pocas palabras, *no puede existir la correferencia sin anáfora*.

4.2.6.5 Ejemplos de anáfora directa

(65) ***María***$_1$ terminó su noviazgo$_2$ con ***Juan***$_3$. *Él*$_3$ se molestó mucho con *ella*$_1$.

En el ejemplo (65) la **anáfora** es **directa** porque las unidades léxicas *Él*$_3$ y *ella*$_1$ son pronombres personales intrínsecos cuya función está predefinida por la gramática del lenguaje. Ambos deben sustituir a la tercera persona del singular; *él* sustituye al género masculino y *ella* al género femenino.

(66) Ayer José$_1$ fue al ***estadio Azteca***$_2$. *Allí* encontró a Juan$_3$ y María$_4$ para ver el partido América$_5$ contra Guadalajara$_6$

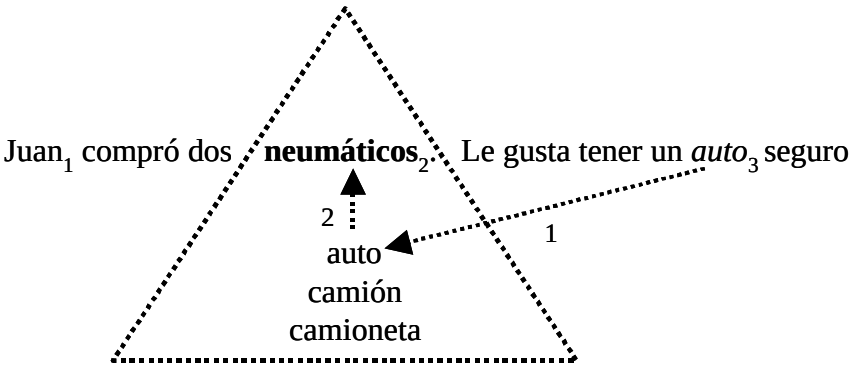
En el ejemplo (66) la **anáfora** es **directa** porque el adverbio de lugar *Allí* funciona como pronombre extrínseco para hacer referencia al **estadio Azteca**$_2$.

(67) Juan$_1$ compró *un* ***carro***$_2$ nuevo pero José$_3$ compró *uno*$_2$ *[carro]* usado$_3$

En el ejemplo (67) la **anáfora** es **directa** porque el número cardinal *uno*$_2$ debido al fenómeno de elipsis funciona como pronombre extrínseco para hacer referencia al concepto de **carro**$_2$. Es importante notar que en cada caso se hace referencia a diferente objeto; primero al carro nuevo de Juan y después al carro usado de José (aunque el fenómeno de anáfora

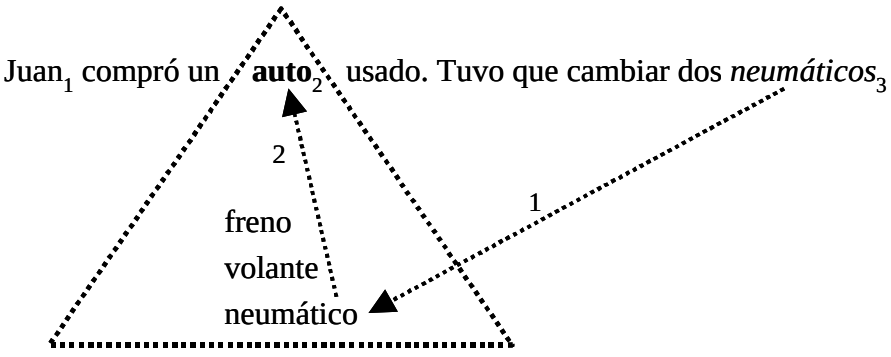
permite hacer referencia al concepto de carro no se presenta el fenómeno de correferencia). En otras palabras, la anáfora utiliza el concepto de *carro* para obtener el sentido o significado de **uno**$_2$; si hace referencia a diferente objeto o no es debido al fenómeno de correferencia; conclusión, ***puede existir anáfora directa sin correferencia*.**

4.2.6.6 Ejemplos de anáfora indirecta



(68)

En el ejemplo anterior la **anáfora** es **indirecta** debido a la relación de holonimia entre *auto*$_3$ que está compuesto o contiene **neumáticos**$_2$ (lo mismo que un camión, camioneta, etc.); este es un ejemplo del primer caso de anáfora indirecta donde la información implícita en el antecedente permite resolver la anáfora; además permite apreciar el uso de un determinante indefinido "un" marcando la presencia de la anáfora indirecta. Los dos **neumáticos**$_2$ (con su información implícita) forman parte del contexto lingüístico cuando en la segunda oración se menciona la anáfora *auto*$_3$; al acceder a la información implícita se encuentra auto como concepto común con *auto*$_3$ (marcado con la flecha 1) y por medio de la relación de holonimia (marcada con la flecha 2) se puede inferir que un *auto*$_3$ de Juan$_1$ va a recibir los **neumáticos**$_2$ que compró.

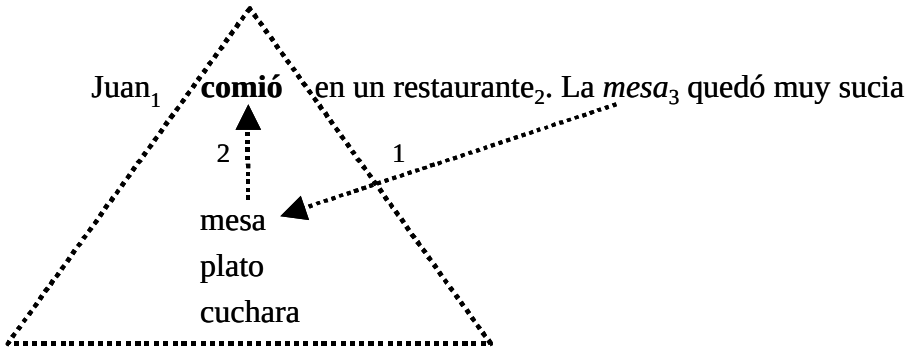


(69)

En el ejemplo anterior la **anáfora** es **indirecta** debido a la relación de meronimia de los componentes o partes del **auto**$_2$ (neumático, volante, freno, etc.); este es un ejemplo del primer caso de anáfora indirecta donde la información implícita en el antecedente permite resolver la anáfora. El **auto**$_2$ (con su información implícita) ya forma parte del contexto lingüístico cuando en la segunda oración se menciona la anáfora *neumáticos*$_3$; al acceder a la información implícita se encuentra neumático como concepto común con *neumaticos*$_3$ (marcado con la flecha 1) y por medio de la relación de meronimia (marcada con la flecha 2) se puede inferir que son componentes del **auto**$_2$ los neumáticos que tuvo que cambiar.

En los ejemplos (68) y (69) analizados la anáfora y el antecedente son expresiones nominales; en los siguientes ejemplos las relaciones se engloban bajo el concepto de rol donde el antecedente puede ser una expresión *nominal* o *verbal*. En este trabajo se denomina rol, al tipo de relación que se establece entre la anáfora y el antecedente de acuerdo a la función semántica (agente, paciente, beneficiario, compañía, lugar, instrumento, etc.) que desempeñan en la oración (como analogía al papel que ha de representar un actor en un escenario). También se repetirá la primera oración, en algunos ejemplos, para ilustrar más claramente los diferentes tipos relaciones que puede establecer la *anáfora* con el mismo **antecedente**.

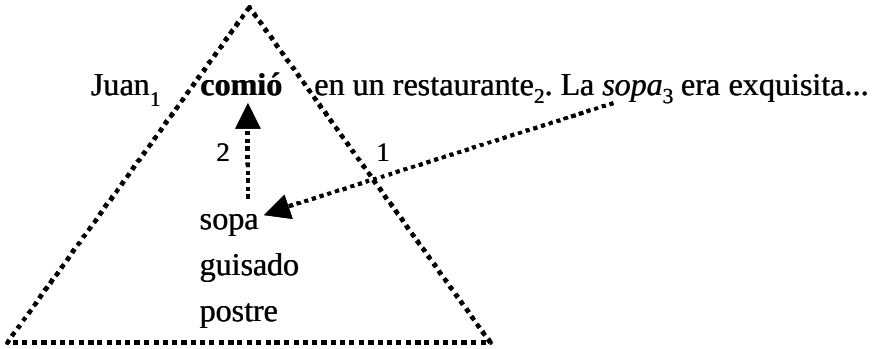
(70)



En el ejemplo anterior la **anáfora** es **indirecta** debido al rol de *mesa*$_3$ como instrumento con el antecedente **comió** porque en general se utiliza una mesa para comer; es un ejemplo del primer caso de anáfora indirecta donde la información implícita en el antecedente permite resolver la anáfora; además el antecedente es una expresión **verbal** y la anáfora una *nominal*. El antecedente **comió** (acto de comer) ya forma parte del contexto lingüístico cuando en la segunda oración se menciona la anáfora *mesa*$_3$; al acceder a la

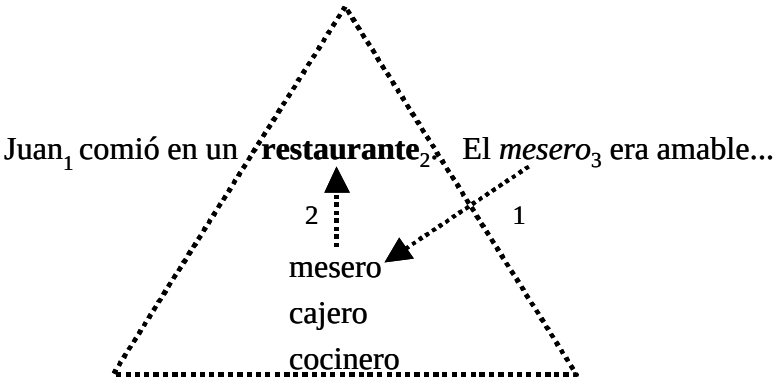
información implícita se encuentra mesa como un mueble utilizado para comer (marcado con la flecha 1) y por medio de la relación de instrumento (marcada con la flecha 2) se puede inferir que es la *mesa*$_3$ donde Juan$_1$ **comió** la que quedó muy sucia.

(71)



En el ejemplo anterior la **anáfora** es **indirecta** debido al rol de *sopa*$_3$ como paciente con el antecedente **comió** porque es un tipo de alimento, ingerido por el agente Juan$_1$; es un ejemplo del primer caso de anáfora indirecta donde la información implícita en el antecedente permite resolver la anáfora; además el antecedente es una expresión **verbal** y la anáfora una *nominal*. El antecedente **comió** (acto de comer) ya forma parte del contexto lingüístico cuando en la segunda oración se menciona la anáfora *sopa*$_3$; al acceder a la información implícita se encuentra sopa como un tipo de alimento (marcado con la flecha 1) y por medio de la relación de paciente (objeto que recibe la acción del verbo, marcada con la flecha 2) se puede inferir que la sopa exquisita era la que Juan comió.

(72)



En el ejemplo anterior la **anáfora** es **indirecta** debido al rol de *mesero*$_3$ como actor con el antecedente **restaurante**$_2$ porque es un tipo del personal para la atención de clientes; es un ejemplo del primer caso de anáfora indirecta donde la información implícita en el

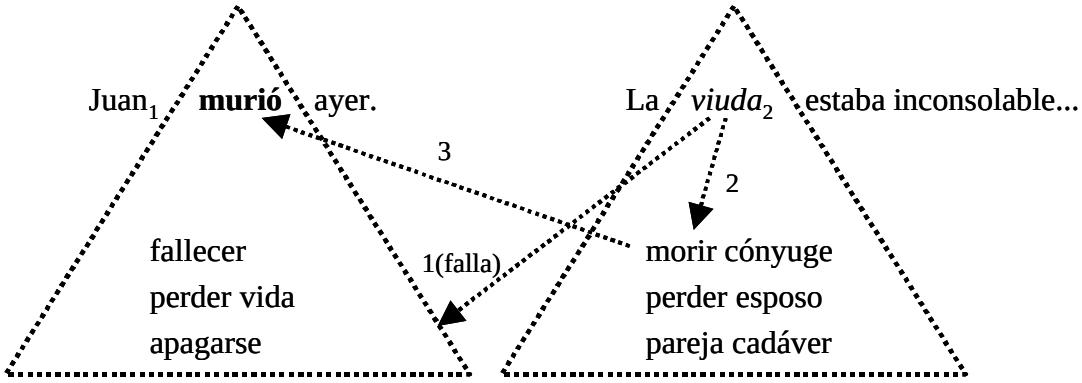
antecedente permite resolver la anáfora; además el antecedente es una expresión **nominal** y la anáfora también es una expresión *nominal*. El antecedente **restaurante**$_2$ ya forma parte del contexto lingüístico cuando en la segunda oración se menciona la anáfora *mesero*$_3$; al acceder a la información implícita se encuentra mesero como un tipo de personal (marcado con la flecha 1) y por medio de la relación de actor (marcada con la flecha 2) se puede inferir que el mesero amable era del restaurante donde Juan$_1$ comió.

Se ha presentado la anáfora indirecta para el caso 1, en los ejemplos (69) al (72). Para presentar el tipo de anáfora indirecta de los casos 2 y 3 es conveniente tomar como base la anáfora directa donde la anáfora y el antecedente están explícitos en la oración; lo anterior permitirá contrastar la necesidad de un proceso de inferencia y del conocimiento común implícito para resolver la anáfora indirecta.

(73) ***Juan***$_1$ murió ayer. *Su*$_1$ viuda$_2$ estaba inconsolable...

En el ejemplo anterior la **anáfora** es **directa** porque el determinativo "su" funciona como pronombre extrínseco posesivo para hacer referencia a **Juan**$_1$; aplicando sólo las reglas gramaticales a "*Su*$_1$ viuda$_2$" se puede determinar que se refiere a la viuda de **Juan**$_1$.

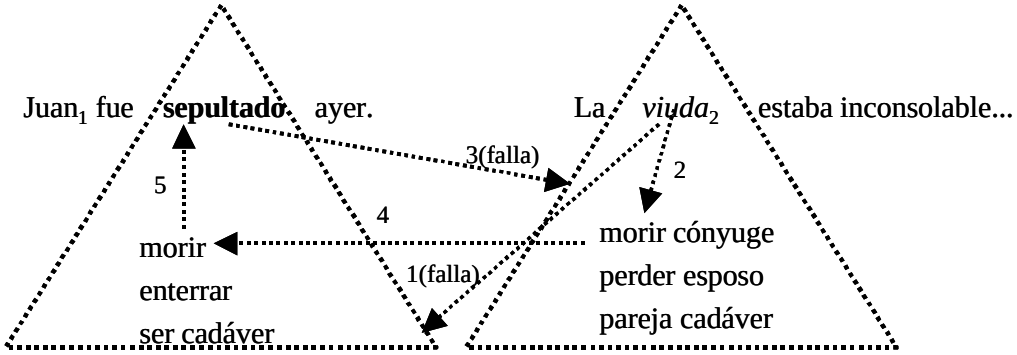
(74)



En el ejemplo anterior la **anáfora** es **indirecta** debido al rol de *viuda*$_2$ como paciente con el antecedente **murió** porque recibe la consecuencia de la muerte de Juan$_1$, ("morir" es un verbo intransitivo); es un ejemplo del segundo caso de anáfora indirecta donde la información implícita en el antecedente no permite resolver la anáfora y es necesario la información implícita de la anáfora para lograrlo; además el antecedente es una expresión **verbal** y la anáfora es una expresión *nominal*. El antecedente **murió** ya forma parte del contexto lingüístico cuando en la segunda oración se menciona la anáfora *viuda*$_2$; al acceder a la

información implícita de **murió** no se encuentra viuda (falla el proceso, marcado con la flecha 1); es necesario obtener la información implícita de *viuda*$_2$ para poder buscar el concepto del antecedente **murió** (proceso marcado con la flecha 2); finalmente se encuentra el concepto del antecedente **murió** en la información implícita de *viuda*$_2$ (morir cónyuge, marcado con la flecha 3). Se puede inferir que la *viuda*$_2$ estaba inconsolable por el hecho de "morir el cónyuge", llamado **Juan**$_1$.

(75)



En el ejemplo anterior la **anáfora** es **indirecta** debido al rol de *viuda*$_2$ como paciente con el antecedente **sepultado** porque recibe la consecuencia de que Juan$_1$ sea sepultado, debido a que ha muerto; es un ejemplo del tercer caso de anáfora indirecta donde la información implícita en el antecedente no permite relacionar la anáfora y la información implícita en la anáfora no permite relacionar el antecedente, *es necesario relacionar entre si la información implícita de ambos para lograrlo*; además el antecedente es una expresión **verbal** y la anáfora es una expresión *nominal*. El antecedente **sepultado** ya forma parte del contexto lingüístico cuando en la segunda oración se menciona la anáfora *viuda*$_2$; al acceder a la información implícita de **sepultado** no se encuentra viuda (falla el proceso, marcado con la flecha 1); es necesario obtener la información implícita de *viuda*$_2$ para poder buscar el concepto del antecedente **sepultado** (proceso marcado con la flecha 2); después se busca el concepto del antecedente **sepultado** en la información implícita de *viuda*$_2$ (proceso marcado con la flecha 3, que también falla); se intenta relacionar la información implícita de ambos hasta encontrar conceptos similares; al tener éxito el proceso (marcado con la flecha 4) se puede terminar de relacionar *viuda*$_2$ con **sepultado** (marcado con la flecha 5). Se puede inferir que la *viuda*$_2$ estaba inconsolable por el hecho de "morir el cónyuge" (llamado **Juan**$_1$)y después de morir "fue sepultado".

¿Qué sucede si el proceso 4 también falla? En este caso, se está encontrando una referencia a nueva información o información adicional; en otras palabras, si no se presentan los fenómenos de correferencia y anáfora indirecta el fenómeno presente es el de referencia.

Resumiendo, se puede deducir que: los fenómenos de referencia y correferencia requieren un objeto del mundo real al que acceden a través de una entidad conceptual del discurso; los fenómenos de anáfora y anáfora indirecta requieren una relación a una entidad conceptual del discurso; ambos fenómenos pueden presentarse en forma directa (explícita) o en forma indirecta (implícita); la forma indirecta requiere un proceso de inferencia, de menor o mayor grado, para resolverla. Debido a que el mismo marcador (determinante) se utiliza para cada una de ellas, no es posible determinar si una expresión referencial requiere un proceso de inferencia para encontrar la relación de anáfora indirecta sin ANTES verificar la posibilidad de correferencia. En otras palabras, la coherencia textual, de acuerdo a el principio de relevancia, obliga a reconocer primero las correferencias directa e indirecta (del menor procesamiento) y SÓLO entonces intentar descubrir la relación de anáfora indirecta existente entre las expresiones nominales precedidas por un determinante.

Respondiendo a la segunda pregunta, ¿qué relación existe entre estos fenómenos?, la respuesta es *la secuencia u orden de resolución:* correferencia directa, correferencia indirecta, anáfora directa, anáfora indirecta, referencia directa, referencia indirecta. En base a esto, ha sido posible diseñar el modelo computacional que se describe a continuación.

4.3 Modelo computacional

La descripción del método se hará de lo general a lo particular, utilizando diagramas y comentarios, para profundizar paulatinamente en las consideraciones que, de una forma u otra, afectan su diseño y ejemplos de texto reales .

Para apoyar la descripción se utilizará la simbología de diagramas de flujo tomando en cuenta las siguientes variaciones:

- cada símbolo representa un proceso u operación local al diagrama, no necesariamente una instrucción equivalente de código

- Un número en la esquina superior izquierda del símbolo indica que este proceso requiere un desglose mayor por medio de un diagrama o figura adicional. El diagrama adicional se identificará con el comentario que se encuentra dentro del símbolo correspondiente

Por ejemplo en la figura 6 (Método General) se observa que sólo cuatro símbolos no están numerados (“Inicio”, “Fin”, “Hasta fin de archivo”, “¿fin de archivo?”) porque representan el ciclo de control general de este proceso para recorrer completamente el archivo. El símbolo marcado con **3,** “¿es correferente?”, indica que para poder obtener la respuesta a esta pregunta será necesario una explicación más detallada por medio de un proceso adicional.

El método general, figura 6, indica que el proceso se lleva a cabo recorriendo el archivo de principio a fin buscando expresiones nominales. Al encontrar una expresión nominal debe definirse un rango común de búsqueda hacia atrás para evaluar cada una de las funciones que puede cumplir dicha expresión en el texto. Después, establece claramente el orden de evaluación a seguir para determinar la función de una expresión nominal: es anáfora indirecta si y sólo si no es correferencia; es una nueva referencia si y sólo si no es correferencia ni anáfora indirecta. Finalmente se obtienen archivos de resultados de la corrida que permiten evaluar el funcionamiento del método.

En otras palabras, detectar primero la existencia de correferencia; la expresión nominal que no sea una correferencia es candidata a ser evaluada buscando la existencia de anáfora indirecta; si no es una anáfora indirecta la expresión es referencia a nuevas entidades extralingüísticas o del mundo real.

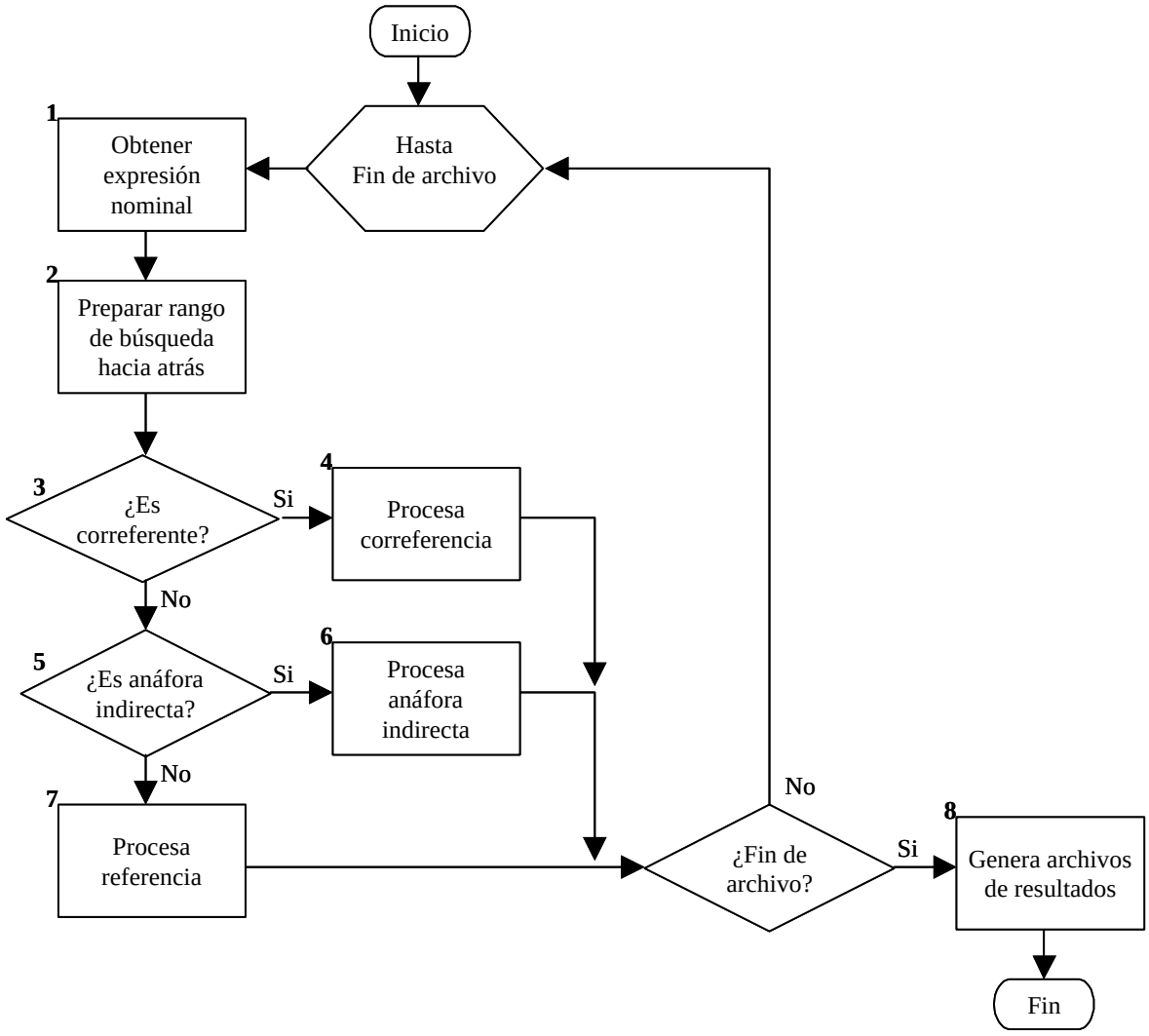


Figura 6 Método General

La observación, base del método, es que la expresión referencial “det + nombre común” se presenta comúnmente en los fenómenos de referencia, correferencia y anáfora indirecta, como se muestra en la tabla 3. Las primeras tres columnas de esta tabla se extraen de la tabla 2 excluyendo los pronombres intrínseco y extrínseco porque es son marcadores diferentes que dan origen a otro tipo de expresión nominal. Sin embargo, deberán tomarse en cuenta cuando sea necesario acoplar un método de resolución de la anáfora directa con éste, para desarrollar, por ejemplo, algún método de evaluación de la coherencia textual. Se le agregan columnas que muestran los requerimientos de solución del tipo de reglas, conocimiento común almacenado en diccionarios y proceso de inferencia.

La tabla 3 permite observar: que el *esfuerzo de procesamiento es menor en la correferencia* que en la anáfora indirecta, lo que es congruente con el uso más común de la correferencia para apoyar la coherencia textual; que el conocimiento (almacenado en diccionarios) y el *proceso de inferencia sólo es necesario en* la referencia, correferencia y anáfora del *tipo indirecta*; y que la *inferencia pragmática* (significación de conceptos y relaciones en función del contexto) *sólo* es necesaria para la solución de la *anáfora indirecta*.

			La solución requiere:		
Descripción	**Tipo**	**Expresión**	**reglas**	**diccionario**	**inferencia**
referencia	directa	nombre propio apelativo det + nombre común	gramática gramática gramática		
	indirecta	det + nombre común	gramática	sinónimos	semántica
correferencia	directa	nombre propio apelativo det + nombre común	gramática gramática gramática		
	indirecta	det + nombre común	gramática	sinónimos	semántica
anáfora	indirecta	det + nombre común	gramática semántica	sinónimos relaciones	semántica pragmática

Tabla 3 Requerimientos de solución

4.4 Obtención de la expresión nominal

La expresión nominal puede ser tan amplia como lo demande el proceso de comunicación como se observó en los ejemplos 51 y 51. El proceso que se muestra en la figura 7 considera los casos más comunes para identificar el núcleo (palabra) de la expresión nominal.

Primero detecta el inicio de la expresión nominal por medio de un determinante; a través del conector “a” va a la selección “¿Otra unidad léxica?”; así continua el proceso y espera encontrar como núcleo un nombre. En caso de encontrar un nombre propio se presupone la presencia de un determinante implícito, por lo que no es necesaria la marca de inicio para identificar el nombre propio como núcleo de la expresión nominal[Gómez, 98].

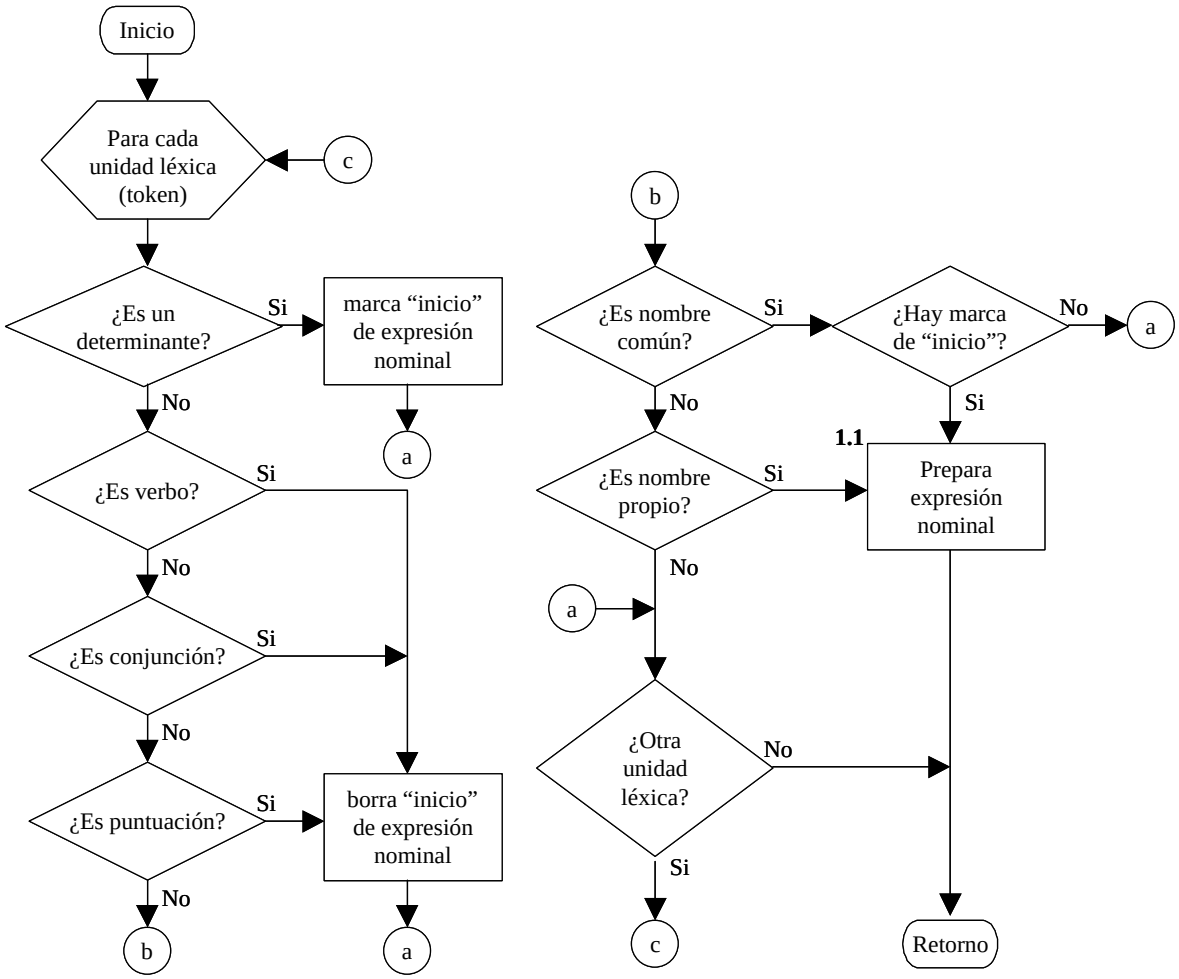


Figura 7 Proceso 1 Obtener expresión nominal

Así pues, la búsqueda del nombre después de encontrar un determinante debe terminar al encontrar un verbo, una conjunción o un signo de puntuación porque finalizan la expresión nominal referencial. Mención especial merecen las preposiciones contraídas *al* (a el) y *del* (de el) que funcionan en este caso como determinantes y así son consideradas.

(76) Veo su foto en *los* ***periódicos****: una* ***mujer*** joven, atractiva, probablemente zamba, esto es, mestiza de negra e india; oscura de color, en fin, como son oscuros *todos los* ***habitantes*** de *las* ***villas*** limeñas, arrabales de miseria en donde se hacinan *cientos* de *miles de* ***personas***.

En el ejemplo (76) y siguientes se muestra en *cursiva* cada determinante y en **negrita** cada nombre para explicar mejor el proceso de funcionamiento. Aplicando el proceso del diagrama al ejemplo (76) se obtienen las cinco expresiones nominales que se agrupan en la tabla 4.

N°	inicio	Expresión	fin
1	*los*	**periódicos**	**periódicos**
2	*una*	**mujer** joven	**mujer**
3	*todos*	los **habitantes**	**habitantes**
4	*las*	**villas** limeñas	**villas**
5	*cientos*	de miles de **personas**	**personas**

Tabla 4 Expresiones nominales del ejemplo 72

En las expresiones nominales obtenidas puede observarse que el inicio siempre es un determinante (1 al 5); el núcleo (nombre) puede ir precedido por más de un determinante o modificador (3 y 5); y que el fin de la expresión se alcanza al identificar el núcleo (1 al 5).

(77) Llegaba *un* **momento** en el que *todo* era bajar y bajar, caer y caer.

(78) ...se *la* quedó mirando fijamente.

El caso de marcar el *inicio* y **borrar el inicio al encontrar un verbo** se observa aplicando el proceso del diagrama a los ejemplos (77) y (78). En el primero trabaja localizando bien la frase nominal "*un* **momento**"; después encuentra el adverbio "*todo"* (homógrafo del determinante) y se marca el inicio; al encontrar el verbo "era" se borra el inicio; encontrando sólo una expresión nominal en esta oración. En el segundo encuentra el determinante "*la*" y se marca el inicio; al encontrar el verbo "quedó" se borra el inicio sin encontrar una expresión nominal.

(79) ..., sobreponiéndose *una* y *otra* **vez** a circunstancias dolorosas y extremas

El caso de marcar el *inicio* y **borrar el inicio al encontrar una conjunción** se observa aplicando el proceso del diagrama al ejemplo (79). Se encuentra el determinante "*una"* y se marca el inicio; al encontrar la conjunción "y" se borra el inicio; se encuentra el determinante "*otra*" y se marca el inicio; finalmente se encuentra el nombre "**vez**" y se logra obtener la expresión nominal "*otra* **vez**".

(80) Vendí *todas las* **libretas** y para mañana no queda ni *una*.

El caso de marcar el inicio y **borrar el inicio al encontrar un signo de puntuación** se observa aplicando el proceso del diagrama al ejemplo (80) Trabaja localizando bien la frase nominal "*todas las* **libretas**"; después encuentra el determinante "*una*" y se marca el inicio; al encontrar el signo de puntuación punto "." se borra el inicio; encontrando sólo una expresión nominal en esta oración.

Una vez localizada una expresión nominal es necesario almacenar su información y controlar su acceso como se muestra con el proceso en la figura 8.

Para mantener la información en memoria de cada expresión nominal (y cuando se requiera de la información "implícita") se utiliza una lista doblemente enlazada. Se necesita una lista doblemente enlazada porque crece con cada inserción al final, de la expresión nominal procesada, y para los procesos de verificación (¿es correferente? y ¿es anáfora indirecta?) se requiere búsqueda hacia atrás ("backtracking").

La verificación de error por falta de memoria es necesaria para poder leer archivos de textos largos. En caso de falta de memoria el espacio se recupera liberando memoria utilizada para almacenar la información "implícita", que fue leída de los archivos de sinónimos y relaciones para evaluar expresiones nominales anteriores a la actual. En otras palabras, se debe manejar una ventana de expresiones nominales, como contexto lingüístico, de tamaño o rango suficiente para la evaluación de los fenómenos de correferencia y anáfora indirecta. Sin el proceso de recuperación de memoria el programa queda limitado a alrededor de 4800 unidades léxicas (tokens) equivalente a 45 KB aprox.

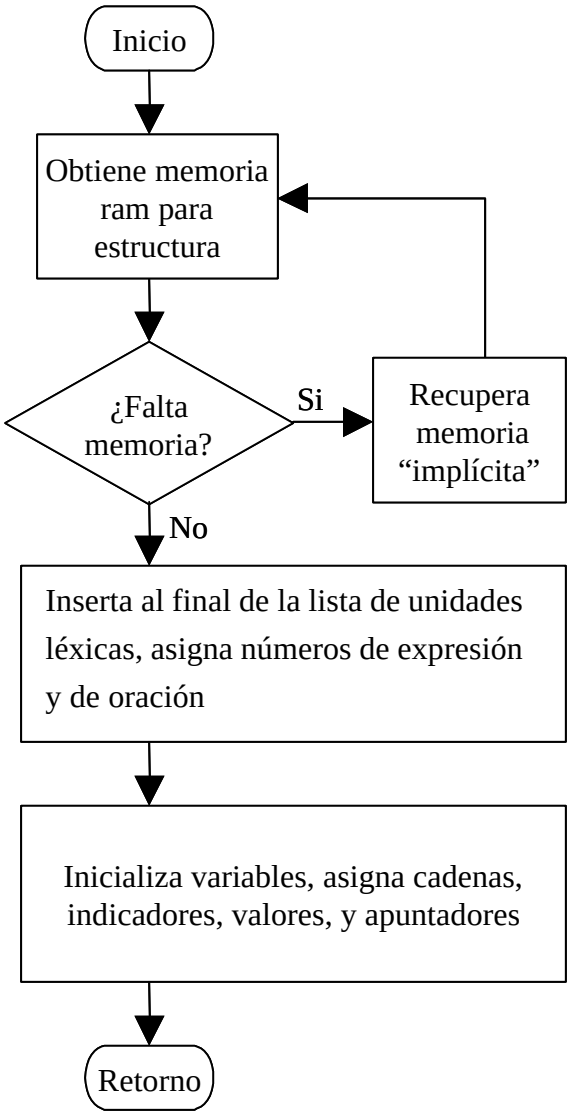


Figura 8 Proceso 1.1 Prepara expresión nominal

4.5 Preparación del rango de búsqueda

Esta operación, mostrada en la figura 9, es independiente y previa a la evaluación de los tres fenómenos porque cada evaluación se hace comparando el mismo número de núcleos de expresiones nominales. Primero se localiza el inicio de la oración actual porque se considera que los fenómenos, de correferencia y anáfora indirecta, se dan entre expresiones nominales de diferentes oraciones. El diseño permite establecer un marcador o bandera y un límite del número de ocurrencias en la búsqueda hacia atrás; esto facilita el aumentar o disminuir la ventana del rango de búsqueda y cambiar la bandera. La verificación de "¿inicio

de lista?” es una protección que opera para las primeras oraciones del texto, sólo cuando es menor el número de ocurrencias de la bandera seleccionada.

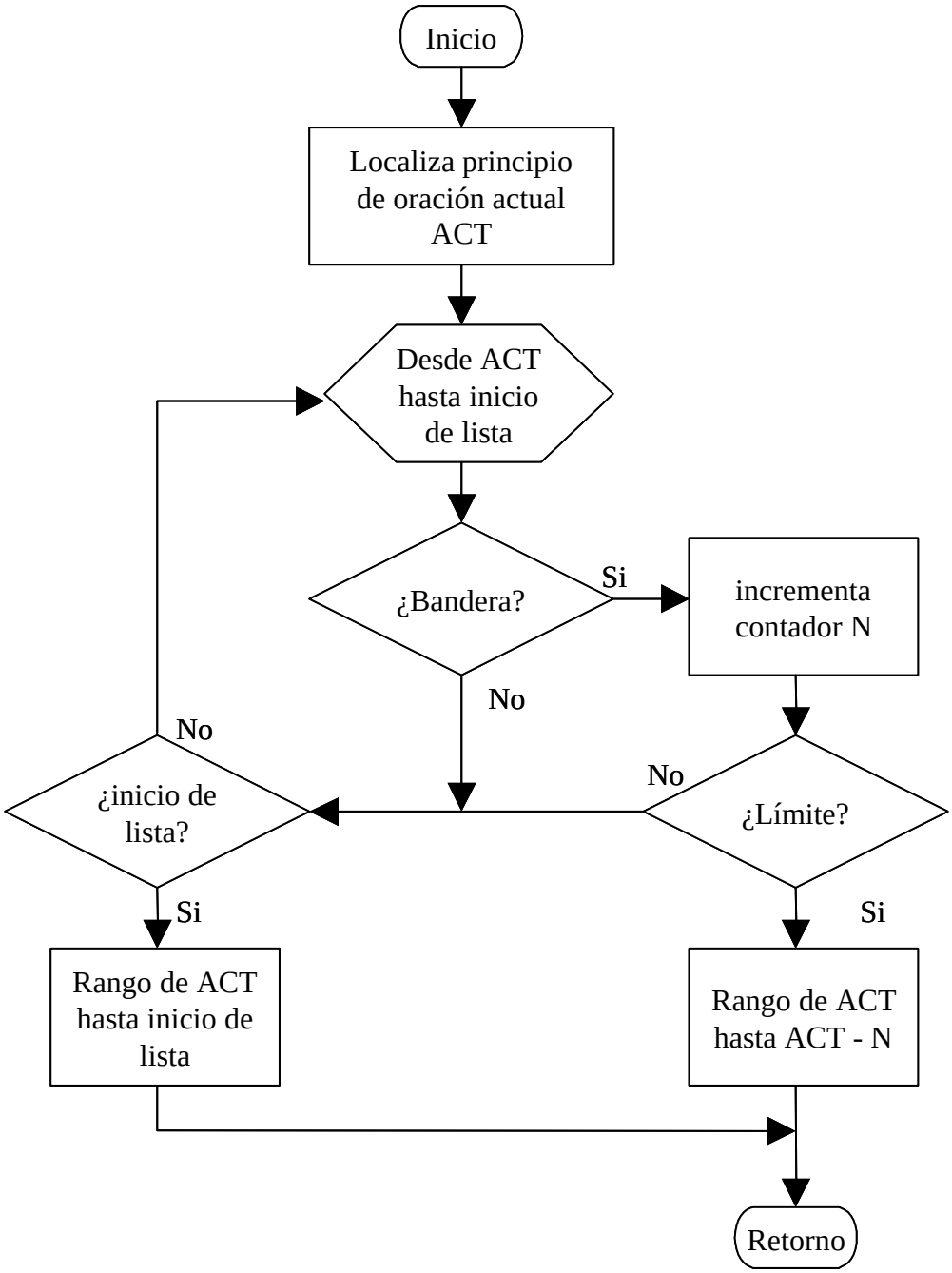


Figura 9 Proceso 2 Preparar rango de búsqueda hacia atrás

4.6 Detección y contextualización de correferencia

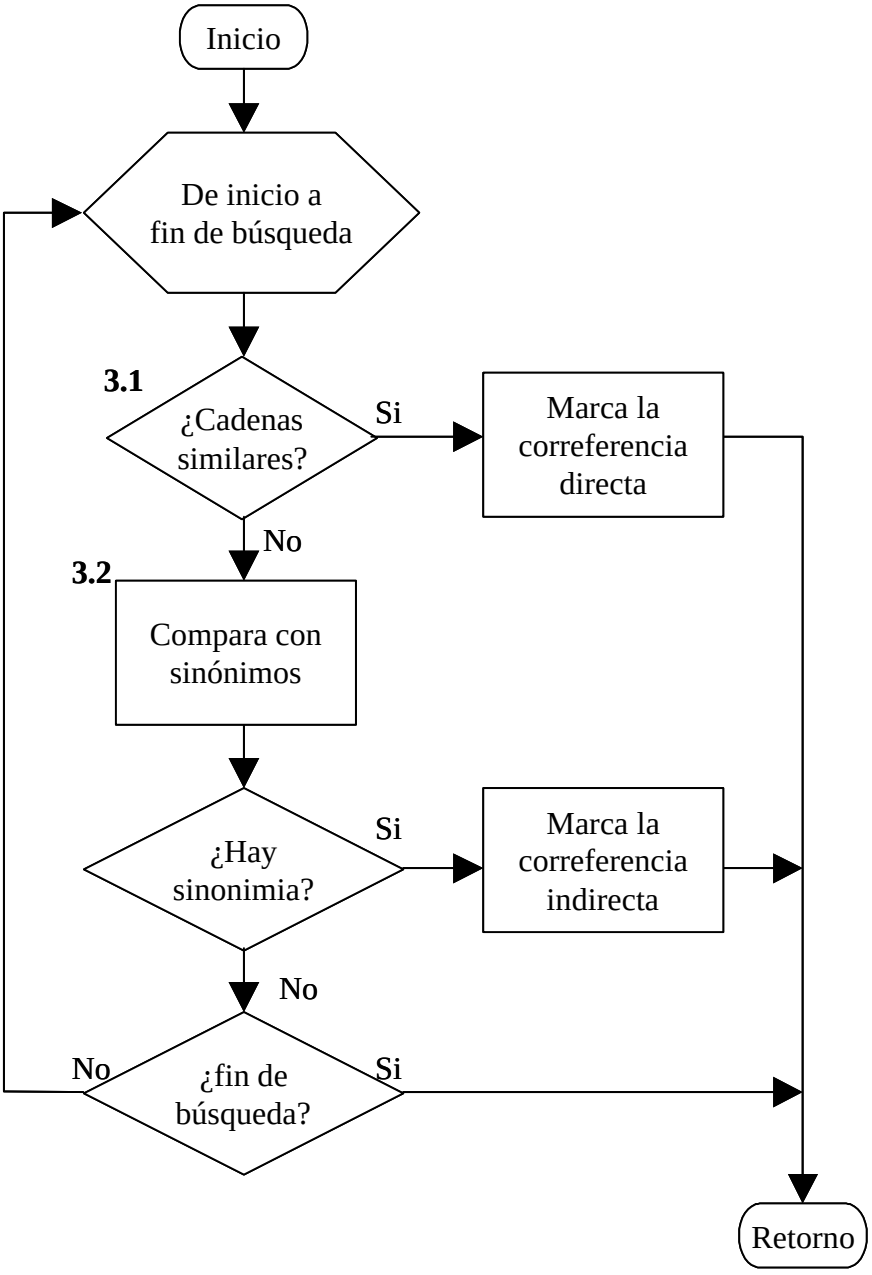


Figura 10 Proceso 3 ¿Es correferente?

En la figura 10 se observa una búsqueda, por todo el rango, para detectar los dos tipos de correferencia: directa e indirecta; marcándolos con un número clave en la estructura del núcleo de la expresión nominal. La detección de correferencia directa se lleva a cabo por comparación de cadenas, como se comentará en la figura 11, y la indirecta utilizando un diccionario de sinónimos, como se comentará en la figura 12.

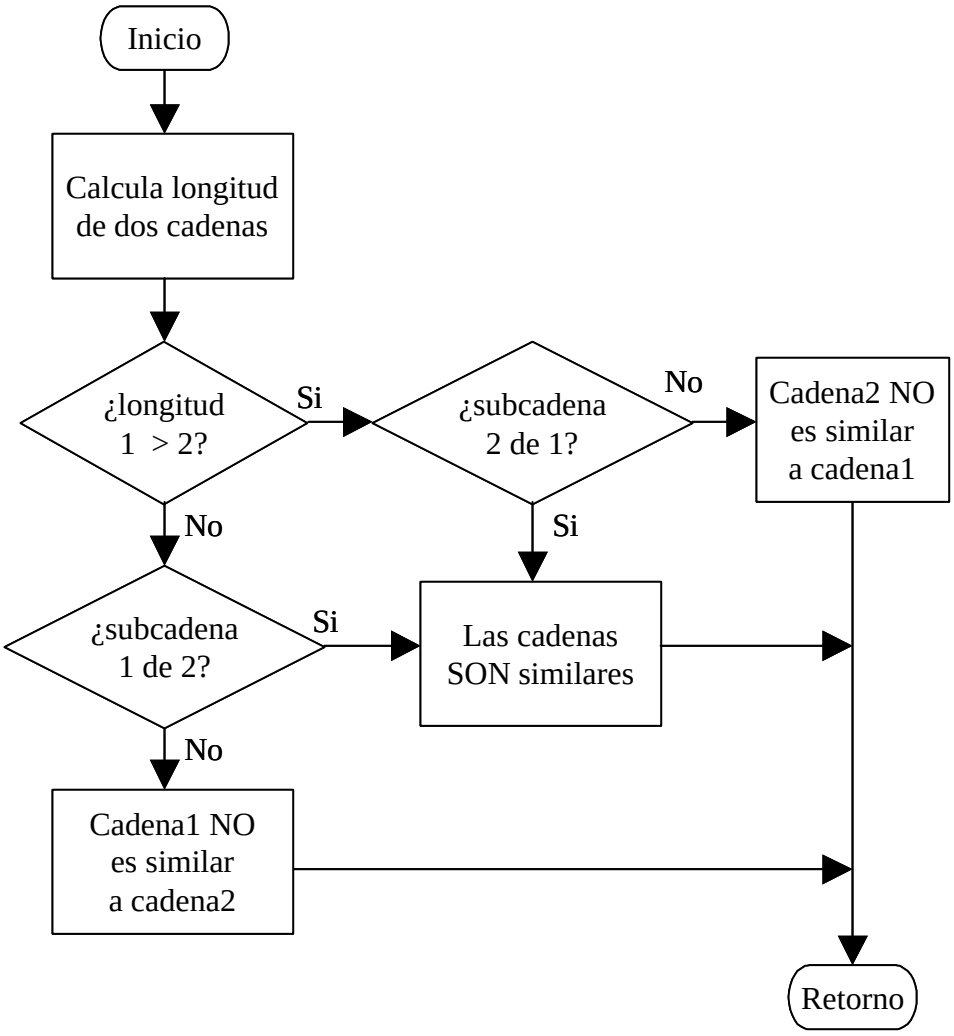


Figura 11 Proceso 3.1 ¿Cadenas similares?

En esta función se hace una comparación aproximada de cadenas por dos razones: la primera es que la comparación de nombres propios implica considerar como iguales María_Elena_Moyano, María_Elena y Moyano (ver ejemplo (76)); la segunda es que en los diccionarios de sinónimos y relaciones se almacena el nombre en su forma masculino singular; por lo tanto, es necesario adecuar la forma léxica para poder comparar por ejemplo a *niño, niña, niñas y niños* con *niño*, como se almacena en el diccionario.

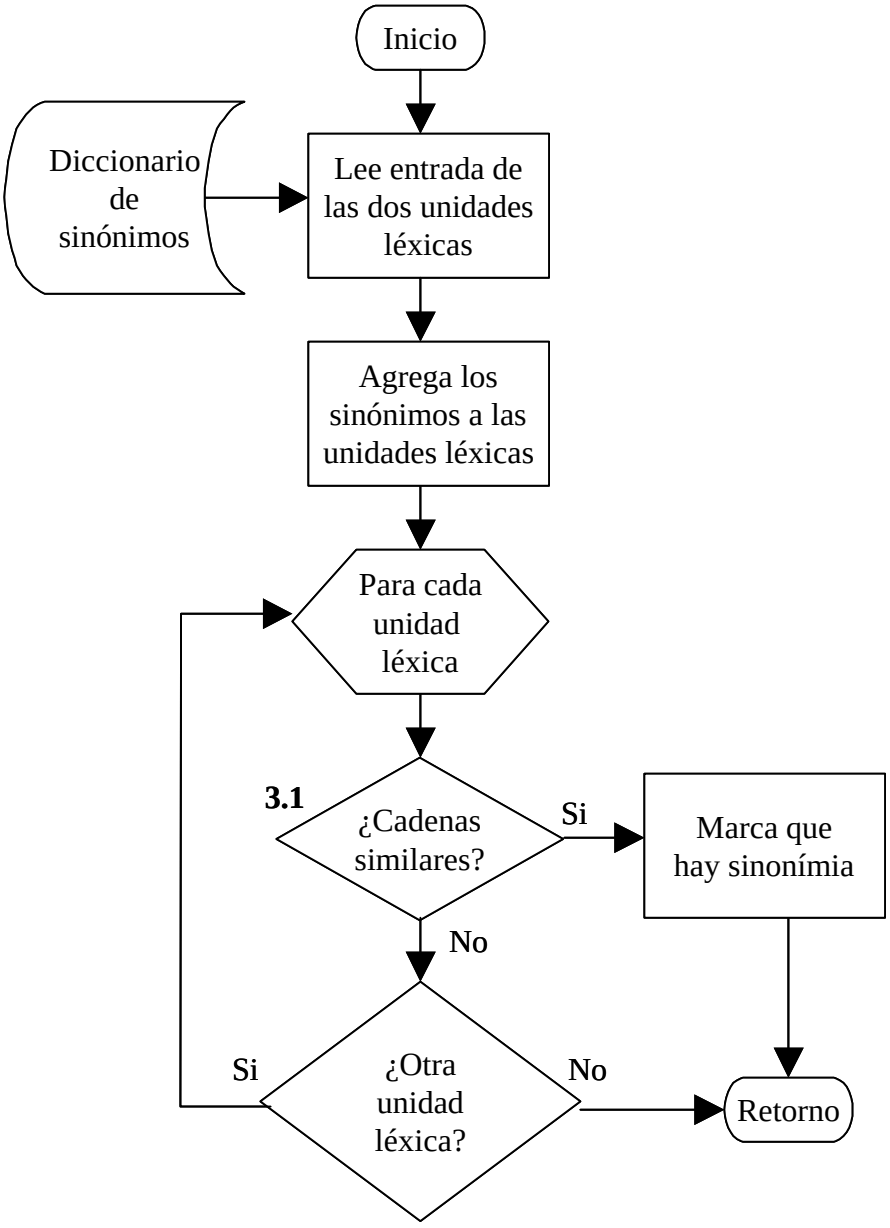


Figura 12 Proceso 3.2 Compara con sinónimos

La figura 12 muestra el proceso de comparación de sinónimos; primero la lectura del diccionario se utiliza para enriquecer el contexto lingüístico agregando la información a la estructura de los núcleos de las dos expresiones nominales. Después se comparan TODOS los sinónimos de una unidad léxica con TODOS los sinónimos de la otra. En otras palabras, en el ciclo se desarrolla una interjección de conjuntos de acepciones de las dos unidades a comparar. La comparación utiliza la misma función “¿cadenas similares?” de la figura 11.

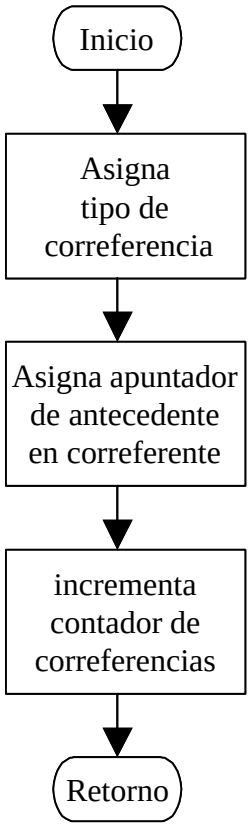


Figura 13 Proceso 4 Procesa correferencia

El proceso de correferencia, mostrado en la figura 13 tiene como función marcar el tipo de correferencia, manejar la asignación de apuntadores en la lista y actualizar estadística de ocurrencias del fenómeno.

(81) Imaginen un pueblo de **chabolas** de varios cientos de miles de personas; caminas y caminas por los mugrientos arenales y la miseria resulta inacabable, inabarcable.

(82) Era media mañana y había bastante gente en las calles, esto es, en las veredas sin urbanizar que habían quedado abiertas entre las *chozas*.

Los ejemplos (81) y (82) son dos líneas de texto, del archivo a14 de CLiC-TALP, que permiten ilustrar el proceso de correferencia. En cada unidad léxica se hace primero una prueba de correspondencia directa (igualdad de forma); después se obtienen del archivo todos los sinónimos de cada forma a comparar y la interjección entre estos dos conjuntos dará como resultado la existencia o no de correferencia indirecta. A continuación se muestra el ejemplo

de impresión cuando el programa encuentra una posible relación de sinonímia (los números a la derecha se añadieron para facilitar la explicación):

Busca referidos por *chozas*	(1)
choza barraca cabaña chamizo	(2)
chabola barraca *choza* chamizo	(3)
posible correferencia entre *chozas* y *chabolas*	(4)

Como puede observarse la entrada *chozas* que se busca (mostrada en la línea 2) existe como sinónimo en la entrada **chabola** (mostrada en la línea 3) que se compara de acuerdo al diccionario de sinónimos utilizados como fuente del conocimiento. Puede darse el caso de que la entrada no exista entre los sinónimos del segundo conjunto pero basta con que exista entre ambos un elemento común para que haya esta relación de sinonímia. Por ejemplo, en el caso anterior se observa que a través de “barraca” o de “chamizo” se puede obtener el mismo resultado porque aparece como sinónimo en ambas entradas del diccionario. La relación sinonímica permite detectar la correferencia indirecta, si no hay correferencia entonces la unidad léxica es candidata a anáfora indirecta.

El diseño desarrollado que facilita establecer una bandera diferente y un límite del número de ocurrencias en la búsqueda hacia atrás, para aumentar o disminuir la ventana del rango de búsqueda; permitió hacer diferentes corridas y poder observar hasta encontrar la bandera y límite más apropiados para este trabajo.

4.7 Detección y contextualización de anáfora indirecta

En la figura 14 se observa una búsqueda, por todo el rango, para detectar la anáfora indirecta marcándola con un número clave en la estructura del núcleo de la expresión nominal. La detección se lleva a cabo utilizando un diccionario de escenarios. Se le denomina diccionario de escenarios porque almacena como entrada nombres y como información los nombres que puedan tener algún tipo de relación, con la entrada, de holonimia, meronimia o rol.

Una vez identificada la unidad léxica se busca en el diccionario de escenarios y se determina si existe o no anáfora indirecta. Es importante recalcar la dependencia vital del

algoritmo con la información del diccionario, sin información suficiente el algoritmo falla, pero con la información necesaria se puede esperar un alto grado de precisión.

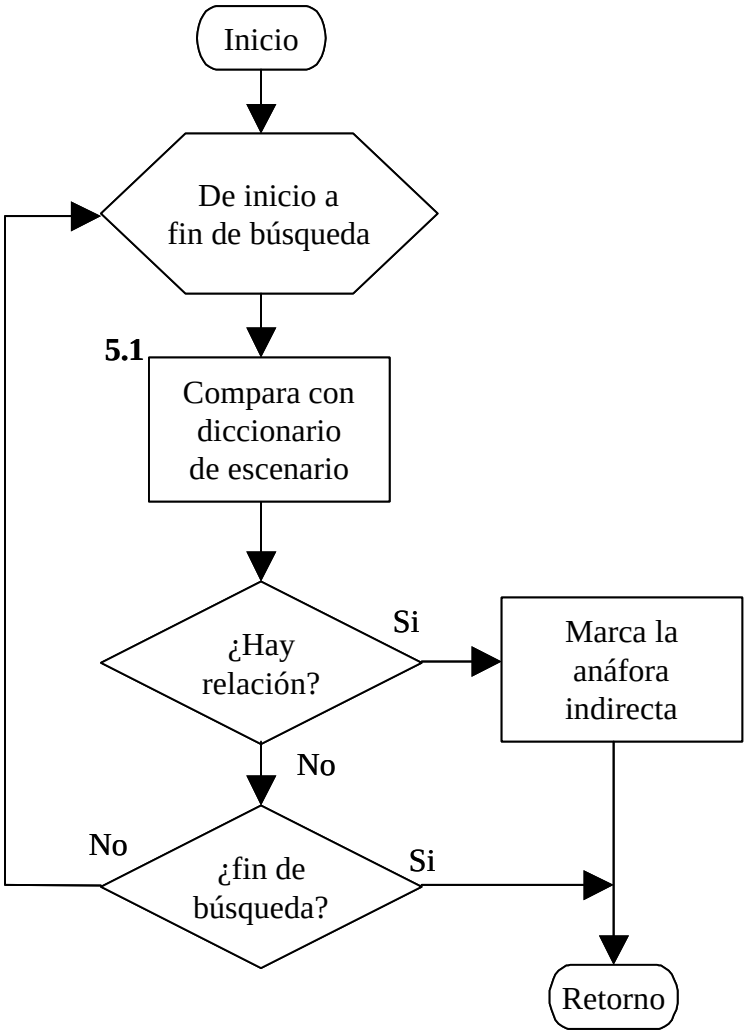


Figura 14 Proceso 5 ¿Es anáfora indirecta?

La figura 15 muestra el proceso de comparación con el diccionario de escenarios; primero la lectura del diccionario se utiliza para enriquecer el contexto lingüístico agregando la información a la estructura de los núcleos de las dos expresiones nominales. Después se comparan TODOS los nombres que tienen relación con una unidad léxica contra TODOS los nombres que tienen relación con la otra. En otras palabras, en el ciclo se desarrolla una interjección de conjuntos de acepciones de las dos unidades a comparar. La comparación utiliza la misma función “¿cadenas similares?” de la figura 11.

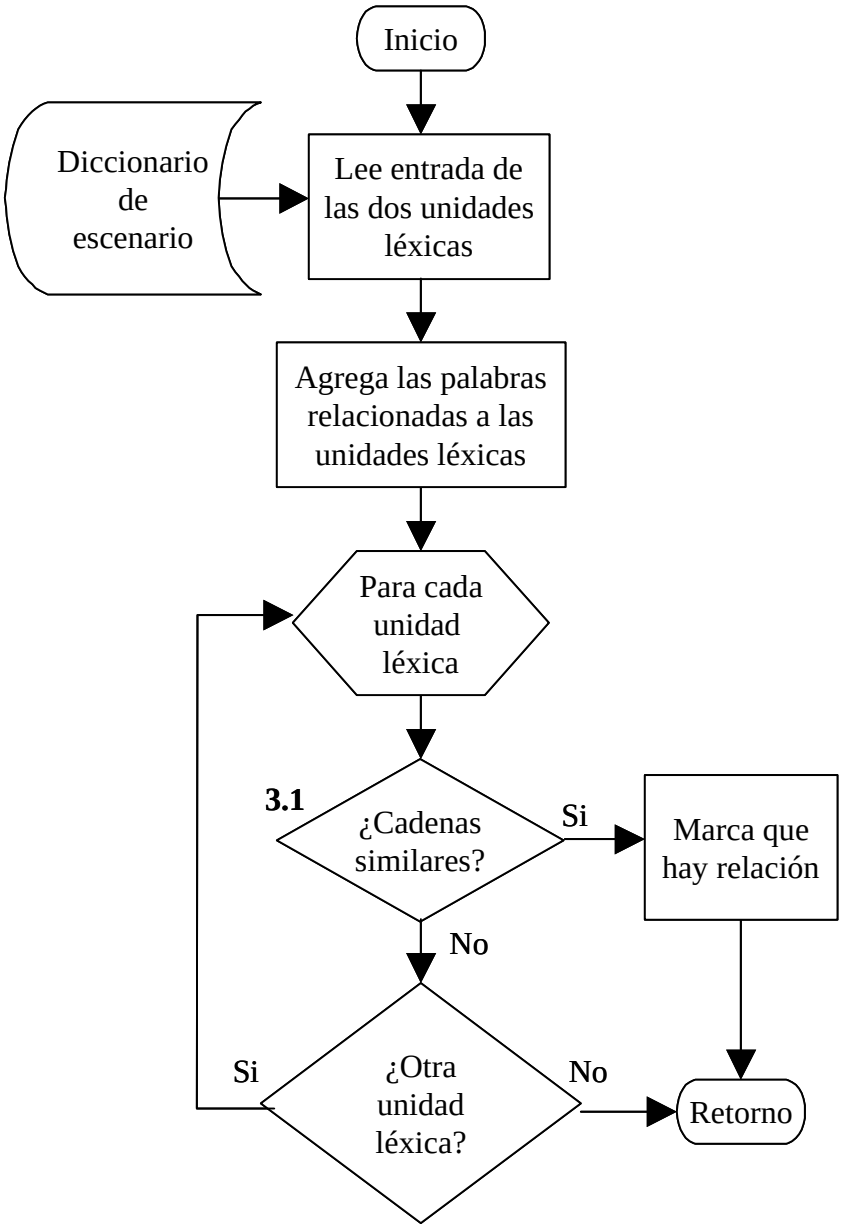


Figura 15 Proceso 5.1 Compara con diccionario de escenarios

El proceso de la anáfora indirecta, mostrado en la figura 16 tiene como función marcar el tipo de anáfora indirecta, manejar la asignación de apuntadores en la lista y actualizar estadística de ocurrencias del fenómeno.

Si con el conocimiento existente en el diccionario de escenarios no es posible detectar la anáfora indirecta se puede esperar con alta probabilidad que la entidad nominal sea una referencia (nueva información) y es necesario agregarla al contexto lingüístico del documento que se está analizando.

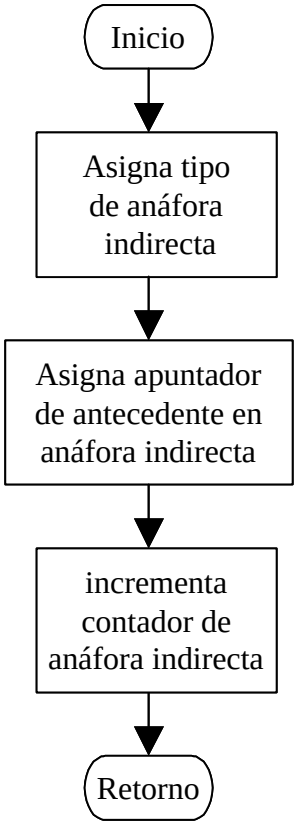


Figura 16 Proceso 6 Procesa anáfora indirecta

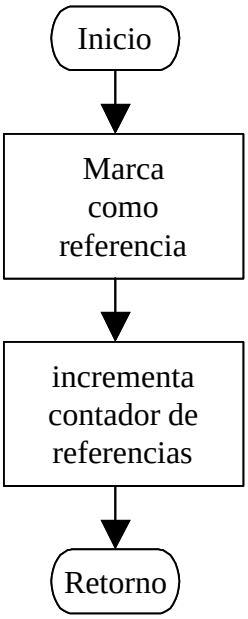


Figura 17 Proceso 7 Procesa referencia

4.8 Contextualización de referencia

El proceso de referencia, mostrado en la figura 17, tiene como función marcar el tipo de referencia y actualizar estadística de ocurrencias del fenómeno. Este proceso prepara el núcleo de la expresión nominal para ser antecedente posible de una expresión nominal posterior a ella en el texto.

4.9 Generación de resultados

El proceso que genera los archivos de resultados, mostrado en la figura 18, tiene como función obtener la información necesaria para evaluar los resultados de la corrida del programa. De acuerdo a esto, la necesidad puede variar y será necesario obtener un solo archivo o la combinación de varios; ésta es la razón de combinar en un argumento “imprime” de entrada al programa, las indicaciones de la información requerida.

Este proceso se documenta al final del método para dar claridad al diagrama aunque en realidad es un proceso inmerso en todo el método. Al inicio del programa se abren los archivos requeridos; las decisiones para grabar la información se van haciendo dentro del código de programación de acuerdo con la situación; y al final se cierran los archivos que se hayan abierto.

En la tabla 5 se muestran ejemplos de valores para el argumento imprime de acuerdo a los archivos deseados, indicando con un 0 el bit desactivado y con un 1 el bit activado o deseado. El valor 3 mostrado en la columna de “**imprime**” es típico para el desarrollo del programa verificando su paso por las funciones y resultados paulatinos de selección; el valor 27 se utiliza para un seguimiento exhaustivo; y el valor por omisión 4 se utiliza cuando el sistema está terminado y sólo se desean observar sus resultados en el archivo tipo htm.

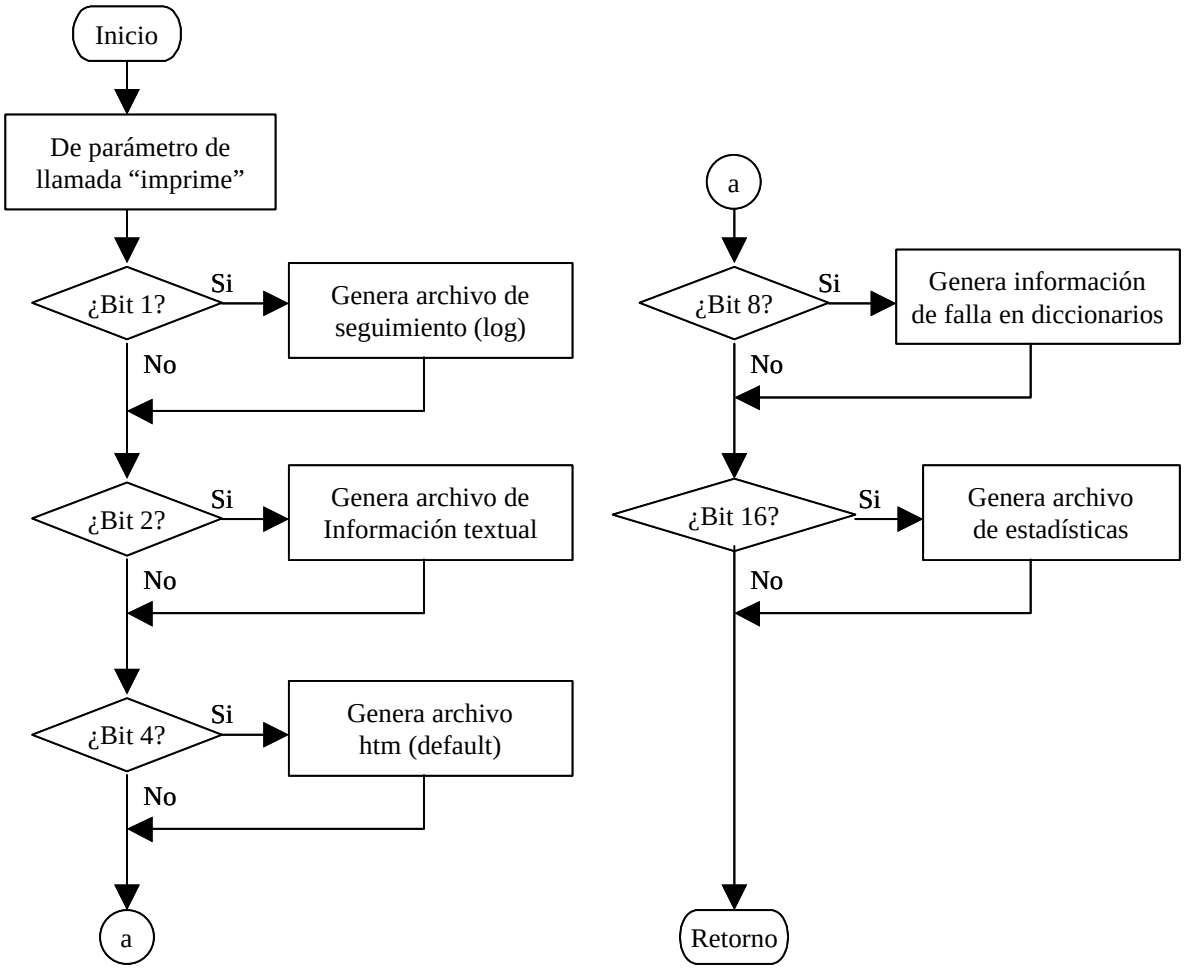


Figura 18 Proceso 8 Genera archivos de resultados

imprime	bit 16 estadística	bit 8 fallas	bit 4 htm	bit 2 texto	bit 1 log
1	0	0	0	0	1
2	0	0	0	1	0
3	0	0	0	1	1
4 (default)	0	0	1	0	0
11	0	1	0	1	1
19	1	0	0	1	1
26	1	1	0	1	0
27	1	1	0	1	1

Tabla 5 Ejemplos de combinaciones de resultados en parámetro “imprime”

5 DESARROLLO DE DATOS LINGÜÍSTICOS

5.1 Introducción

En este capítulo se describe el proceso que se siguió al seleccionar los recursos que permitieron implementar el sistema para poder evaluar el modelo desarrollado. Primero, se describen las razones para seleccionar el corpus más adecuado y disponible para correr las pruebas. La decisión se tomó después del análisis, apoyado en programas específicos de verificación y el análisis visual y manual necesario. En segundo lugar, ante la necesidad de probarlo con texto libre, se comenta la inclusión: a) del etiquetador TnT, al cual se le entrenó para el Español; b) del diccionario de sinónimos del Laboratorio de Lenguaje Natural del CIC-IPN que fue necesario corregir y validar manualmente; c) del diccionario de escenarios construido obteniendo la información de relaciones del diccionario semántico EuroWordNet en Español. En el capítulo siguiente se presentarán los resultados obtenidos.

5.2 Selección del corpus a utilizar

Para la obtención del corpus lingüístico se tenían inicialmente las siguientes alternativas:

a) Lecturas de Libros de Texto Gratuitos en Español obtenidos de la página Web de la Secretaría de Educación Pública en México. Se esperaba como ventajas la seguridad (relativa) de ser textos bien redactados (sin faltas de ortografía) y revisados, con buen nivel de lenguaje; se tendría que verificar el permiso para utilizarlo en el trabajo de investigación (Copyright). Después de un análisis de los textos se observó: una redacción y vocabulario demasiado elemental; NULA presencia del fenómeno de anáfora hasta el cuarto año de primaria; a partir del quinto año de primaria uso de anáfora directa por medio de pronombres (muy

elemental); casi nula presencia del fenómeno de anáfora indirecta (menor al 1 %). Se descartó, de acuerdo a lo anterior, por no ser representativo del fenómeno a analizar.

b) LexEsp (Léxico informatizado del español) – es un corpus en Español etiquetado con alrededor de 5 millones de palabras en texto libre, compilado por: el Departamento de Sicología de la Universidad de Oviedo, el Grupo Lingüístico de la Universidad de Barcelona y el Grupo para el tratamiento del Lenguaje de la Universidad Politécnica de Cataluña, recogido entre los años 1978 y 1995. Tiene como ventajas el poder reportar resultados con un corpus que ha sido utilizado por otros investigadores [Muñoz, 2000]; al ser etiquetado facilita el desarrollo del prototipo inicial; otra ventaja es que se tiene el permiso para utilizarlo en el trabajo de investigación (Copyright). Se tomó una muestra al azar representativa de 15 archivos y desarrollaron programas para efectuar la revisión observando: errores de captura del texto (palabras mal escritas, puntuación, etc); el etiquetado fue realizado con un etiquetador automático con errores ante palabras ambiguas; incongruencia entre el manual y los archivos en las claves de etiquetado; y la presencia de hasta 14 temas diferentes en cada archivo sin separación marcada. Se descartó por los errores mencionados y además porque para el proyecto es deseable que cada archivo contenga un solo artículo o tema debido a la importancia del contexto lingüístico.

c) Corpus en español que se ha ido compilando en el Laboratorio de Lenguaje Natural del CIC-IPN de noticias en periódicos publicadas en el Web, en texto libre. Tiene como ventaja el poder reportar resultados con un corpus que ha sido utilizado por otros investigadores [Galicia-Haro et al., 1999]. Su desventaja es que no es un texto libre de errores (ortográficos y de redacción); la presencia de temas diferentes en cada archivo sin separación marcada; y no está etiquetado. Se descartó por los errores mencionados y además porque para el proyecto es deseable que cada archivo contenga un solo artículo o tema debido a la importancia del contexto lingüístico.

Ante la situación mostrada se localizó un corpus etiquetado “Corpus CLiC-TALP que se planea esté formado por 1,000,000 de palabras etiquetadas, desambiguadas y validadas manualmente”, con fecha de la ultima actualización en Enero del 2002. Los propietarios del corpus son la Universidad Politécnica de Cataluña y la Universidad de Barcelona en España que ofrecen la cesión gratuita de derechos a investigadores. Se contactó a los responsables y obtuvo una muestra representativa del corpus validada manualmente (el proyecto está en proceso).

El corpus CLiC–TALP proviene de dos fuentes diferentes. Por una parte recoge una muestra representativa (de 500.000 palabras) de un corpus de prensa de 7 millones de palabras cedido por el periódico *La Vanguardia*. Por otra, recoge una muestra (también de 500.000 palabras) del corpus LexEsp, que es un corpus de 5 millones de palabras, representativo del español estándar escrito porque presenta varios estilos narrativos, procedentes de distintas fuentes (literatura, prensa, etc.) e incluye también muestras tanto del español peninsular como del de América. Recoge un número reducido de palabras por obra y no más de tres obras por autor. Las fuentes son las que aparecen en la tabla 6.

Fuentes	**Porcentaje**
Narrativa	40
Divulgación científica	10
Ensayo	10
Prensa	25
Semanarios	10
Prensa deportiva	5

Tabla 6 Fuentes de LexEsp

Recoge muestras de 329 novelas con unas 6.000 palabras por obra, aproximadamente. Las revistas de divulgación científica utilizadas han sido *Muy interesante*, *Mundo científico* e *Investigación y ciencia*, así como algunos artículos de Divulgación publicados en suplementos de periódicos como *El País* y *ABC*. Los fragmentos de ensayo provienen de unas 88 obras, a razón de unas 5.700 palabras por obra. La parte procedente de prensa se ha obtenido de *El País, ABC, El Mundo, El Periódico, Diario 16, El Independiente* y *La Vanguardia*. Hay que reseñar que esta parte se compone de otras tres: editoriales (15%), articulistas (50%) y noticias (35%). Los semanarios utilizados han sido *Cambio 16, Interviu, Época* y *Tiempo*. Por último

la parte de la prensa deportiva proviene de las publicaciones *As, Marca* y *Mundo Deportivo.* La parte del corpus CLiC-TALP extraída de LexEsp aparece en diferentes archivos cuyos nombres contienen una letra inicial seguida de un número. La letra se corresponde con las distintas fuentes utilizadas, de modo que es posible conocer el tipo de texto. En la tabla 7 se presenta la relación entre el nombre del archivo y su contenido.

Letra Inicial	Contenido
a	Articulistas
e	Ensayo
d	prensa deportiva
dc	Divulgación científica
c	suplementos de ciencia
ed	Editoriales
n	Noticias
r	Semanarios
t	Narrativa

Tabla 7 Contenido de la parte de LexEsP

Se desarrollaron programas de verificación y después de analizarlo se encontraron errores de captura (menor al 2%) y errores de documentación; estos errores se notificaron a Montserrat Civit (responsable y contacto) quien corrigió el manual agradeciendo las observaciones sugeridas. Se observó que en cada archivo había cuando mucho tres temas, pudiendo dividirse manualmente hasta obtener un solo artículo o tema. Por lo anterior se consideró más adecuado como alternativa, además de ofrecer la ventaja de evitar el preprocesamiento para obtener la información morfosintáctica de las expresiones lingüísticas para la prueba del prototipo inicial.

5.3 Adecuación y entrenamiento del etiquetador TnT

El etiquetador de partes de la oración es un módulo requerido como etapa de preprocesamiento cuando se necesita trabajar con texto libre, aunque no fue necesario para el prototipo inicial porque se trabajo directamente con el corpus etiquetado CLiC-TALP. El etiquetador es un programa que acepta texto libre o no preparado como entrada y añade a cada unidad léxica (token) una etiqueta que especifica sus propiedades gramaticales como

categoría, número, persona, etc. Para este trabajo se localizó y obtuvo uno de los mejores y más populares —TnT tagger— que había sido probado exhaustivamente para el Inglés y Alemán pero no para el Español.

El etiquetador TnT (TnT es el acrónimo de Trigrams'n'Tags) es de tipo estadístico que puede ser entrenado para diferentes lenguajes y conjuntos de etiquetas dependiendo de un corpus etiquetado. Su exactitud promedio está alrededor del 96.5% para palabras conocidas y del 87% para palabras desconocidas. Fue necesario realizar una evaluación con el fin de conocer que tan adecuado es para el Español y cuanto afectaría al rendimiento global del sistema obteniendo un rendimiento del 96.5%, igual al promedio, para palabras conocidas y del 79.8% para palabras desconocidas [Morales y Gelbukh, 2003].

La diferencia del 16.7% entre la precisión obtenida con palabras conocidas y desconocidas hizo necesario analizar la causa de los errores de etiquetado en el Español encontrando tres causas principales:

(**I**) El orden más libre de palabras en **la posición del adjetivo** por ejemplo:

(83) Juan compró un **nuevo** carro (otro carro, tal vez usado; posición poco usada)

(84) Juan compró un carro **nuevo** (un carro recién fabricado; posición más usada)

A continuación se presentan ejemplos de la corrida de TnT (resaltando la ocurrencia del error con **negrita**).

Ejemplo 1	
Etiquetado TnT	**Debe ser**
los DA0MP0	los DA0MP0
sórdidos **NCMP000**	sórdidos AQ0MP0
arenales AQ0CP0	arenales **NCMP000**

Ejemplo 2	
Etiquetado TnT	**Debe ser**
precarios **NCMP000**	precarios AQ0MP0
tenderetes AQ0CP0	tenderetes **NCMP000**
de SPS00	de SPS00
cartón NCMS000	cartón NCMS000

(**II**) **Ambigüedad por homografía** (formas iguales de palabras) con diferentes significados o funciones gramaticales, por ejemplo:

(85) Yo **bajo**$_1$ con el hombre **bajo**$_2$ a tocar el **bajo**$_3$ **bajo**$_4$ las escaleras (**bajo**$_1$ verbo, **bajo**$_2$ adjetivo, **bajo**$_3$ nombre de instrumento musical y **bajo**$_4$ adverbio de lugar)

(86) Deja el **sobre**$_1$ que **sobre**$_2$ **sobre**$_3$ la mesa (**sobre**$_1$ nombre, **sobre**$_2$ verbo y **sobre**$_3$ adverbio de lugar)

A continuación se presentan dos ejemplos de la corrida de TnT (resaltando la ocurrencia del error con **negrita)**.

Ejemplo 3	
Etiquetado TnT	**Debe ser**
y CC	y CC
recuerdo NCMS000	**recuerdo VMIP1S0**
aquel DD0MS0	aquel DD0MS0
almuerzo NCMS000	almuerzo NCMS000
conmovedor AQ0MS0	conmovedor AQ0MS0

En el ejemplo 3, la forma de palabra **recuerdo** puede ser el verbo recordar en tiempo pasado, primera persona del singular o un nombre que denota la memoria de un hecho del pasado.

Ejemplo 4	
Etiquetado TnT	**Debe ser**
se P0300000	se P0300000
sujeta VMIP3S0	sujeta VMIP3S0
la DA0FS0	la DA0FS0
cola NCFS000	cola NCFS000
con SPS00	con SPS00
la DA0FS0	la DA0FS0
boca NCFS000	boca NCFS000
y CC	y CC
rueda NCFS000	**rueda VMIP3S0**
como CS	como CS
una DI0FS0	una DI0FS0
rueda NCFS000	*rueda* NCFS000

En el ejemplo 4, hay dos ocurrencias de la forma de palabra **rueda** en la misma oración; en la primera ocurrencia es el verbo rodar en tiempo presente, tercera persona del singular, mal identificado por TnT; en la segunda ocurrencia es un nombre bien identificado por TnT.

(**III**) **Afijos o terminaciones iguales** de las palabras con diferente función

Ejemplo 5	
Etiquetado TnT	**Debe ser**
Comimos VMIS1P0	Comimos VMIS1P0
un DI0MS0	un DI0MS0
arroz AQ0CS0	**arroz NCMS000**
con SPS00	con SPS00
pollo NCMS000	pollo NCMS000

El error de TnT se debe a que el afijo –oz es común para nombres (arroz, voz) y adjetivos (atroz, feroz, portavoz).

Ejemplo 6	
Etiquetado TnT	**Debe ser**
la DA0FS0	la DA0FS0
nariz NCFS000	nariz NCFS000
en SPS00	en SPS00
forma NCFS000	forma NCFS000
de SPS00	de SPS00
taco NCMS000	taco NCMS000
de SPS00	de SPS00
billar **VMN0000**	billar **NCMS000**

El error de TnT se debe a que el afijo –ar denota el infinitivo de verbos de la primera conjugación pero también es común en nombres (billar, azúcar, telar) y adjetivos (espectacular, lumbar, estándar, molecular).

En el ejemplo 7 el afijo –ía se utiliza para el tiempo pasado simple del modo indicativo en la segunda conjugación para la primera y tercera persona del singular. En el ejemplo 8 el afijo –aba se utiliza para la primera y tercera persona del singular en el modo indicativo y tiempo pasado imperfecto de la primera conjugación. El error de TnT es en la identificación de

la persona (intercambio de primera y tercera persona). Estos ejemplos muestran la clase de ambigüedad más común manifestada en la conjugación de verbos en Español; el problema aumenta con el uso de cortesía o político del “ustedes” en lugar de “vosotros” cada vez más común en el Español moderno porque los afijos son iguales para la segunda y tercera persona del plural, como se puede observar en la tabla 8 donde se han concentrado los casos de ambigüedad.

Ejemplo 7	
Etiquetado TnT	**Debe ser**
Pero CC	Pero CC
la DA0FS0	*la DA0FS0*
existencia NCFS000	*existencia NCFS000*
de SPS00	de SPS00
dos DN0CP0	dos DN0CP0
recién RG	recién RG
nacidos AQ0MPP	nacidos AQ0MPP
en SPS00	en SPS00
la DA0FS0	la DA0FS0
misma DI0FS0	misma DI0FS0
caja NCFS000	caja NCFS000
sólo RG	sólo RG
pod**ía VMII1S0**	pod**ía VMII3S0**
deberse VMN0000	deberse VMN0000
a SPS00	a SPS00
un DI0MS0	un DI0MS0
descuido NCMS000	descuido NCMS000
de SPS00	de SPS00
fábrica NCFS000	fábrica NCFS000

Ejemplo 8	
Etiquetado TnT	**Debe ser**
Me PP**1**CS000	*Me* PP**1**CS000
obsesion**aba VMII3S0**	obsesion**aba VMII1S0**
la DA0FS0	la DA0FS0
imagen NCFS000	imagen NCFS000
del SPCMS	del SPCMS
pobre AQ0CS0	pobre AQ0CS0
Niño_Dios NP00000	Niño_Dios NP00000
rechazado AQ0MSP	rechazado AQ0MSP

				Afijos de los modelos de Conjugación		
Modo	**Tiempo**	**Persona**	**Número**	**1st**	**2nd**	**3rd**
Indicativo	presente	2	plural	-an	-en	-en
		3	plural	-an	-en	-en
	pasado simple	1	singular	-aba	-ía	-ía
		3	singular	-aba	-ía	-ía
		2	plural	-ban	-ían	-ían
		3	plural	-ban	-ían	-ían
	pasado indefinido	2	plural	-ron	-ieron	-ieron
		3	plural	-ron	-ieron	-ieron
	futuro simple	2	singular	-rán	-erán	-irán
		3	singular	-rán	-éran	-irán
	condicional simple	1	singular	-ría	-ería	-iría
		2	singular	-ría	-ería	-iría
		2	plural	-rían	-erían	-irían
		3	plural	-rían	-erían	-irían
Subjuntivo	presente	1	singular	-e	-a	-a
		3	singular	-e	-a	-a
		2	plural	-en	-an	-an
		3	plural	-en	-an	-an
	pasado simple	1	singular	-ara	-iera	-iera
		3	singular	-ara	-iera	-iera
		2	plural	-aran	-ieran	-ieran
		3	plural	-aran	-ieran	-ieran
	futuro simple	1	singular	-are	-iere	-iere
		3	singular	-are	-iere	-iere
		2	plural	-aren	-ieren	-ieren
		3	plural	-aren	-ieren	-ieren

Tabla 8 Ambigüedad en la conjugación de verbos

El error tipo (I) por posición del adjetivo, ver ejemplo (84), es poco usual por lo que se considera que no repercutirá significativamente en los resultados a obtener cuando se integre al sistema; el error tipo (II), ejemplos (85) y (86), es común a otras lenguas (Inglés por lo menos) y dependerá del entrenamiento ("casos conocidos") para que "aprenda" a discriminarlos; el error tipo(III), ejemplos de TnT 5 y 6, necesariamente depende del entrenamiento por lo que es necesario aumentar el vocabulario del etiquetador, número de archivos de entrenamiento, para solucionarlo; finalmente, el error tipo (III), ejemplos de TnT 7 y 8, no puede ser solucionado por el etiquetador en la situación actual y se requiere de una

etapa posterior de análisis apoyado en la concordancia con el sujeto de la oración para etiquetar correctamente el texto [Morales y Gelbukh, 2003].

Como se ha observado, los errores de etiquetado repercuten directamente la precisión del método con archivos de texto libre, lo que abre un área de oportunidad para mejorar el preprocesamiento de etiquetado y como consecuencia el método de resolución de la anáfora indirecta. Mención especial requiere el error tipo (III), de los ejemplos de TnT 7 y 8, que limita el uso del etiquetador para la resolución de la anáfora directa donde es indispensable la concordancia del sujeto con el verbo; sin embargo, no afecta al método de resolución de la anáfora indirecta desarrollado porque sólo utiliza la categoría gramatical.

5.4 Preparación del diccionario de sinónimos

El diccionario de sinónimos inicial se recibió del Laboratorio de Lenguaje Natural del CIC-IPN; este diccionario había sido obtenido con un scanner a partir del diccionario Océano de sinónimos. La revisión que se hizo mostró que estaba pendiente la verificación del proceso y corrección de errores. Ejemplos de los errores más comunes se presentan en la tabla 9 donde se observa que algunos errores habían quedado marcados con la letra **q** (sinónimos en fila 1, 2 y 4) y otros sólo podrían ser detectados por verificación visual directa (entrada en fila 3). Se puede observar también que hay una entrada para el singular y otra para el plural (ÉLITE y ÉLITES en filas 1 y 2); que algunos sinónimos están formados por una expresión idiomática (BUEN**AqS**OCIEDAD en filas 1 y 2); que existen errores de detección (ÉMU**I**O en lugar de ÉMU**L**O en fila 3; PR**EqD**ILECTO en lugar de PR**ED**ILECTO; **qC**ONO en lugar de **ÍC**ONO en fila 4); que el diccionario utiliza mayúsculas mientras que el texto libre “normalmente” utiliza minúsculas y sólo en la primera palabra de una oración, nombres propios, abreviaturas o siglas utiliza mayúsculas.

Se hizo la verificación y corrección de errores, apoyado con programas y manualmente, para lograr un diccionario que: utilice minúsculas; considere la limitante del sistema de archivos de texto (en computadoras personales de 1022 caracteres) y en lo posible mantenga una sola entrada para nombres, masculino singular, e infinitivo para verbos; no contenga expresiones idiomáticas porque las comparaciones se hacen con una palabra núcleo

de la expresión nominal. Se alcanzó la corrección de los errores en la tabla 9 obteniendo los resultados mostrados en la tabla 10.

La corrección redujo el número de entradas, de palabras y de palabras / entrada en el diccionario; además, si se considera “idealmente” la consulta al diccionario de una palabra que se encuentre a la mitad tendremos también una reducción en el número de accesos lógicos, al archivo en disco, al reducir el número de entradas (17% aprox.), como se muestra (ver 15816 en la fila **Corregido** de la tabla 11).

N°	ENTRADA	SINÓNIMOS
1	ÉLITE	ÉLITES BUENAqSOCIEDAD CÍRCULO CÍRCULOS CREMA CREMAS
2	ÉLITES	ÉLITE BUENAqSOCIEDAD CÍRCULO CÍRCULOS CREMA CREMAS DISTINCIÓN DISTINCIONES ELEGANCIA LAqFLOR LOqMEJOR MUNDANERÍA SELECCIÓN SELECCIONES
3	ÉMUIO	ADVERSARIA ADVERSARIO ADVERSARIOS ANTAGONISTA ANTAGONISTAS COMBATIENTE
4	PREFERIDO	PREqDILECTO ELEGIDO PRIVILEGIADO FAVORITO PRIVADO PROTEGIDO qCONO

Tabla 9 Errores en diccionario de sinónimos

N°	ENTRADA	SINÓNIMOS
1 y 2	élite	círculo crema distinción elegancia selección
3	émulo	adversario antagonista competidor contrario rival opuesto contendiente contrincante
4	preferido	predilecto elegido privilegiado favorito privado protegido ícono

Tabla 10 Corrección de diccionario de sinónimos

	entrada	**palabra**	**acceso**	**Pal/Ent**
Inicial	38245	260986	19123	6.82
Corregido	31632	223028	15816	6.76
Final	31632	223028	586	6.76

Tabla 11 Modificaciones al diccionario de sinónimos

letra	entrada	palabra	acceso	Pal/Ent
a	**3858**	27575	1929	7.15
b	849	6182	425	7.28
c	3857	**27577**	1929	7.15
d	3163	21571	1582	6.82
e	3274	22833	1637	6.97
f	1036	7673	518	7.41
g	750	5609	375	**7.48**
h	662	4848	331	7.32
i	2192	14851	1096	6.78
j	270	2014	135	7.46
k	**6**	**31**	3	5.17
l	787	5806	394	7.38
m	1524	10858	762	7.12
n	346	2427	173	7.01
ñ	14	96	7	6.86
o	563	3972	282	7.06
p	2638	18519	1319	7.02
q	126	827	63	6.56
r	1863	13030	932	6.99
s	1544	10546	772	6.83
t	1195	8409	598	7.04
u	144	1015	72	7.05
v	697	4904	349	7.04
w	11	45	6	4.09
x	18	60	9	**3.33**
y	49	342	25	6.98
z	196	1408	98	7.18
Suma	31632	223028		
Mínimo	6	31	**3**	3.33
Máximo	3858	27577	**1929**	7.48
Promedio	1172	8260	**586**	6.76

Tabla 12 Análisis del diccionario de sinónimos

Sin embargo, el problema del acceso secuencial de archivos continúa presente, lo que provoca lentitud de procesamiento del programa. Para reducir el tiempo de acceso en la búsqueda secuencial se dividió el archivo de sinónimos en varios archivos, de acuerdo a la letra inicial de la palabra de entrada (incluyendo las acentuadas en caso de iniciar con vocal), obteniendo los valores mostrados en la tabla 12. Este análisis permitió visualizar una

reducción del tiempo de acceso al manejar de esta forma el diccionario porque: en el **peor** de los casos, de nombres que inician con la letra “a”, se tienen 1989 accesos lógicos contra 15816 sin la división del diccionario logrando reducir casi 8 veces el número de accesos (15816/1989 = 7.95; ver la fila **Máximo** de la tabla 12); en el **mejor** de los casos, letra “k”, con sólo 3 accesos lógicos al disco (de acuerdo al tamaño de “buffer” y memoria caché de los discos actuales, con sólo un acceso físico) se obtiene la información requerida (ver la fila **Mínimo** de la tabla 12); y el caso **promedio** con 585 accesos lógicos (ver la fila **Final** de la tabla 11 y la fila **Promedio** de la tabla 12). Con esta forma de manejar el diccionario se logro reducir el tiempo de procesamiento de más de 8 minutos por archivo a menos de 2 minutos por archivo.

5.5 Construcción de diccionario de escenarios

Para la construcción del diccionario de escenarios se tomó como base la información contenida en el diccionario semántico EuroWordNet LE2-4003 WP6.5. Esta versión del diccionario semántico, propiedad del Laboratorio de Lenguaje Natural del CIC-IPN, no se suministra con bibliotecas de funciones para acceder directamente la información (sólo se suministran para WordNet en Inglés en la versión 1.5). Para utilizarlo en Español fue necesario utilizar el archivo de exportación (Export file) en formato de texto ASCII; en la tabla 13 se muestra como ejemplo una entrada del archivo.

Primero se analizó la documentación y se seleccionaron las relaciones necesarias para relacionar núcleos de expresiones nominales (nombres); después se desarrollaron programas de acceso y extracción. El proceso de extracción se validó con un programa de prueba para contar las relaciones, permitiendo probar las rutinas de acceso, y se obtuvieron los resultados mostrados en la tabla 14; en esta tabla se hace la comparación contra las cantidades en la documentación de EuroWordNet observando las diferencias mostradas en la tercera columna.

```
0 @5106@ WORD_MEANING
  1 PART_OF_SPEECH "n"
  1 VARIANTS
    2 LITERAL "órgano"
      3 SENSE 4
    2 LITERAL "órgano_vegetal"
      3 SENSE 1
  1 INTERNAL_LINKS
    2 RELATION "has_hyperonym"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "cosa"
          5 SENSE 1
    2 RELATION "has_hyponym"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "retoño"
          5 SENSE 2
    2 RELATION "has_hyponym"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "follaje"
          5 SENSE 2
    2 RELATION "has_hyponym"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "cabillo"
          5 SENSE 1
    2 RELATION "has_hyponym"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "offset"
          5 SENSE 4
    2 RELATION "has_hyponym"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "raíz"
          5 SENSE 2
    2 RELATION "has_hyponym"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "lámina"
          5 SENSE 2
    2 RELATION "has_hyponym"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "estructura_reproductiva"
          5 SENSE 1
    2 RELATION "has_hyponym"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "ascocarpo"
          5 SENSE 1
    2 RELATION "has_hyponym"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "esporocarpo"
          5 SENSE 1
  1 EQ_LINKS
    2 EQ_RELATION "eq_synonym"
      3 TARGET_ILI
        4 PART_OF_SPEECH "n"
        4 WORDNET_OFFSET 7977350
    2 EQ_RELATION "eq_has_hyperonym"
      3 TARGET_ILI
        4 PART_OF_SPEECH "n"
        4 WORDNET_OFFSET 7976849
```

Tabla 13 Ejemplo de formato de WordNet en Español

RELACIÓN	Documentado	Contado	Diferencia
has_holo_madeof	110	108	2
has_holo_member	427	426	1
has_holo_part	1929	1923	6
has_hyperonym	24608	24507	101
has_hyponym	24608	24507	101
has_mero_madeof	110	108	2
has_mero_member	427	426	1
has_mero_part	1929	1923	6
TOTAL	**54148**	**53926**	**220**

Tabla 14 Relaciones obtenidas de WordNet en Español

Las diferencias observadas hicieron necesario verificar la posibilidad de error en los programas, rutinas o en la información del diccionario de EuroWordNet. El análisis condujo a encontrar diferentes tipos de errores en el diccionario que provocan estas diferencias (y corroboran que los programas y rutinas están trabajando bien); además, permitió observar errores adicionales que afectarán, de una forma u otra, la información obtenida para la resolución de la anáfora indirecta; ejemplos de estos errores enumerados en la primera columna, resaltados con **negrita** en ENTRADA y RELACIONES, se muestran en la tabla 15.

N°	ENTRADA	RELACIONES
1	añil	has_hyperonym añil
2	añil	has_hyperonym color
3	añil	has_hyperonym color
4	día_de_la_independencia	has_holo_part julio
5	día_de_la_bandera	has_holo_part junio
6	nuclio	has_hyperonym palabra
7	lugar	has_hyponym P
8	P	has_hyperonym lugar
9	Méjico	has_mero_member mejicano

Tabla 15 Ejemplo de errores de WordNet en Español

Un tipo de error **1** es encontrar alguna relación incorrecta, otro tipo de errores, **2** y **3**, es encontrar entradas duplicadas; en ambos casos el proceso rechaza como incorrecta la relación duplicada, durante la creación del diccionario de escenarios; estos tres tipos de errores son la causa de las diferencias (reducción en el número de relaciones en un 0.4%).

Los errores **4** y **5** son errores de contexto porque para el Inglés de Estados Unidos de Norteamérica ambas entradas están relacionadas con los meses de **julio** y **junio** pero para cualquier otro país la relación es falsa; por ejemplo para México deberían ser septiembre y febrero respectivamente (16 de septiembre y 24 de febrero).

El error tipo **6** representa errores de captura en la entrada provocando que esta relación se pierda porque en el diccionario existe **núcleo** que es lo correcto y no **nuclio** (mal escrito y sin acento); en otras palabras, el programa que busque las relaciones posibles para la entrada núcleo encontrará la entrada relacionada con "*órgano célula cromosoma cromatina centro átomo conjunto importancia mecanismo disco sumista cognición*" pero desconectada de la relación de hiponimia con "*palabra*" debido a este error de captura. Un caso diferente existe

con las entrada **médula** y **medula**; ambas acepciones son permitidas por el Diccionario de la Real Academia Española y los desarrolladores de EuroWordNet trataron acertadamente de mantener las variantes registradas pero no lo lograron porque falta esta duplicidad en las entradas sistema_nervioso_central y estructura_neurológica.

Los errores tipo **7** y **8** representan errores de adecuación al idioma porque esta relación no existe (al menos en México) ya que "lugar has_hyponym P" proviene del Inglés "P = Parking" y tiene su equivalente en Español con "E = estacionamiento".

El error tipo **9** se comete por no observar que el diccionario de la Real Academia Española admite la acepciones con "x" y con "j", México y Méjico respectivamente, por lo tanto, se deberían registrar y mantener ambas variantes pero no sucede así; México y todas sus variantes: mexicano, mexiquense, etc. no existen en el diccionario EuroWordNet. Un texto libre donde se registre la acepción con "x" fallará irremediablemente.

El análisis de los errores anteriores alerta sobre: el impacto en la precisión del sistema desarrollado; la necesidad de desarrollar recursos propios para el Español más confiables; y la necesaria participación de México en el desarrollo de sistemas que se promueven como el estándar "de facto" en el mundo.

El siguiente paso, para obtener el diccionario de escenarios, fue la extracción de las relaciones seleccionadas; se muestran los totales en la columna **Relaciones** de la tabla 18 y un ejemplo en la tabla 16. La entrada órgano tomada como ejemplo, muestra cuatro diferentes sentidos de la palabra de acuerdo al contexto utilizado: periodístico, botánico, musical, etc. Habiendo observado que el sistema de resolución de la anáfora indirecta requiere sólo "saber" que existe una relación para establecer el enlace se decidió reducir el tamaño de este diccionario con una estructura y forma de manejo similar a la del diccionario de sinónimos; lo anterior, permitió además utilizar las subrutinas ya desarrolladas.

El ejemplo de la tabla 16 muestra cuatro entradas con treinta y siete relaciones; después del proceso de reducción en la tabla 17 se muestra una sola entrada equivalente con las mismas treinta y siete relaciones; en el primer caso se requerían setenta y ocho cadenas para almacenar la información y en el segundo caso se requieren sólo cuarenta (una reducción del 49% sólo para esta entrada del diccionario).

ENTRADA	RELACIONES
órgano	has_hyperonym boletín
órgano	has_hyperonym cosa has_hyponym retoño has_hyponym follaje has_hyponym cabillo has_hyponym offset has_hyponym raíz has_hyponym lámina has_hyponym estructura_reproductiva has_hyponym ascocarpo has_hyponym esporocarpo
órgano	has_hyperonym instrumento_de_viento has_mero_part pedal has_mero_part teclado
órgano	has_hyperonym parte_del_cuerpo has_hyponym órgano_eréctil has_hyponym órganos_reproductores has_hyponym centriolo has_hyponym condriosoma has_hyponym nucléolo has_hyponym núcleo has_hyponym órgano_secretatorio has_hyponym cristalino has_hyponym órgano_del_habla has_hyponym lengua has_hyponym receptor has_hyponym víscera has_hyponym órgano_externo has_hyponym órgano_efector has_hyponym órgano_vital has_hyponym músculo has_hyponym ventosa has_hyponym patas has_hyponym oviscapto has_hyponym cilio has_mero_part lóbulo

Tabla 16 Ejemplo de entradas obtenidas de WordNet en Español

Resumiendo, la reducción del diccionario se apoyó en la selección de las relaciones necesarias para el diccionario de escenarios; los tamaños se muestran en la tabla 18 donde se observa que EuroWordNet contiene en total 72,508 entradas (de relaciones individuales palabra-palabra) en el formato mostrado en la tabla 13; la primera selección redujo el número

de entradas a 27,959 con el formato de la tabla 16; y finalmente se obtuvieron 853 entradas en el formato de la tabla 17.

ENTRADA	RELACIONES
órgano	boletín cosa retoño follaje cabillo offset raíz lámina ascocarpo estructura_reproductiva esporocarpo instrumento_de_viento pedal teclado parte_del_cuerpo órgano_eréctil órganos_reproductores centriolo órgano_secretatorio cristalino órgano_del_habla núcleo condriosoma nucléolo lengua receptor víscera órgano_externo órgano_efector órgano_vital músculo ventosa patas oviscapto cilio lóbulo

Tabla 17 Ejemplo de entradas del diccionario de escenarios

	TAMAÑO		
DESCRIPCIÓN	EuroWordNet	Relaciones	Final
Cadenas	217,488	91,917	3,210
Entradas	72,502	27,959	853
Tamaño (KB)	2,444	944	33

Tabla 18 Reducción del diccionario de escenarios

El proceso total de reducción, que se muestra en la tabla 18, permite observar una disminución drástica (alrededor del 98%) del almacenamiento necesario. Lo anterior se explica verificando el tamaño de cadenas de la tabla 16; las cadenas cosa, raíz, cilio, etc. (que permanecen en el diccionario) son cadenas de menor tamaño que has_hyponym (que se elimina); además de considerar la reducción del número de entradas en el diccionario al agrupar todas las relaciones en una sola entrada.

6 ANÁLISIS DE RESULTADOS

6.1 Introducción

En este capítulo se describe el procedimiento empleado para determinar la validación de las ideas expuestas anteriormente. Primero se describe el tipo de evaluación a efectuar y las métricas seleccionadas: precisión, especificidad y concordancia.

En segundo lugar, se presenta la metodología experimental para determinar el tamaño apropiado de la muestra que permite probar la hipótesis planteada con el prototipo inicial. En tercer lugar se describe el proceso y resultados obtenidos de las corridas del sistema utilizando archivos de texto etiquetado y diccionarios específicos (ad hoc) de sinónimos y escenarios; se presentan los resultados de las corridas en los archivos seleccionados y verificados manualmente. Después se presenta el análisis para determinar el tamaño apropiado para la ventana de búsqueda hacia atrás. Finalmente se comentan los resultados de los experimentos realizados con archivos de texto libre.

6.2 Métricas seleccionadas

La evaluación a realizar es intrínseca o de categorización porque juzga la calidad o efectividad del sistema en la asignación automática de la expresión nominal, a la categoría de correferencia, anáfora indirecta o referencia, comparando los resultados con la asignación o verificación manual de un "experto" humano. Para la medición de resultados se seleccionó como métricas primarias la ***precisión*** y *la* ***especificidad*** (del Inglés precision y recall respectivamente). Entendiendo como ***precisión*** la habilidad del sistema de identificar ***sólo*** *los elementos relevantes* (o pertenecientes a la categoría), y como ***especificidad*** la de identificar ***todos*** *los elementos relevantes* [Salton, 1989]. Sin embargo, estas métricas, por si solas, no permiten apreciar el funcionamiento total del sistema; es necesario pues utilizar una métrica

adicional, la concordancia (del Inglés agreement) que puede calcularse con base en los datos obtenidos. Entendiendo como ***concordancia*** la capacidad del sistema de identificar los elementos, tanto relevantes como no relevantes, "de acuerdo con el experto". Para explicar mejor estás métricas se utilizará una tabla para decisiones de clasificación que se muestra en la tabla 19.

		EXPERTO		
		Si	**No**	
SISTEMA	**Si**	a	b	k = a + b
	No	c	d	m = c + d
		r = a + c	s = b + d	n = a + b + c + d

Tabla 19 Contingencias para decisiones de clasificación

A continuación, de acuerdo a la tabla 19, se establecen las fórmulas para las tres métricas propuestas.

$$precisión = \frac{a}{k} \qquad especificidad = \frac{a}{r} \qquad concordancia = \frac{a+d}{n}$$

Donde:

a = la proporción de elementos **asignados** a la categoría por el sistema **y que son** miembros de esa categoría

b = la proporción de elementos **asignados** a la categoría por el sistema **y que no son** miembros de esa categoría

c = la proporción de elementos **no asignados** a la categoría por el sistema **y que son** miembros de esa categoría

d = la proporción de elementos **no asignados** a la categoría por el sistema **y que no son** miembros de esa categoría

k = suma de todos los elementos **asignados** a la categoría **por el sistema**

m = suma de todos los elementos **no asignados** a la categoría **por el sistema**

r = suma de todos los elementos **asignados** a la categoría **por el experto**

s = suma de todos los elementos **no asignados** a la categoría **por el experto**

n = suma de todos los elementos **considerados** en la evaluación

Algunos autores proponen métricas adicionales de acuerdo al efecto que se desea apreciar: por ejemplo digresión o identificación incorrecta de elementos de una categoría. Entendiendo como ***disgresión*** (del Inglés fallout) la discrepancia del sistema al identificar *elementos relevantes* falsos y cuya fórmula es:

$$disgresión = \frac{b}{s}$$

Para el trabajo desarrollado se consideran adecuadas las tres métricas inicialmente mencionadas y se calcularán las necesarias para apreciar mejor los resultados obtenidos.

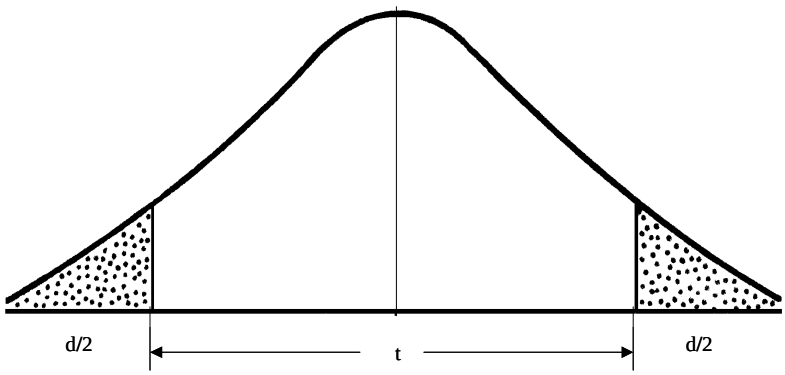
6.3 Tamaño de la muestra

Reconociendo como base: que un parámetro poblacional correcto sólo puede obtenerse por el estudio de toda la población; que aún en este caso la certeza del valor está sujeta al proceso e instrumentos de medición; que, acorde con lo anterior, el muestreo siempre lleva involucrado un error en la estimación del parámetro; que el muestreo es necesario en la experimentación por la imposibilidad económica, física y de tiempo; se estimó el tamaño de la muestra representativa para la evaluación del prototipo inicial.

Considerando que la coherencia textual se observa entre oraciones consecutivas y que el párrafo puede ser tomado como la unidad mínima de coherencia textual; se imprimieron los archivos juntos, como un solo archivo consecutivo, y se revisaron los primeros 100 párrafos con el objetivo de detectar la presencia de expresiones nominales (al menos una) compuestas por “det + nombre común” que presenten el fenómeno de correferencia o anáfora indirecta. La mayoría de los párrafos presentaron el fenómeno de correferencia o anáfora indirecta y se encontraron sólo 9 párrafos que contenían correferencia, pero lo llevaban a cabo usando pronombres. Con este análisis visual de los archivos se tuvo un panorama de la proporción de las muestras en la población ***p***; en otras palabras, la población de la que podrían hacerse inferencias al presentar los fenómenos de correferencia y anáfora indirecta y poder obtener resultados a partir de una muestra específica. Se puede calcular una estimación inicial de la proporción muestral ***p*** de párrafos que presenta correferencia o anáfora indirecta en un archivo a procesar como:

$$p = \frac{91}{100} = 0.91$$

Considerando un muestreo simple con distribución normal para este número de archivos el tamaño adecuado de la muestra n [Cochran et al. 1977] puede obtenerse de:



$$n = \frac{t^2 pq}{d^2}$$

Donde:

t = la abscisa de la curva de la distribución normal que corta un área de riesgo α en las colas. En otras palabras, es el valor de la desviación estándar, de la distribución normal, correspondiente a la probabilidad de confianza deseada.

d = margen de error (varianza esperada) de la precisión *en la proporción muestral* estimada ***p***

q = $1 - \boldsymbol{p}$

Habiendo obtenido un estimado de ***p*** se puede calcular el valor de ***q*** en q = 1 - p = 0.09. Deseando mantener un margen de error menor al 10% el valor de ***d*** = 0.10, y para un nivel de confianza del 95% el valor de ***t*** = 1.96 [ver pag 157, Spiegel, 1976], se calcula el valor de ***n***, obteniendo que el número de muestras debe ser aproximadamente 32.

$$n = \frac{(1.96)^2 (0.91)(0.09)}{(0.10)^2} = (384.16)(0.91)(0.09) = 31.46 \cong 32$$

Considerando que:

- La necesidad de obtener la información al menor costo implica terminar a tiempo el proyecto porque requerirá validación manual [Mendenhall et al. 1986].
- La distribución muestral de medias, proporciones y medianas se ajusta mucho a una normal para ***n*** igual o mayor a 30 incluso para poblaciones no normales [Spiegel 1976].
- Es el primer estudio al respecto y los resultados obtenidos pueden mejorarse junto con el prototipo en un proyecto posterior.

Se puede considerar como adecuado analizar 32 archivos seleccionando sólo los del género de articulistas (ver tabla 7). En el anexo A se presenta una tabla que muestra las características y estadísticos básicos de los documentos seleccionados.

6.4 Resultados con el prototipo

El programa de evaluación para determinar la posibilidad de correferencia se corrió sobre los 38 documentos de articulistas dando libertad de búsqueda hacia atrás hasta siete verbos obteniendo los resultados mostrados en la tabla 20 donde también se muestra el tiempo de procesamiento en segundos.

Archivo	Expresiones nominales			Tiempo
	Correferencia	No correferente	Total	Segundos
a1	64	245	309	336.03
A2	8	25	33	85.47
A4	9	22	31	233.76
A10a	36	80	116	113.53
A10b	29	80	109	113.42
A11a	36	76	112	119.85
A11b	21	81	102	115.78
A12	68	151	219	250.85
a13a	35	119	154	163.45
a13b	24	103	127	127.38
a13c	16	52	68	66.9
a14	**52**	**82**	**134**	**148.84**

Archivo	Expresiones nominales			Tiempo
	Correferencia	No correferente	Total	Segundos
a15a	28	108	136	226.32
a15b	24	53	77	24.32
a15c	13	58	71	95.79
a18	8	87	95	90.9
a19	4	10	14	8.13
a20	8	65	73	106.5
a21a	33	147	180	279.46
a21b	7	53	60	55.42
a21c	13	43	56	57.12
a22a	14	87	101	77.23
a22b	5	33	38	33.06
a23a	28	113	141	223.55
a23b	58	112	170	173.95
a24	106	270	376	640.82
a25a	52	230	282	415.79
a25b	22	83	105	147.2
a26a	27	133	160	191.3
a26b	65	112	177	302.86
a26c	49	79	128	134.13
a27	17	65	82	88.43
a28a	40	107	147	214.1
a28b	31	102	133	173.57
a28c	21	24	45	49.38
a29	8	33	41	45.75
a30a	19	98	117	144.13
a30b	2	19	21	12.8
Suma	1102	3446	4548	5887
Promedio	29	91	120	154.93

Tabla 20 Resultados de una corrida general

Los comentarios de resultados se harán sobre las corridas de programas con un documento del corpus seleccionado (a14) que tiene las características mostradas en la tabla 21; se muestra en el anexo d: ejemplo del archivo de entrada, y el texto en formato normal de lectura en el anexo c: .

Descripción	Cantidad	Porcentaje
Adjetivos	58	7.00
Adverbios	42	5.06
Determinantes	134	16.19
Nombres	200	24.16
Verbos	143	17.27
Pronombres	46	5.56
Conjunciones	65	7.85
Preposiciones	134	16.19
Numerales	3	0.36
Números	3	0.36
Palabras Totales	**828**	**100.00**

Tabla 21 Características del documento a14

Se utilizó un diccionario específico donde cada "entrada" de palabra está relacionada con las palabras que pueden ser sinónimos a ella. Una vez detectada o marcada la unidad léxica, obteniendo una expresión referencial, se convierte en un correferente potencial por lo que se buscan los posibles candidatos referentes anteriores, desde la oración previa hacia el inicio del texto; se determina el grado de satisfacción por similitud hasta lograr un nivel de satisfacción preestablecido. Si se logra, significa que existe la relación correferencial de otra forma se supone inexistente.

El programa en una corrida libre marcó 134 nombres precedidos por un determinante. Detectó una posible correferencia (relación de sinonímia) en 52 de estos nombres con algún nombre que lo antecede en la búsqueda libre de todo el contexto lingüístico. Al verificar el número de correferencias reales en el texto (verificación manual) se encontraron sólo 21. Ante esta situación se decidió restringir la búsqueda hacia atrás (tomando en cuenta que la coherencia se da entre oraciones consecutivas) y se encontró que al restringirla a quince nombres se detectaban sólo 24 con posible relación de sinonímia dentro de los cuales se encontraban los 21 correferentes verificados manualmente, en esta situación se alcanza una precisión del 87.50%; los resultados se concentran en la tabla 22.

Corrida	Total	Correferentes	No-corref	Real	Precisión %	Concordancia %
Libre	**134**	**52**	**82**	**21**	40.38	76.87
Restringida	27	24	3	21	87.50	88.88

Tabla 22 Resumen de resultados en a14 con diccionario específico

		EXPERTO		
		Si	**No**	
SISTEMA	**Si**	21	31	**52**
	No	0	82	**82**
		21	113	**134**

Tabla 23 Ejemplo de cálculo de métricas

En la tabla 23 se substituyen los valores obtenidos en la corrida, marcados con **negrita**, para mostrar un ejemplo del cálculo de las métricas del primer renglón de la tabla 22.

$$precisión = \frac{21}{52} = .4038 \qquad especificidad = \frac{21}{21} = 1 \qquad concordancia = \frac{21+82}{134} = .7687$$

Relacionando la información, del primer renglón de la tabla 22 con las características del documento en la tabla 21, se puede observar que existen 46 pronombres que "normalmente" contienen correferencias por medio del fenómeno de anáfora directa con lo cual esperaríamos 134 – (46+21) = 67 posibilidades de: anáforas indirectas o referencias que sean parte de la información complementaria del documento. La especificidad es alta (100%) porque se utilizó un diccionario de sinónimos construido específicamente para este documento; en este caso la precisión y la concordancia aumentan al restringir la búsqueda hacia atrás.

Estos resultados animaron la implantación del algoritmo de resolución de anáfora indirecta, para trabajar con nombres comunes; con el programa se hizo una corrida libre y una corrida restringida a diez verbos, obteniendo los resultados que se muestran en la tabla 24. Cabe mencionar que en la verificación manual se encontraron 23 casos de anáfora indirecta en el texto por lo que la precisión y la concordancia de la anáfora indirecta se calculan con respecto a este concepto.

		Programa			Real		Ana Ind en %	
Corrida	Total	Cor	AInd	No-cor	Cor	AInd	Precisión	Concord
Libre	134	65	27	42	21	23	85.19	97.01
Restringida	134	25	25	83	21	23	92.00	98.51

Tabla 24 Resultados en a14 con anáfora indirecta

La especificidad es alta (100%) porque se utilizó un diccionario de sinónimos construido específicamente para este documento y lo mismo puede decirse de la precisión y la concordancia. La validez de estas pruebas radica en probar el modelo y que el algoritmo es adecuado, aunque altamente dependiente de la información apropiada en el diccionario de escenarios.

Buscando que el sistema pueda trabajar con otro archivo diferente a la muestra seleccionada, se repitió el experimento para la evaluación de correferencias con el mismo archivo (a14) utilizando el diccionario de sinónimos del Laboratorio de Lenguaje Natural del CIC-IPN capturado por medio de un scanner, esperando obtener resultados iguales o muy parecidos. El programa marcó 134 nombres precedidos por un determinante. Detectó una posible relación de sinonímia en 73 de estos nombres con algún nombre que lo antecede en la búsqueda libre de todo el contexto lingüístico. Al tener validadas sólo 21 correferencias reales, se decidió restringir la búsqueda hacia atrás y se encontró que al restringir el inicio de búsqueda a la oración previa y hasta cuatro nombres anteriores se detectaban sólo 29 con posible relación de sinonímia dentro de los cuales se encontraban los 21 correferentes verificados manualmente, en esta situación se alcanza una precisión del 72.42%; los resultados se concentran en la tabla 25.

Corrida	Total	Correferentes	No-corref	real	Precisión %	Concordancia %
Libre	134	73	61	21	28.77	61.19
Restringido	110	29	81	21	72.42	92.73

Tabla 25 Resumen de resultados con diccionario del LLN CIC-IPN

La diferencia de resultados, disminución drástica de la precisión (22 y 15 puntos porcentuales), al cambiar el diccionario de sinónimos ha obligado a revisar el diccionario de sinónimos del Laboratorio de Lenguaje Natural encontrando errores debidos al proceso de captura por medio de un scanner, su proceso de corrección fue descrito en la sección 5.4.

Después de corregir el diccionario de sinónimos y obtener un diccionario de escenarios, descrito en la sección XXX, la atención se concentró en las pruebas para determinar el tamaño de la ventana de búsqueda hacia atrás para poder trabajar con texto libre.

6.5 Tamaño de ventana de búsqueda

En la sección anterior se mencionaron corridas libres en la búsqueda de la correferencia o anáfora indirecta potencial “desde la posición actual hasta el inicio del archivo” y corridas restringidas en función de el número de ocurrencias de una bandera definida “ 7 verbos, 15 nombres, 10 verbos, 4 nombres”. Estas corridas tenían como intención mantener el contexto lingüístico en memoria para lograr que los resultados incluyeran todas correferencias y anáforas indirectas, detectadas en la verificación manual, porque los diccionarios, de sinónimos y escenarios, contenían la información completa.

La intención original “*incluir el contexto lingüístico (los antecedentes necesarios) que satisfaga las correferencias y anáforas indirectas*” sigue siendo válida y esto puede lograrse de dos formas:

- almacenar todas las unidades léxicas, estructura e información implícita (de sinónimos y escenarios) y mantener el sistema dentro del límite físico de 4800 unidades léxicas (o tokens) o 45 KB aproximadamente del tamaño de archivo a procesar
- modelar al lector humano que almacena en el contexto lingüístico sólo la información relevante (nombres propios; enlaces correferenciales y de anáforas indirectas; y los ultimos N grupos de unidades léxicas necesarios para satisfacer las correferencias y anáforas indirectas.

La primera alternativa fue apropiada para el desarrollo del prototipo inicial porque permitió hacer corridas con el tamaño necesario de la ventana; pero para poder alcanzar la segunda meta o alternativa se plantean dos problemas: ¿cuál es la bandera o marcador más adecuado? y ¿cuál es el tamaño N apropiado de la ventana de búsqueda?

Para plantear mejor el problema de determinar la bandera adecuada se utilizará parte del texto del archivo a14 con “oraciones” numeradas (considerando como oración la separación de puntuación conocida como “punto”). Se han marcado con **negrita** los verbos y con *cursiva* los nombres de la expresión nominal.

1. Cuando **escribo** esto la *Madre_Coraje* peruana **acaba** de **ser** reventada por los *senderistas*.

2. **Veo** su *foto* en los *periódicos*: una *mujer* joven, atractiva, probablemente zamba, esto_**es**, mestiza de negra e india; oscura de color, en_fin, como **son** oscuros todos los *habitantes* de las *villas* limeñas, *arrabales* de miseria en donde se **hacinan** cientos de miles de *personas*.
3. **Son**, en su **mayoría**, **indígenas** que **bajaron** de los *Andes* **huyendo** del *hambre*, del *atraso* y la *tuberculosis*; **quisieron llegar** a la *ciudad*, pero **quedaron varados** en las *afueras*, a una decena de *kilómetros*, en los sórdidos *arenales* que **rodean** *Lima*, en donde **plantaron** sus *chabolas*, precarios *tenderetes* de cartón y *cajones* astillados.
4. El *liderazgo* de *María_Elena* **nació** de aquella *miseria* y de una increíble *voluntad* de superación.
5. De la *generosidad*, de la *inteligencia*, del *tesón*.

Una revisión general de las cinco oraciones permite observar el diferente tamaño y composición, como se muestra en el resumen de la tabla 26.

Oración	Verbos	Nombres	Puntuación	palabras
1	3	2	1	14
2	4	7	10	45
3	9	14	10	53
4	1	4	1	16
5	0	3	3	8

Tabla 26 Resumen de elementos de oraciones

El sólo signo de puntuación “punto” (“punto y seguido” o “punto y aparte”) contiene diferente número de nombres de acuerdo a la extensión de las expresiones nominales en la oración; además de la posibilidad de confusión con el punto que acompaña a las abreviaturas (Dr., Sr., etc). Se pueden reconocer contrastes entre la oración 1 y 4; la oración 4 con sólo un verbo agrupa hasta 4 nombres mientras la oración 1 con 3 verbos sólo agrupa a 2 nombres. Las oraciones 2 y 3 tienen el mismo número de signos de puntuación y casi el mismo número de palabras pero la oración 3 tiene mayor número de verbos y nombres. La oración 5 no tiene verbo explícito, debido al fenómeno de elipsis verbal, sin embargo contiene hasta 3 nombres y 3 signos de puntuación.

Los comentarios del párrafo anterior presentan un panorama confuso para determinar el marcador adecuado. Con el fin de observar el comportamiento de cada tipo de bandera se decidió hacer corridas de diferente tamaño utilizando uno de los archivos de texto libre, “Contra la guerra” (ver tabla 27), obteniendo los resultados que se grafican de la figura 19 a la figura 22; los resultados completos se pueden apreciar tabulados en el Anexo G.

En todas las figuras se puede observar que conforme aumenta el tamaño de la ventana (número de signos de puntuación, nombres, verbos o el “punto”) aumenta el número de elementos detectados (correferencia y anáfora indirecta) hasta alcanzar un valor constante porque han sido incluidas todas las relaciones que pueden ser detectadas.

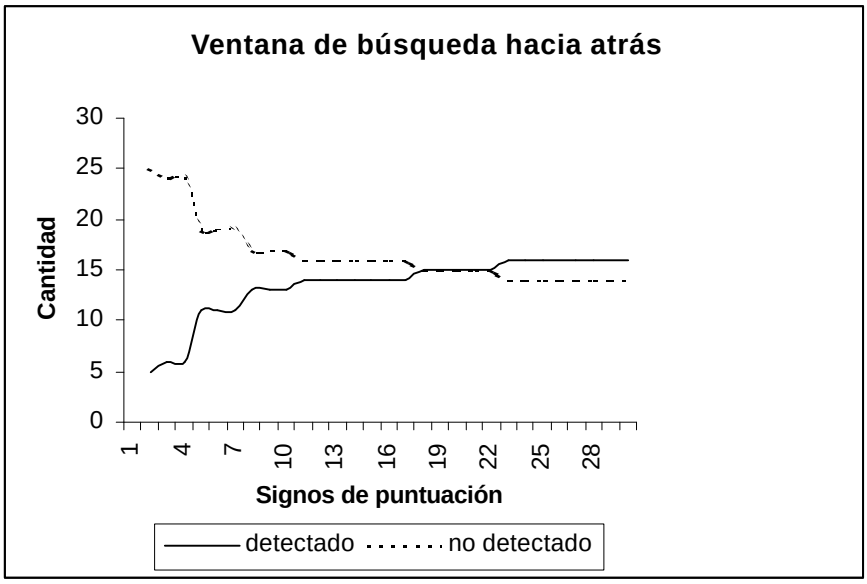


Figura 19 Evaluación como bandera de signos de puntuación

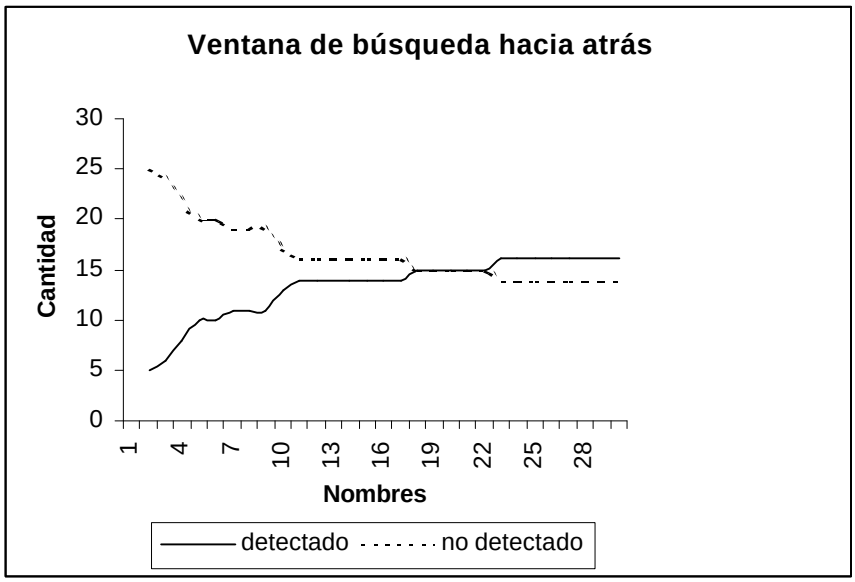


Figura 20 Evaluación como bandera de los nombres

En otras palabras, una ventana de tamaño mayor a este límite, de N banderas encontradas, incrementa sin necesitarlo el tiempo de procesamiento y la memoria requerida para mantener el contexto lingüístico; una ventana menor a este límite afectará adversamente a la precisión, especificidad y concordancia obtenidas en la corrida (por fallas en la detección).

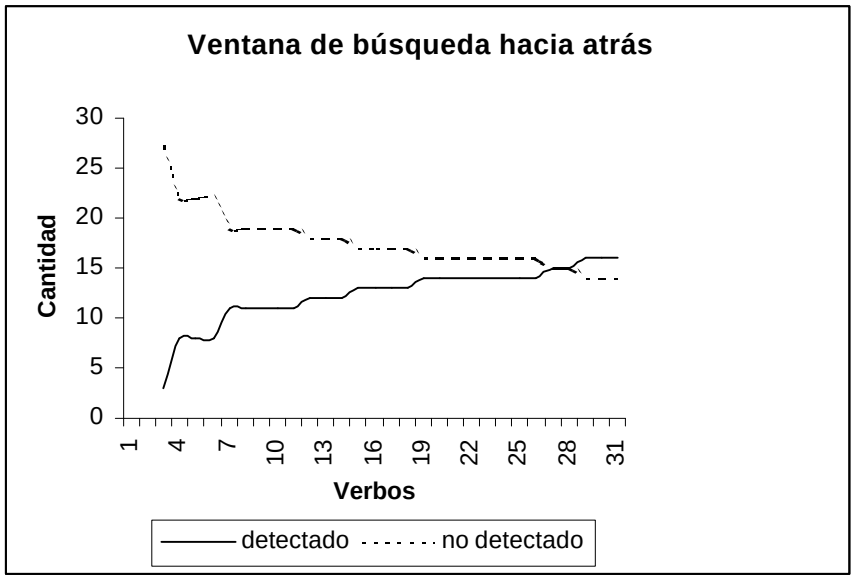


Figura 21 Evaluación como bandera de los verbos

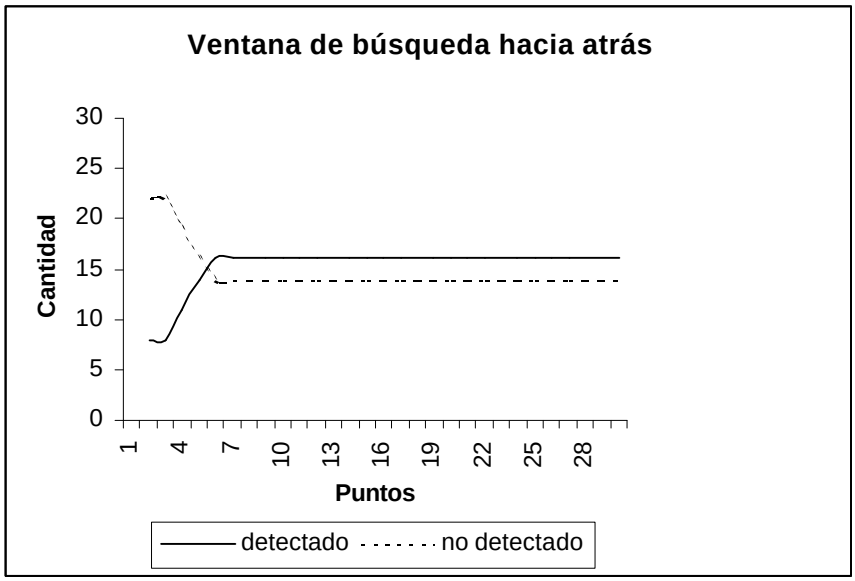


Figura 22 Evaluación como bandera de los puntos

Observando las dos primeras gráficas, figura 19 y figura 20, se puede apreciar que los signos de puntuación y los nombres aumentan su precisión con el tamaño de la ventana hasta llegar a un punto donde se mantiene constante dentro de un rango (18 a 21 y 18 a 22 respectivamente); despues de este punto ya no se incrementa la precisión a pesar del aumento de tamaño de la ventana.

La gráfica de verbos, figura 21, tiene un comportamiento similar a las de signos de puntuación y nombres aumenta su precisión con el tamaño de la ventana hasta llegar a un punto donde se mantiene constante dentro de un rango de menor variación (26 a 27); despues de este punto ya no se incrementa la precisión a pesar del aumento de tamaño de la ventana.

Observando las tres primeras gráficas, figura 19 a figura 22, se puede apreciar que los signos de puntuación, los nombres y los verbos tienen un rango de variación y el "punto" tiene un corte preciso al llegar a **seis** y se mantiene constante a partir de este valor. Con esta información ya se puede responder a las preguntas: ¿cuál es la bandera o marcador más adecuado? y ¿cuál es el tamaño N apropiado de la ventana de búsqueda?

La bandera más adecuada "*es la que tiene un punto de corte más definido antes de que se mantenga constante el número de elementos detectados*".

El tamaño apropiado de N sería **seis**, de acuerdo al resultado obtenido, pero hay que tomar en cuenta dos cosas: la recomendación de sicolingüístas que, conforme a los resultados de sus experimentos, recomiendan al menos **siete** como la capacidad del procesamiento de información del ser humano, lo que influye en la redacción y lectura de textos [Miller, 1956]; además es necesario considerar el riesgo de que existan abreviaturas, que afecten al etiquetador y repercuta en los resultados de la corrida; por lo tanto, es necesario aumentarlo a un valor que asegure "*incluir el contexto lingüístico suficiente para satisfacer las correferencias y anáforas indirectas presentes en el texto"*; el tamaño apropiado elegido es **nueve** como primera aproximación.

Resumiendo, la bandera más adecuada es el "punto" y el tamaño apropiado de la ventana es **nueve**. Lo anterior, concuerda con el conocimiento lingüístico de que "la coherencia textual se presenta entre oraciones consecutivas"; el "punto" agrupa ideas completas independientemente de la complejidad de la oración; permite modelar un "contexto

lingüístico" emulando al lector humano; e implementarlo, sin exceder la capacidad de memoria, disponible en las computadoras actuales.

6.6 Resultados con archivos de texto libre

Para las pruebas de texto libre se tomaron tres archivos recibidos por e-mail en febrero del 2003; dos son comentarios contra la guerra y el tercero una reflexión sobre el cerebro con las características presentadas en la tabla 27.

Documento	Palabras	Párrafos	Líneas	KB
Contra la guerra	208	17	23	2
Instituto Oriente	327	20	44	3
El cerebro	583	26	64	4

Tabla 27 Características de Archivos para prueba de texto libre

Los resultados completos de las corridas se presentan en el Anexo H y un resumen de resultados se muestra en la tabla 28, donde se incluye la evaluación manual del primer archivo resaltado en **negrita** y las abreviaturas son: **cd** = correferencia directa, **ci** = correferencia indirecta, **ai** = anáfora indirecta, **det** = detectados (suma de cd, ci, y ai), **ndet** = no detectados (nuevas referencias o falla del método), **nr** = no referenciados (nombres no precedidos por determinante) y **com** = total de nombres comunes marcados en el texto.

archivo	**eval**	**cd**	**ci**	**ai**	**det**	**ndet**	**nr**	**com**
contra	Auto	6	7	3	16	14	5	35
contra	**Man**	**8**	**5**	**2**	**15**	**14**	**6**	**35**
io	Auto	9	11	3	23	28	21	72
cerebro	Auto	12	28	23	63	24	33	120

Tabla 28 Resultados de corridas para prueba de texto libre

		Programa			Real		Anáfora Indirecta	
Archivo	**Total**	**Cor**	**AInd**	**ndet**	**Cor**	**AInd**	**Precisión**	**Concord**
Contra	35	13	3	14	13	2	66.67%	97.14%

Tabla 29 Evaluación del archivo en prueba de texto libre

En la tabla 29 se presentan los valores de precisión y concordancia obtenidos. Con las abreviaturas: **Cor** = correferencia, **Aind** = anáfora indirecta, **ndet** = no detectados (nuevas referencias o falla del método) y **Concord** = Concordancia.

A continuación se presenta la salida del programa ante la presencia correferencia directa, correferencia indirecta y anáfora indirecta donde se marcan los errores detectados.

Archivo f_tnt\contra.tts
(3) * guerra ←cd-(17) guerra
(8) * mundo ←cd-(78) mundo
(17) * guerra ←ci-(42) problema
(39) * paz ←cd-(91) paz
(42) * problema ←ci-(151) mal
(42) * problema ←ci-(177) momento ERROR: del modelo
(52) * guerra ←cd-(66) guerra
(78) * mundo ←ai-(108) pueblo
(91) * paz ←ci-(236) abrazo
(97) * favor ←ci-(135) nombre ERROR: del modelo
(103) * gracias ←ci-(156) derechos
(127) * malos ←ai-(221) vecino ERROR: del modelo
(169) * partes ←ai-(196) lado

En los casos de “ERROR: del modelo” la información se encuentra en el diccionario y la relación existe; además la ventana de búsqueda hacia atrás incluye esta palabra; dando una discrepancia con el lector humano. Estos son los tipos de errores que falta por resolver.

A continuación se muestra una parte del texto extraído de la salida del programa en formato htm. Se han marcado con subrayado y **negrita** las palabras involucradas en la situación que provoca el error.

El **problema** (42) es por qué no sucede lo mismo ante cualquier guerra (52) o pisoteo (54) de los derechos (57) humanos.
Además creo que debería quedar claro que el marchar por la paz (91) no significa que estemos a **favor** (97) de Sadam(99), pues " gracias (103)" a él su pueblo (108) se está hundiendo.
¿ Vivimos realmente en un mundo (118) de buenos y malos ? ¿ Están los **malos** (127)legitimados para permitir que , en su **nombre** (135), mueran inocentes ? ¿ Está tan claro lo que es el bien (148) y el mal (151)? Hemos elaborado los derechos (156) humanos y parece que no sirven para nada, pues en todas partes (169) se siguen atropellando.
Quizá llegue el **momento** (177) en que muchos millones (181) nos manifestemos en contra de este atropello(188), que a menudo sucede a nuestro lado(196).
Como dice la sabiduría (209) popular "es fácil ver la paja (216) en el ojo(219) del **vecino**(221), sin darse cuenta(225) de la viga (228) en nuestro propio ojo(232)".

En la Tabla 30 se presentan los tiempos registrados para el tipo de bandera “punto” y diferentes tamaños de ventana (2 a 22 “puntos”). Estos valores fueron obtenidos en una computadora LapTop Pentium IV, a 1.2G y 256 MB en RAM en un archivo de 583 palabras de texto libre incluyendo todo el proceso de conversión de texto libre a archivo si etiquetas, etiquetado de archivo y evaluación de anáfora indirecta.

Se puede apreciar la variación del tiempo con el cambio del tamaño de ventana, sin embargo la principal ganancia fue la reducción de tiempo lograda con la reducción del tamaño de almacenamiento de los diccionarios al simplificar la estructura, evitar repetición de entradas y duplicidad de información.

Archivo El Cerebro bandera “Punto”			
Tamaño	**Corrida**		**Formato**
Ventana	**Hora final**	**Duración**	**mm:ss.dd**
23	6:44:18		44:18.0
22	6:44:36	00:17.6	44:35.6
21	6:44:53	00:17.4	44:53.0
20	6:45:10	00:17.4	45:10.4
19	6:45:28	00:17.3	45:27.6
18	6:45:45	00:17.1	45:44.8
17	6:46:02	00:17.0	46:01.8
16	6:46:19	00:16.9	46:18.7
15	6:46:35	00:16.7	46:35.4
14	6:46:52	00:16.2	46:51.7
13	6:47:07	00:15.8	47:07.4
12	6:47:23	00:15.4	47:22.8
11	6:47:37	00:14.5	47:37.3
10	6:47:51	00:14.1	47:51.3
9	**6:48:05**	**00:13.5**	**48:04.8**
8	6:48:17	00:12.4	48:17.2
7	6:48:28	00:11.1	48:28.3
6	6:48:38	00:09.7	48:38.0
5	6:48:46	00:08.0	48:46.0
4	6:48:53	00:06.7	48:52.7
3	6:48:58	00:05.0	48:57.7
2	6:49:01	00:03.4	49:01.1

Tabla 30 Duración de corrida para diferentes tamaños de ventana

7 CONCLUSIONES

7.1 Resultados obtenidos

El trabajo de investigación realizado ha permitido desarrollar un método de resolución de la anáfora indirecta que trabaja con un buen nivel de precisión y concordancia, hasta del 92% y 98.51% respectivamente en el prototipo inicial (ver tabla 24), cuando existe información completa disponible en los diccionarios de sinónimos y escenarios. Cuando se utilizó con texto libre la precisión bajó hasta 66.67%, sin embargo el nivel de concordancia se mantuvo en un nivel semejante de 97.14% (ver tabla 29).

Estos indicadores permiten visualizar que el contexto lingüístico basado en el modelo de escenario logra la resolución automática de la anáfora indirecta nominal (meta de este trabajo) con una fuerte dependencia de la información suministrada (compilar automáticamente esta información es un área de oportunidad para trabajos futuros).

Se logró descubrir la relación existente entre la sintaxis, la semántica y la pragmática, observando que: la sintaxis *sólo marca las expresiones definidas*; apoyándose en la semántica y en la pragmática es posible determinar el tipo de referencia existente (correferencia directa o indirecta) y la relación anafórica indirecta en un texto. Esta situación llevó a realizar un algoritmo de resolución de correferencias, basado en un diccionario de sinónimos, como requisito previo a la detección de la anáfora indirecta para poder resolverla, por medio de un diccionario de escenarios.

Así pues, no se puede hablar de marcadores específicos de cada fenómeno sino de una imbricada red de relaciones que sólo puede resolverse con un algoritmo que de forma integral modele el proceso de lectura del receptor. En otras palabras, el discurso debe verse como un "*conocimiento del receptor que se ve enriquecido paulatinamente con la información recibida y conforme avanza el proceso de lectura*".

7.2 Aportaciones

Las aportaciones específicas de este trabajo son:

- Descubrir la interrelación existente de los fenómenos de correferencia (directa e indirecta) y la anáfora indirecta porque ambas utilizan expresiones nominales referenciales para manifestar su presencia.
- Descubrir el orden de evaluación requerido para discriminar el fenómeno presente que sirve como base para detectar la presencia de la anáfora indirecta.
- Desarrollar un nuevo método basado en el modelo de escenario ampliado con el contexto lingüístico que modela al lector humano, necesario para la resolución de la anáfora indirecta y las correferencias.
- Desarrollar un conjunto de programas que integrados permiten la creación de diccionarios sin repetición de entradas o duplicidad de información logrando así reducir el tamaño de los archivos y los tiempos de acceso a disco; como consecuencia reducir el tiempo de procesamiento.
- Desarrollar un conjunto de programas que permiten la extracción de información del diccionario semántico EuroWordNet en Español desde sus archivos de exportación de información.
- Construir un diccionario de escenario a partir de la información semántica almacenada en el diccionario de EuroWordNet en Español con las relaciones de holonimia, meronimia, y rol necesarias para este trabajo.

De este trabajo surgieron 4 publicaciones, tres son ponencias para congresos internacionales y una es un reporte técnico del estado del arte en anáfora indirecta publicado en el CIC-IPN (ver Ponencias y Publicaciones en apartado 0).

7.3 Recomendaciones y sugerencias para el trabajo futuro

Para mejorar la precisión del sistema desarrollado es necesario mejorar la etapa de preprocesamiento, por lo que se requiere:

- Mejorar los etiquetadores actuales, porque las técnicas utilizadas han llegado al límite, con el conocimiento lingüístico (semántico y pragmático) adicional que permita lograr una mayor automatización y precisión. Se espera desarrollar trabajo conjunto a futuro con Brants Thorsten sobre el etiquetador TnT.
- Estas mejoras se espera realizarlas con tecnologías de aprendizaje incremental basada en casos y continuar trabajando en conjunto con Montserrat Civit aprovechando las bondades de su trabajo con los corpus etiquetados en Español. Se espera poder trabajar en equipo con el Dr. Aurelio López López del INAOE para apoyar el desarrollo de un etiquetador para el Español.

Para mejorar la base de la información implícita que aumente el poder de resolución del sistema desarrollado es necesario mejorar y automatizar la construcción de diccionarios, por lo que se requiere:

- Desarrollar una biblioteca de rutinas de procesamiento de cadenas y texto multilingüe que trabaje inicialmente para el Español. Las bibliotecas actuales tienen problemas para el ordenamiento y comparación de cadenas con símbolos o letras específicos del Español (ñ, í, ü, etc). Esto obligó a revisar y realizar parte del trabajo de elaboración de diccionarios manualmente.
- Involucrarse en el desarrollo de estándares en la construcción de diccionarios semánticos y algoritmos de extracción del significado de textos libres.

Para mejorar el rendimiento global del sistema desarrollado es necesario mejorar los algoritmos de reconocimiento de entidades referenciales, por lo que se requiere:

Investigar a mayor profundidad el reconocimiento de entidades en las expresiones referenciales para poder diferenciarlas cuando se refieren a objetos diferentes del mundo real.

Es necesario ampliar el concepto de identificación para que el sistema pueda manejar objetos similares del mundo real en el mismo texto, por ejemplo:

> El **perro**$_1$ negro de Juan$_2$ mordió al **perro**$_3$ café de Pedro$_4$. El veterinario$_5$ dijo que el **perro**$_3$ debería conservar el vendaje$_6$ por una semana$_7$ para evitar una infección$_8$.

En esta oración el sistema trabaja bien porque la búsqueda hacia atrás encuentra primero al **perro**$_3$. Pero en el siguiente ejemplo fallaría:

> El **perro**$_1$ café de Pedro$_2$ fue mordido por el **perro**$_3$ negro de Juan$_4$. El veterinario$_5$ dijo que el **perro**$_1$ debería conservar el vendaje$_6$ por una semana$_7$ para evitar una infección$_8$.

En esta oración el sistema trabaja mal porque la búsqueda hacia atrás encuentra primero al **perro**$_3$ y el perro mordido es el **perro**$_1$. Es necesario desarrollar un algoritmo que identifique al objeto con todos los componentes de la frase nominal para poder evitar una identificación falsa al depender sólo del núcleo de la expresión.

Además es necesario distinguir las frases nominales atributivas de las referenciales y recuperar también la información implícita debida al fenómeno de elipsis. Se está planeando trabajar en conjunto en esta área con Hiram Calvo, aspirante al doctorado en el Laboratorio de Lenguaje Natural del CIC-IPN.

De este trabajo surgieron 4 publicaciones, tres son ponencias para congresos internacionales y una es un reporte técnico del estado del arte en anáfora indirecta publicado en el CIC-IPN (ver Ponencias y Publicaciones en apartado 0).

GLOSARIO

La inclusión de este glosario de términos se considera necesaria por las siguientes razones:

(**I**) Los autores utilizan diferentes términos para referirse al mismo fenómeno, de acuerdo a la teoría lingüística que promueven, por lo que es necesario dar claridad al uso particular en el texto; Mel'èuk por ejemplo lo hace en las páginas 21, 49, 53, 105, 163, etc.; se muestra un ejemplo de la página 49:

> "What are called here Semantic-Communicative Oppositions are also known as **Discourse Functions** [Chafe, 1994] or **Pragmatic Functions** [Dik, 1981:127-156; Bossong, 1989:28" [Mel'èuk, 2001:49].
> "Las que aquí son llamadas oposiciones semántico-comunicativas también son conocidas como **Funciones del Discurso** [Chafe, 1994] o **Funciones Pragmáticas** [Dik, 1981:127-156; Bossong, 1989:28" [Mel'èuk, 2001:49].

(**II**) Actualmente algunas definiciones y conceptos, en el ambiente científico, se encuentran en desarrollo y por lo tanto no se incluyen en los diccionarios generales del Español o Inglés.

(**III**) Para homogenizar y dar claridad al "sentido" de las palabras, utilizadas por el autor para los lectores, tomando como base una muestra seleccionada de referencias:

Diccionario de la Lengua Española. Edición de la Real Academia Española 1992; en la versión electrónica de 1995 en CD Versión 21.1.0 de Espasa Calpe, S.A.

Mel'èuk Igor. Communicative Organization in Natural Language.
John Benjamins Publishing Company. Philadelphia, USA. 2001

Utrecht institute of Linguistics OTS. Utrecht University.
http://www2.let.uu.nl/UiL-OTS/Lexicon/
Marzo 11, 2003

International Linguistics Department. Summer Institute of Linguistics (Dallas, TX)
http://www.sil.org/linguistics/glossary/
Marzo 11, 2003

The American Heritage® Dictionary of the English Language Fourth Edition. 2000.
http://www.bartleby.com/
Marzo 11, 2003

(**IV**) Para apoyar el enriquecimiento lingüístico del ambiente computacional, incluyendo términos relacionados que apoyan la comprensión de las descripciones.

Nota: En la **Descripción** se marcan con **negrita** los términos incluidos en este glosario.

ablaut

De origen Alemán, es la variación vocálica radical que coincide con una oposición gramatical. Ejemplos:

Alemán.- Hand: Händ***e*** formación de número *singular:plural*
Español.- perr***o***: perr***a*** formación del género *masc:fem*

alegoría

tropo que consiste en una serie de metáforas relacionadas entre sí; es una **metáfora** continuada. Ejemplo:

es mar la noche negra; *mar* → este conjunto de
la nube es una concha, *concha* → términos están
la luna es una perla. *perla* → *referidos al mar*
José Juan Tablada

anáfora

Es un mecanismo un mecanismo de **economía lingüística** para hacer **referencia** de una **entidad** *anáfora* (o referente) a una entidad *antecedente* (o referido) que ya ha sido mencionada en el texto. Aparece una **entidad** lingüística que *debe ser vinculada* con otra ya mencionada. Ejemplo:

Juan baña a la *niña* y María *la* seca con la toalla.

antecedente

Un **antecedente** es una **entidad** referenciada por *otra* que la precede o sigue. El **antecedente** suministra la información necesaria para interpretar correctamente a la *otra*. Ejemplos:

El amigo *que* te presente ayer ...
(amigo es **antecedente** de *que*)
Si necesitas *una*, hay toallas en el closet.
(toallas es **antecedente** de *una*)

antonomasia

tropo que consiste en emplear el nombre propio por el apelativo (o apodo) y viceversa. Ejemplos:

Salomón → para referirse a una *persona sabia*
el arcángel → para referirse a San Miguel arcángel

catáfora

Es un mecanismo un mecanismo de **economía lingüística** para hacer **referencia** de una **entidad** *anáfora* (o referente) a una entidad *antecedente* (o referido) que va a ser mencionada en el texto. Aparece una **entidad** lingüística que *debe ser vinculada* con otra por mencionar. Ejemplo:

Cerca de *él*, *Juan* vio una serpiente.

contexto (lingüístico)	En lingüística, se entiende por contexto el *conjunto de conocimientos y creencias compartidos* por los interlocutores de un proceso comunicativo que son *necesarios para producir e interpretar sus enunciados*. En otras palabras, es el entorno lingüístico del cual depende el sentido y el valor de una **palabra** o **frase** considerados. Ejemplo: 1) Terminamos rápido la pesca porque encontramos un gran *banco*. 2) Espérame a las 12:00 en el *banco* para que vayamos de compras. 3) Descansa en el *banco* mientras me anudo la corbata. Donde en: 1) *banco* = conjunto de peces que van juntos en gran número 2) *banco* = inmueble de una institución financiera 3) *banco* = asiento con respaldo o sin él para una o más personas
correferencia	Es *la relación* que existe entre dos **frases** nominales que se interpretan *refiriéndose a una misma entidad*; en otras palabras, es la **referencia** en una **expresión** al mismo **referente** de otra **expresión**. La correferencia es convencionalmente denotada por coindexado en las representaciones lingüísticas. Ejemplos: Lisa$_1$, dijo que ella$_1$ vendría... (*Lisa*$_1$ y *ella*$_1$ se refieren a la misma entidad) *Tu*$_2$ dijiste, que *tu*$_2$ vendrías... (ambos *tu*$_2$ se refieren a la misma entidad) Un *carro*$_3$ chocó frente a la casa$_4$. El *vehículo*$_3$ quedó desecho. (*carro*$_3$ y *vehículo*$_3$ hacen referencia a la misma entidad)
dada, antigua (información)	Se dice que la información es “dada” (en Inglés given) cuando el emisor asume que *cierta información está implícita en el escenario del receptor porque ha sido establecida en el discurso*.
deixis (deítica, o)	Función que desempeñan ciertos elementos lingüísticos o paralingüísticos *señalando lo que está presente en la comunicación*. Se realiza mediante ciertos elementos lingüísticos como: *esta, esa, aquella* que indican un objeto; *yo, vosotros* que indican personas; *allí, arriba* que indican un lugar; o un tiempo, como *ayer, ahora*. El señalamiento puede referirse a otros elementos del discurso (ejemplo 1) o presentes sólo en la memoria (ejemplo 2). 1) *Invité a tus hermanos y a tus primos, pero* ESTOS *no aceptaron* 2) AQUELLOS *días fueron magníficos*
determinante	es la **unidad léxica** que precede al nombre *para especificar su **referencia***, incluyendo la cantidad. Ejemplo: *Todos esos* carros están en venta *Todos* especifica la cantidad total de carros *esos* indica el lugar donde se encuentran relativo al hablante

diacrónico	Palabra compuesta por el prefijo dia- que significa "*a través de*" y cronos que significa "*tiempo*". Opuesto a **sincrónico**, en lingüística se aplica a los hechos y relaciones que se desarrollan, suceden o analizan a través del tiempo, considerando su evolución.
economía lingüística	forma de suprimir o sustituir elementos repetidos de un texto que se dan como entendidos por el receptor (**elipsis**, **anáfora**, **catáfora**).
elipsis	**figura de construcción**, que es un mecanismo de **economía lingüística**, para la *omisión* de palabras repetidas (que se sobreentienden) dentro de una frase. No aparece ninguna **entidad** lingüística que deba ser vinculada con un **antecedente**, simplemente se deja un vacío. Ejemplos: *Mi amigo* me saludó cuando [*mi amigo*] entró. Juan *toca* el piano; María [*toca*] la guitarra. Miguel *es tonto y perezoso*; Francisco, no [*es tonto y perezoso*].
elipsis catafórica	se da en el caso de que el componente omitido (elidido) "*aparezca*" después de la posición que debía ocupar. Ejemplo: Si [*Juan*] gana en la lotería, *Juan* se compra un piano.
entidad (del discurso)	Es *el concepto* concreto o abstracto asociado a una **expresión** lingüística que puede actuar como **antecedente** para una **referencia**. Ejemplo: Juan perdió su cartera en la fiesta Las entidades pueden ser "*Juan*", "*su cartera*" , "*la fiesta*", "*perdió su cartera*", "*en la fiesta*" y la oración completa "*Juan perdió su cartera en la fiesta*". Cada entidad puede ser referenciada por una oración posterior en el texto. "Juan$_1$" → ¿Cómo *le*$_1$ pasó? "su cartera$_2$" → Debería cuidar*la*$_2$ mejor. "la fiesta$_3$" → ¿*La*$_3$ del sábado pasado? "perdió su cartera"$_4$ → Es terrible que te pase *eso*$_4$ "en la fiesta"$_5$ → *Ahí*$_5$ es común que suceda. "Juan perdió su cartera en la fiesta"$_6$ → Para él *esto*$_6$ fue horrible.

escenario	En el modelo de escenario, la idea básica es que la interpretación de la anáfora indirecta se encuentra siempre referida a un espacio o dominio mental apropiado de **referencia**. Es importante recordar que un escenario es la parte del teatro construida y dispuesta convenientemente para que en ella se puedan colocar las decoraciones y representar las obras dramáticas o cualquier otro espectáculo; de esta forma el modelo de escenario intenta representar en la mente del receptor la parte **explícita** del lenguaje por los actores y objetos en el contexto lingüístico como la escena y la parte **implícita** por el **rol** que representan y el conjunto de elementos que sugiere a los espectadores la escena mientras ocurre en el escenario.
explícito, a (o información emitida)	La información "perceptible" que recibe el receptor por el medio de la representación adecuada al proceso de comunicación. En el caso del texto escrito se refiere *al conjunto de símbolos de puntuación y letras utilizados de acuerdo a la estructura secuencial* en **palabras**, sintagmas, cláusulas, oraciones, líneas párrafos, etc.
expresión(lingüística)	Es lo que en una **unidad léxica**, corresponde sólo al *significante* oral o escrito; en otras palabras, la información **explícita** del texto.
figura de construcción	Consiste en la *alteración del orden normal* de la frase, o de las reglas de concordancia en cuanto a género, número y persona. Entre las figuras de construcción se incluyen el **hipérbaton**, el **pleonasmo**, la **elipsis** y la **silepsis**.
foco	ver: rema
frase	Es un conjunto de **palabras** capaz de desempeñar como un todo una función sintáctica (sujeto, verbo, complemento, etc). También conocida como **sintagma** o grupo sintáctico está integrada por componentes que realizan diferentes funciones: núcleo, **determinante**, modificador, etc. La categoría de la frase se define exclusivamente por la categoría de su núcleo así tenemos: frase nominal si su núcleo es un nombre o sustantivo, frase verbal si su núcleo es un verbo, etc. Ejemplo: 1) Hay *mesas* → *mesas* es un nombre que desempeña la función de complemento directo 2) Hay *mesas llenas de papeles* → ***mesas*** *llenas de papeles* aquí el núcleo es ***mesas*** por lo que la frase es una frase nominal y también desempeña la función de complemento directo

hipérbaton

figura de construcción que altera el orden normal de los elementos de una oración. Ejemplos:

Verbo antes del Sujeto → Vuela *el águila* a gran altura.
Adjetivo antes del Sustantivo → No es tan fiero el *león* como lo pintan
Adverbio antes del Verbo → Tranquilamente *volvió* a su casa
Complemento antes del Verbo → Con sus amigos *es* muy generoso

hiperonimia

una relación que denota *una categoría superior* de una clase más particular en una estructura jerárquica (inverso de **hiponimia**). Ejemplo:

león, tigre, gato → pertenecen a la clase *felino*

hiponimia

una relación que denota *una subcategoría* de una clase más general en una estructura jerárquica (inverso de **hiperonimia**). Ejemplos:

león → es un tipo de felino
tigre → es un tipo de felino

holonimia

una relación que denota *ser compuesta o que contiene a* otra (inverso de **meronimia**). Ejemplos:

el árbol → *tiene* hojas
el libro → *tiene* hojas de papel

implícito, a

(o información conocida)

El emisor asume que *cierta información* (significado, codificación, reglas, etc) *es conocida por el receptor*, o posible de obtener por un proceso de inferencia o deducción, en un proceso de comunicación porque ha sido establecida en el discurso, es parte del conocimiento común o es parte del **contexto** extralingüístico.

locución

Es una **unidad léxica** de dos o más palabras que funciona como elemento oracional. Su sentido unitario no siempre se justifica, como suma del significado normal de los componentes. Ejemplos:

1) adjetiva .- La que funciona como complemento de un nombre a manera de adjetivo: de tomo y lomo; de pacotilla; de rompe y rasga.
2) adverbial .- La que funciona como adverbio: de antemano; de repente.
3) conjuntiva .- La que funciona como conjunción: por consiguiente; con tal que; a pesar de.
4) interjectiva .- La que funciona como una interjección: ¡Ay de mí!; ¡válgame Dios!
5) prepositiva .- La que funciona como preposición: en pos de, para con, en torno a.

meronimia — una relación que denota *ser componente o parte de* otra (inverso de **holonimia**). Ejemplos:

la puerta *de* la habitación
el motor *del* coche

metáfora — **tropo** donde se pasa del sentido directo al figurado mediante una comparación. Ejemplo:

vejez → el atardecer de la vida
amanecer → el despertar del día

metátesis — (thesis = del Griego "colocación") consiste en cambiar de lugar algún sonido o vocablo. Ejemplo:

murcié*galo* → murcié*lago*

metonimia — **tropo** que consiste en llamar una cosa con el nombre de otra, tomando los efectos por las causas y viceversa. Ejemplos:

las canas → por la vejez
tiene un golpe en la pierna → por tiene una contusión por el golpe que le han dado en la pierna
perder la cabeza → por perder el juicio

onomatopeya — la **palabra** que imita un sonido característico. Ejemplos:

guau → ladrar del perro (en Francés: *ouaoua*; Alemán: *wauwau*)
bee → balido de la oveja
tic tac → sonido del reloj
glu glu → chorro de agua

palabra — Es la representación gráfica de un conjunto de sonidos articulados que expresan una idea. En otras palabras, es una **unidad léxica** (significante) asociada con un concepto (significado).

pleonasmo — **figura de construcción** que consiste en usar **palabra**s redundantes que contribuyen a dar fuerza a la expresión. Ejemplo:

Lo vi *con mis propios ojos*

pronombre — Son **unidades léxicas** que señalan o apuntan a significados dependientes del **contexto**. Ejemplo:

1) {Juan$_1$ regañó a María$_2$ injustamente.}$_3$
2) José$_4$ comentó que *eso*$_3$ no está bien.
3) *Ella*$_2$ está dispuesta a perdonar*lo*$_1$ si *él*$_1$ pide disculpas.
eso$_3$ señala el hecho "Juan$_1$ regañó a María$_2$ injustamente"
Ella$_2$ señala a una 3ª persona de género femenino en este caso María$_2$
lo$_1$ y *él*$_1$ señala a una 3ª persona de género masculino; podría ser Juan$_1$ o José$_4$ y sólo el contexto permite determinar que es Juan$_1$ porque es el que ofendió a María$_2$.

referencia	Es *la relación* simbólica que una **expresión** lingüística establece con el objeto que representa. En este fenómeno una **frase** nominal en una oración *es asociada con algún* ***objeto*** en el mundo real, su **referente**. Ejemplo: *Juan* pateó el *balón* con *fuerza* anotando un *gol*. *Juan* hace referencia a un individuo, varón, etc.; b*alón* hace referencia al tipo de pelota que se utiliza para jugar fútbol; f*uerza* hace referencia al concepto que permite cambiar el estado de reposo o movimiento de un cuerpo (la pelota); g*ol* hace referencia a la unidad de puntaje en un partido de fútbol.
referente, referido	Es el **objeto** *al que una expresión lingüística hace referencia*. Ejemplos: Juan → individuo, varón, ... pelota → objeto esférico usado en deportes ... perro → mamífero doméstico de la familia cánidos ... vergüenza → turbación del ánimo ocasionada por ...
rema	*Es la parte de la oración que contiene la información del discurso expresada por el emisor.* También conocida como *comentario* sobre el **tema**; **foco**; núcleo; información nueva o adicional; nueva información. Algunos autores utilizan *foco* como sinónimo de *rema.* Otros utilizan *foco* para distinguir, de entre la nueva información, la de más alto interés o *la más interesante* desde el punto de vista comunicativo; es en este sentido como se utiliza *foco* en este trabajo.
rol (papel)	Del Francés *rôle*, en Inglés role, en Español sinónimo de *papel*, es la parte o personaje de una obra de teatro que ha de representar un actor en un **escenario**; en pocas palabras la función que se desempeña. En este trabajo se denomina así, al tipo de relación que se establece entre expresiones de acuerdo a la función semántica (agente, paciente, lugar, instrumento, etc.) que desempeñan en la oración.
silepsis	**figura de construcción** que consiste en hacer concordar un masculino con un femenino y un singular con un plural o viceversa. Ejemplo: su Santidad es justo y bondadoso *fem masc masc* la mitad de los soldados murieron en el frente *singular plural*
sincrónico	Opuesto a **diacrónico**, en lingüística se aplica a los hechos y relaciones que se desarrollan, suceden o analizan en un momento o período dado de su existencia histórica, sin atender a su evolución.

sinécdoque

tropo que consiste en designar al *todo* con el nombre de una de las partes, o viceversa. Ejemplos:

todo x partes	todo México lo supo → muchos mexicanos lo supieron
género x especie	María tiene un felino en casa → María tiene un gato en casa
especie x género	el pan de cada día → los alimentos diarios
concreto x abstracto	Cervantes manejó la pluma → Cervantes cultivó la literatura
singular x plural	ama siempre la virtud → ama siempre las virtudes

sinonímia;

sinónimo, a

una relación que denota *semejanza o similitud de significado* (del griego ***sin*** = semejanza y ***ónyma*** = nombre).

Algunos autores distinguen entre sinónimos “puros” y “cercanos”.

Los sinónimos puros son palabras provenientes de lenguajes distintos con el mismo significado.

Un ejemplo: ***prólogo*** .- del griego pro = antes y logos = discurso, “antes del discurso”; ***prefacio***.- del latín pre = antes y fari = decir, “antes de decir”; t*exto inicial de un discurso* y en general de lo que precede a algo. Se utiliza indistintamente como título de un apartado en la edición de libros.

Otro ejemplo: *carro* .- del latín carrus; *coche* del turco cochí = carruaje; *tipo de vehículo que sirve para transportar personas*.

Juan chocó ayer su *carro*. El *coche* quedó desecho.

Los sinónimos cercanos son palabras que de acuerdo al contexto (oración) tienen significados semejantes. Se utilizan para hacer referencia a la misma entidad sin repetir la misma palabra. Por ejemplo:

Juan chocó ayer su *carro*. El *vehículo* quedó desecho.

En este trabajo no se toma en cuenta esta distinción ya que no influye en el tratamiento del fenómeno.

sintagma

ver: **frase**

tema	*Es la parte de la oración donde el emisor presenta sobre lo que se va a hablar y de esta forma lo comparte con el receptor*. En otras palabras, la información *sobre la que se da información adicional*; es el asunto o materia del discurso; es aquello de lo que se habla o escribe y a lo que se deben subordinar todos y cada uno de los enunciados del texto. Algunos autores utilizan **tópico** como sinónimo de *subtema* viendo el tópico como tema "local" a la oración para distinguirlo del tema "global" del documento. Otros autores utilizan **tópico** como sinónimo de *tema* y así se utiliza en este trabajo.
tópico	ver: **tema**
tropo	Empleo de palabras en sentido distinto del "*normal*" que les corresponde, pero que *tienen con éste alguna semejanza o relación*. Los principales tropos son: la **metáfora**, la **alegoría**, la **sinécdoque**, la **metonimia** y la **antonomasia**.
umlaut	(de origen Alemán) la modificación del timbre de una vocal bajo la influencia de una vocal vecina; recibe también los nombres de inflexión y de mutación.
unidad léxica (UL)	Es el conjunto de **unidades morfológicas** con *una función y un significado* predefinidos en un texto. La unidad léxica más conocida es la **palabra** que consta de una unidad morfológica; un nombre propio tiene al menos una; la **locución**, en cambio, consta al menos de dos. Ejemplo: Juan_Treviño come picante a_partir_de los tres años función y significado: Juan_Treviño → nombre propio; hace referencia a una persona come → verbo; hace referencia a la acción de ingerir alimento picante → nombre común; hace referencia al condimento con chile a_partir_de → locución prepositiva; en este caso significa "desde" Nota: El nombre propio y la **locución** prepositiva se representan unidos por "_" para apreciar mejor la composición de la unidad léxica.

unidad morfológica (UM)	Es el conjunto, no vacío, de símbolos con *una función predefinida* en un texto. En el ambiente de lenguajes de programación y análisis sintáctico se le conoce como “token” (una instancia de **expresión** lingüística) e incluye a todas las expresiones terminales y no terminales: signos de puntuación, operadores, identificadores, etc. Su función puede ser sólo estructural, por ejemplo: 1) Sufragio efectivo *no**,*** reelección 2) Sufragio efectivo***,*** *no* reelección En estas frases la diferencia se observa en que la posición de la “coma”, para agrupar el “*no*”, cambia totalmente el sentido: en la primera se *rechaza el sufragio efectivo* y se desea la reelección; mientras que en la segunda se *desea el sufragio efectivo* y se rechaza la reelección. Cabe hacer notar que el espacio (o blanco) *funciona únicamente como separador* de UM’s y por lo tanto no se le considera como una UM en el proceso de etiquetado.

PONENCIAS Y PUBLICACIONES

Ponencias en Congresos ó Seminarios

Evaluation of TnT tagger for Spanish	Fourth Mexican International Conference on Computer Science (ENC 2003) Apizaco, Tlax, México	Sep, 2003
A method for indirect anaphora detection	XI International Computing Conference DF, México	Nov 2002
Coherencia Textual: un reto para la lingüística computacional	9º Congreso Internacional de Investigación en Ciencias Computacionales Puebla, México	Oct 2002
Resolución de la Polisemia en el análisis automático de Texto	Ciclo de Seminarios de Investigación CIC-IPN México, DF	Jun 2000
Indexado de Imágenes Textuales No Nítidas	IV Coloquio Nacional para el Fomento y Desarrollo de la Investigación Puebla, México	Nov 1999
Sistemas para el Trabajo Colaborativo	Academia de Sistemas y Computación del I.T. de Puebla. Puebla Pue., México	Abr 1997
Diseño Curricular: Especialidad en Sistemas Expertos	Semana de desarrollo Académico del I.T. de Puebla. Puebla Pue., México	Abr 1997
Desarrollo de tutoriales para Informática	IV Reunión Nacional de Seguimiento Curricular en I.T. Chihuahua II. Chihuahua, México	Ago 1995
Propuesta de especialidad en Sistemas Expertos	III Reunión Nacional de Seguimiento Curricular en I.T. de la Costa Grande. Guerrero, México	Oct 1994

Publicaciones

Evaluation of TnT tagger for Spanish	Fourth Mexican International Conference on Computer Science (ENC 2003) Apizaco, Tlax, México	Sep, 2003
A method for indirect anaphora detection	XI International Computing Conference DF, México	Nov 2002
Coherencia Textual: un reto para la lingüística computacional	9º Congreso Internacional de Investigación en Ciencias Computacionales Puebla, México	Oct 2002
La anáfora indirecta en la lingüística computacional	Informe Técnico Centro de Investigación en Computación IPN ISBN 970-18-7908-2	Mar 2002
Indexado de Imágenes Textuales No Nítidas	IV Coloquio Nacional para el Fomento y Desarrollo de la Investigación Puebla, México	Nov 1999

REFERENCIAS

1. Alexandrov, M., Gelbukh, A., y Makaganov, P. (2000) On Metrics for Keyword-based Document Selection and Classification. In Conference on Intelligent Text Processing and Computational Linguistics CICLing-2000. Centro de Investigación en Computación. Instituto Politécnico Nacional. February 13-19, 2000. México DF, México. 373-389
2. Aone, C. y Bennett, S. (1995) Evaluating automated and manual acquisition of anaphora resolution rules. In: Proceedings of ACL'95, 122-129
3. Aone, C. y McKee, D. (1993) A language-independent anaphora resolution system for understanding multilingual texts. In: Proceedings of the 31st Annual Meeting of the ACL (ACL'93) The Ohio State University, Columbus, Ohio, 156-163
4. Baldwin, B. (1997) CogNIAC: high precision correference with limited knowledge and linguistic resources. In: Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution Madrid, Spain 38-45
5. Carbonell, J. y Brown R. (1988) Anaphora Resolution: a Multi-Strategy Approach. In: Proceedings of the 12. International Conference on Computational Linguistics (COLING'88), Vol. I, Budapest, Hungary 96-101
6. Carter, D. (1987) Interpreting Anaphora in Natural Language Texts. Ellis Horwood, Chichester
7. Carter, D. (1990) Control issues in anaphor resolution. In: Journal of Semantics, 7, 435-454
8. Cerdá, Massó Ramón (1975) Lingüística Hoy. Colección “Hay que saber”. 3ª Edición, Teide. Barcelona, España
9. Chafe, W. (1987) Cognitive Constraints in Information Flow. In R. Tomlin (Ed.), Coherence and Grounding in Discourse (pp 21-51) Benjamins, Amsterdam
10. Chafe, W. (1994) Discourse, Consciousness, and Time. Chicago–London: The University of Chicago Press. 327 pp.
11. Chomsky, Noam (1986) Knowledge of language: Its Nature, Origin and Use. Praeger, New York
12. Clark, Herbert H. y Haviland, Susan E. (1977) Comprehension and the given-new contrast. In R. Freedle (Ed.), Discourse Production and Comprehension.
13. Cochran, W.G. (1977) Sampling Techniques. John Wiley & Sons. USA
14. Collins, M. (1997) Three generative, lexicalised models for statistical parsing. In: Proceedings of the 35th Annual Meeting of the ACL (ACL'97) Madrid, Spain 16-23

15. Cornish, Francis (1999) Anaphora, Discourse, and Understanding. Oxford University Press New York
16. Daelemans, W.; Zavarel, J.; van der Slot, K.; y van den Bosch, A. (1999) TIMBL: Tilburg Memory Based Learner, version 2.0. Reference guide, ilk technical report ILK, Tilburg University 99-01
17. Dagan, Ido e Itai, Alon (1990) Automatic processing of large corpora for the resolution of anaphora references. In: Proceedings of the 13th International Conference on Computational Linguistics (COLING'90), Vol. III, 1-3, Helsinki, Finland 1-3
18. Dagan, Ido e Itai, Alon (1991) A statistical filter for resolving pronoun references. In: Y.A. Feldman, Y.A, Bruckstein, A (eds): Artificial Intelligence and Computer Vision, Elsevier Science Publishers B.V. (North-Holland) 125-135
19. Diccionario de la Lengua Española. Edición de la Real Academia Española 1992; en la versión electrónica de 1995 en CD Versión 21.1.0 de Espasa Calpe, S.A.
20. Erku, F. y Gundel, J. K. (1987) The pragmatics of indirect anaphors. In J. Verschueren & M. Bertuccelli-Papi (Eds.), The pragmatic perspective: Selected papers from the 1985 International Pragmatics Conference (pp. 533-545) Amsterdam: John Benjamins.
21. Evans, R. (2000) A Comparison of Rule-Based and Machine Learning Methods for Identifying Non-nominal It. In: Natural Language Processing-NLP2000, Second International Conference Proceedings, Lecture Notes in Artificial Intelligence, Springer-Verlag, 233-242
22. Ferrandez, A.; Palomar. M.; y Moreno L. (1997) Slot unification grammar and anaphora resolution. In: Proceedings of the International Conference on Recent Advances in Natural Language Proceeding (RANLP'97) Tzigov Chark, Bulgaria 294-299
23. Fillmore, Charles (1982) Toward a descriptive framework for spatial deixis. In R. Jarvella and W. Klein (ed.) Speech, Place and Action. John Wiley and Sons, Chichester. 31-59
24. Fox, B. A. (1987) Discourse structure and anaphora: written and conversational English. Cambridge University Press, Cambridge. USA
25. Fraurud, K. (1996) Cognitive ontology and NP form. In T. Fretheim & J. K. Gundel (Eds.), Reference and referent accessibility (pp. 193-212) Amsterdam: John Benjamins
26. Fretheim, T. y Gundel, J. K. (Eds.) (1996) Reference and referent accessibility. Amsterdam: John Benjamins.
27. Fukumoto, F.; Yamada, H.; y Mitkov, R. (2000) Resolving overt pronouns in Japanese using hierarchical VP structures. In: Proceedings of Corpora and NLP Monastir, Tunisia. 152-157
28. Galicia-Haro Sofía N., Bolshakov I. A. y Gelbukh A. F. (1999) Un modelo de descripción de la estructura de las valencias de verbos españoles para el análisis automático de textos
29. Garrod, Simon C. y Sanford, Anthony J. (1994) Resolving Sentences in a discourse Context. In M.A. Gernsbacher (Ed.) Handbook of Psycholinguistics. Academic Press, London. 675-98

30. Ge, N.; Hale, J.; y Charniak, E. (1998) A statistical approach to anaphora resolution. In: Proceedings of the Workshop on Very Large Corpora. Montreal. Canada. 161-170
31. Gelbukh, Alexander (2000) Computational Processing of Natural Language: Tasks, Problems and Solutions. Congreso Internacional de Computación en México DF., Nov 15-17, 2000
32. Gelbukh, Alexander y Sidorov Grigori (1999) A Thesaurus-based Method for Indirect Anaphora Resolution. Revised version of On Indirect Anaphora Resolution In: Proceedings of PACLING-99
33. Gelbukh, Alexander y Sidorov Grigori (2001) La estructura de dependencias entre las palabras en un diccionario explicativo del español: resultados preliminares. Por publicar.
34. Gernsbacher, Morton Ann (1997) Coherence Cues Mapping during Comprehension. In: Processing Interclausal Relationships. Editado por Costermans Jean y Fayol Michel. Lawrence Erlbraum Associates Publishers. New Jersey, USA.
35. Guzmán Arenas Adolfo (1999) Finding the main themes in a Spanish document. Journal Expert Systems with Applications, Vol 14, Nº 1/2. January/February, 139-148
36. Hahn, U.; Strube, M.; y Markert, K. (1996) Bridging textual ellipses. Proceedings of the 16th International Conference on Computational Linguistics (pp 496-501)
37. Hawkins J.A. (1978) Definiteness and indefiniteness. Humanities Press, Atlantic Highlands, New Jersey, USA.
38. Hirst, Graeme (1981) Anaphora in Natural Language Understanding. Springer Verlag, Berlin
39. Hobbs, J. R. (1976) Pronoun resolution. Research Report 76-1. New York: Department of Computer Science, City University of New York
40. Hobbs, J. R. (1978) Resolving pronoun references. Lingua, 44 339-352.
41. Huang Yang (1994) The Syntax and Pragmatics of Anaphora: a study with special reference to Chinese. Cambridge University Press, Cambridge, USA
42. Huang Yang (2000) Anaphor: A Croos-linguistic Approach. Oxford University Press, New York, USA
43. Kameyama, M. (1997) Recognizing referential links: an information extraction perspective. In: Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution Madrid, Spain 46-53
44. Kempson Ruth (1988a) Grammar and conversational principle. In Newmeyer (1988:ii, 139-63)
45. Kempson Ruth (1988b) Logical Form: the grammar cognition interface. In Journal of linguistics, 24:393-431.
46. Kempson, Ruth (1982) Teoría Semántica. Editorial Teidé. Barcelona, España
47. Kennedy, C. y Boguraev, B. (1996) Anaphora for everyone: pronominal anaphora resolution without a parser. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING'96) Copenhagen, Denmark 113-118
48. Krahmer, Emiel y Piwek, Paul (2000) Varietes of Anaphora

49. Kurohashi, Sadao; Murata, Masaki; Yata Yasunori; Shimada Mitsunobu; y Nagao Makoto (1998) Construction of Japanese Nominal Semantic Dictionary using “A NO B” Phrases in Corpora. Kyoto University
50. Lappin, S., Leass, H. (1994) An algorithm for pronominal anaphora resolution, Computational Linguistics, 20(4), 535-561
51. Levinson, Stephen C. (1989) A Review of relevance. In Journal of Linguistics. 25:455-72
52. Matsui, Tomoko (1993) Bridging reference and the notions of “topic” and “focus”. Lingua, 90:49-68.
53. Matsui, Tomoko (1995) Bridging and relevance Ph.D. dissertation, University College London.
54. Mel’èuk Igor. (2001) Communicative Organization in Natural Language. John Benjamins Publishing Company. Philadelphia, USA.
55. Mendenhall, W., Scheaffer, R.L., y Wackerly, D.D. (1986) Estadística Matemática con Aplicaciones. Grupo Editorial Iberoamérica. México.
56. Miller, G.A. (1956) The Magical Number Seven or Minus Two: Some Limits on Our Capacity of Processing Information, Psychological Rev. 63 81-97.
57. Minsky Marvin L. (1975) A Framework for representing Knowledge. In P. Watson (Ed.) The Psychology of Computer Vision. McGraw Hill, New York
58. Mitkov, R. (1995) An uncertainty reasoning approach for anaphora resolution. In: Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'95), Seoul, Korea 149-154
59. Mitkov, R. (1997) Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches. In: Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution, Madrid, Spain 14-21
60. Mitkov, R. (1998a) Evaluating anaphora resolution approaches. In: Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC'2) Lancaster, UK
61. Mitkov, R. (1998b) Robust pronoun resolution with limited knowledge. In: Proceedings of the 18.th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference Montreal, Canada 869-875
62. Mitkov, R. (2000a) Pronoun resolution: the practical alternative. Paper presented at the Discourse Anaphora and Anaphor Resolution Colloquium (DAARC), Lancaster, UK (1996) Also appeared in: Botley, S., McEnery, T. (eds): Corpus-based and computational approaches to discourse anaphora. John Benjamins, Amsterdam/Philadelphia1 89 -212
63. Mitkov, R. (2000b) Towards more consistent and comprehensive evaluation in anaphora resolution. In: Proceedings of LREC'2000, Athens, Greece, 1309-1314
64. Mitkov, R. (2000c) Towards more consistent and comprehensive evaluation of robust anaphora resolution algorithms and systems. Invited talk. In: Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000), Lancaster, UK

65. Mitkov, R. (2001) Oustanding Issues in Anaphora Resolution. In: Proceedings of Second International Conference, CICLing 2001, Mexico City, México, 18-24 February. Alexander Gelbukh (Ed.) Lecturer Notes in Computer Science LNCS 2004 Springer 110-125
66. Morales, Carrasco Raúl (1999) Indexado de imágenes textuales no nítidas. 4º Coloquio Nacional para el Fomento y Desarrollo de la Investigación en Ciencias de la Ingeniería. Instituto Tecnológico de Puebla, Puebla, México Nov, 26.
67. Morales, C.R. y Gelbukh, A (2003) Evaluation of TnT Tagger for Spanish. Fourth Mexican International Conference on Computer Science (ENC 2003) 8-12 September 2003, Apizaco, Tlaxcala, México (in press).
68. Muñoz, R.; Saiz-Noeda; M., Suárez; y A., Palomar, M. (2000) Semantic approach to bridging reference resolution. In: Proceedings of the International Conference Machine Translation and Multilingual Applications (MT2000) Exeter, UK.
69. Murata, Masaki y Nagao, Makoto (1996) Indirect Reference in Japanese Sentences. In DAARC96 - Discourse Anaphora and Resolution Colloquium. Edited by Simon Philip Botley and Julia Glass
70. Murata, Masaki y Nagao, Makoto (2000) Indirect reference in Japanese sentences. In: Botley, S., McEnery, T. (eds): Corpus-based and computational approaches to discourse anaphora. John Benjamins, Amsterdam/Philadelphia 211-226
71. Orasan C.; Evans R.; y Mitkov R. (2000) Enhancing Preference-Based Anaphora Resolution with Genetic Algorithms, In Proceedings of NLP'2000, Patras, Greece 185-1
72. Palomar M.; Saiz-Noeda, M.; Muñoz, R.; Suárez, A; Martínez-Barco, P.; y Montoyo, A. (2001) In: Proceedings of Second International Conference, CICLing 2001, Mexico DF, México, 18-24 February. Alexander Gelbukh (Ed.) Lecturer Notes in Computer Science LNCS 2004 Springer 125-139
73. Preuß S.; Schmitz, B.; Hauenschild, C.; y Umbach, U. (1994) Anaphora Resolution in Machine Translation. Studies in Machine Translation and Natural Language Processing. In: Ramm, W.(ed) : (Vol. 6 "Text and content in Machine Translation: Aspects of discourse representation and discourse processing"): Luxembourg: Office for Official Publications of the European Community 29-52
74. Prince, Ellen F. (1981) Toward a taxonomy of given-new information. In Cole (1981:223-55)
75. Rich, E. y LuperFoy S. (1988) An Architecture for Anaphora Resolution. In: Proceedings of the Second Conference on Applied Natural Language Processing (ANLP-2), Austin, Texas, U.S.A. 18-24
76. Rumelhart, David E. (1980) Schemata: the basic building blocks of cognition. In R. Spiro; B. Bruce and W. Brewer (Ed.) Theoretical issues in Reading Comprehension. Erlbaum. Hillsdale, New Jersey, USA
77. Salton, G. (1989) Automatic Text Processing. Addison Wesley. Reading Massachusetts
78. Sanford, Anthony J. y Garrod, Simon C. (1981) Understanding Written Language (Chichester: John Wiley and Sons)

79. Schank, Roger C. y Abelson, Robert P. (1977) Scripts, Plans, Goals and Understanding. Lawrence Erlbaum, Hillsdale, New Jersey.
80. Schmelkes, Corina (1993) Manual de presentación de anteproyectos y trabajos de investigación. Oxford University Press. México
81. Sidner, Candece (1979) Toward a computational theory of definite anaphora comprehension in English. Technical report No. AI-TR-537. MIT Press, Cambridge, Massachussetts
82. Sidner, Candece (1983) Focusing and Discourse. Discourse Processes. 6: 107-30
83. Sidner, Candece (1983) Focusing and Discourse. Discourse Processes. 6: 107-30
84. Sidorov Grigori y Gelbukh, Alexander (1999) Demostrative Pronouns as Markers of Indirect Anaphora.
85. SIL (Summer Institute of Linguistic) Glossary of linguistic terms (en línea) http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/contents.htm
International Linguistics Department. Summer Institute of Linguistics (Dallas, TX)
http://www.sil.org/linguistics/glossary/
Marzo 11, 2003
86. Sowa, John F. (1984) Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley Publishing Company. NY, USA.
87. Sperber, Dan y Wilson, Deirdre (1995) Relevance: Comunication and Cognition. 2nd edition. Basil Blackwell. Oxford
88. Spiegel, M.R. (1976) Probabilidad y Estadística. McGrawHill. México.
89. Tanev, H. y Mitkov, R. (2000) LINGUA - a robust architecture for text processing and anaphora resolution in Bulgarian. In: Proceedings of the International Conference on Machine Translation and Multilingual Applications (MT2000), Exeter, UK. 20.1-20.8.
90. Tetreault, J. R. (1999) Analysis of Syntax-Based Pronoun Resolution Methods. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Maryland, USA. 602-605
91. The American Heritage® Dictionary of the English Language Fourth Edition. 2000.
http://www.bartleby.com/
Marzo 11, 2003
92. Uchida Hiroshi; Zhu Meiying; y Della Senta Tarcisio (1999) The UNL, a Gift for a Millennium http://www.unl.ias.unu.edu/
93. Utrecht institute of Linguistics OTS. Utrecht University.
http://www2.let.uu.nl/UiL-OTS/Lexicon/
Marzo 11, 2003
94. Wu Victor, Manmatha R. y Riseman Edward.(1997) Finding Text in Images in ACM Digital Library, Philadelphia, PA, USA, 1997

ANEXOS

Anexo A: Unidades Léxicas Determinantes

En esta tabla se presentan las unidades léxicas identificadas como determinantes y se lista su categoría (artículo definido, cardinal, etc.). Se marcan con una **x** las funciones adicionales, detectadas y más comunes, que puede cumplir en la oración (ver 4.2.3).

Unidad léxica	Funciones			
	Determinante	**Adjetivo**	**Pronombre**	**Nombre**
1. el	artículo			
2. la	artículo		x	
3. las	artículo		x	
4. lo	artículo		x	
5. los	artículo		x	
6. cientos	cardinal		x	
7. uno, dos,etc.	cardinal		x	
8. miles	cardinal		x	
9. millones	cardinal		x	
10. aquel	demostrativo	X		
11. aquella	demostrativo	X		
12. aquellas	demostrativo	X		
13. aquellos	demostrativo	X		
14. esa	demostrativo	X		
15. esas	demostrativo	X		
16. ese	demostrativo	X		
17. esos	demostrativo	X		
18. esta	demostrativo	X		
19. estas	demostrativo	X		
20. este	demostrativo	X		
21. estos	demostrativo	X		
22. semejante	demostrativo	X		x
23. semejantes	demostrativo	X		x
24. tal	demostrativo			
25. tales	demostrativo			
26. cada	distributivo	X	x	

Unidad léxica	Funciones			
	Determinante	Adjetivo	Pronombre	Nombre
27. sendas	distributivo			x
28. sendos	distributivo			
29. algún	indefinido			
30. alguna	indefinido		x	
31. algunas	indefinido		x	
32. algunos	indefinido		x	
33. ambas	indefinido		x	
34. ambos	indefinido		x	
35. bastante	indefinido	X	x	
36. bastantes	indefinido	X	x	
37. cierta	indefinido	X		
38. ciertas	indefinido	X		
39. cierto	indefinido	X		x
40. ciertos	indefinido	X		
41. cualquier	indefinido			
42. cuanta	indefinido	X	x	
43. cuantas	indefinido	X	x	
44. cuantísima	indefinido	X	x	
45. cuantísimas	indefinido	X	x	
46. cuantísimo	indefinido	X	x	
47. cuantísimos	indefinido	X	x	
48. cuanto	indefinido	X	x	x
49. cuantos	indefinido	X	x	
50. demasiada	indefinido	X	x	
51. demasiadas	indefinido	X	x	
52. demasiado	indefinido	X	x	
53. demasiados	indefinido	X	x	
54. más	indefinido	X	x	
55. menos	indefinido	X	x	
56. mucha	indefinido	X	x	
57. muchas	indefinido	X	x	
58. muchísima	indefinido	X	x	
59. muchísimas	indefinido	X	x	
60. muchísimo	indefinido	X	x	
61. muchísimos	indefinido	X	x	
62. mucho	indefinido	X	x	
63. muchos	indefinido	X	x	
64. ningún	indefinido			
65. ninguna	indefinido		x	
66. ningunas	indefinido		x	
67. ningunos	indefinido		x	
68. otra	indefinido		x	
69. otras	indefinido		x	
70. otro	indefinido		x	

Unidad léxica	Funciones			
	Determinante	Adjetivo	Pronombre	Nombre
71. otros	indefinido		x	
72. poca	indefinido	X	x	
73. pocas	indefinido	X	x	
74. poco	indefinido	X	x	
75. pocos	indefinido	X	x	
76. poquísima	indefinido	X	x	
77. poquísimas	indefinido	X	x	
78. poquísimo	indefinido	X	x	
79. poquísimos	indefinido	X	x	
80. tanta	indefinido	X	x	
81. tantas	indefinido	X	x	
82. tantísima	indefinido	X	x	
83. tantísimas	indefinido	X	x	
84. tantísimo	indefinido	X	x	
85. tantísimos	indefinido	X	x	
86. tanto	indefinido	X	x	X
87. tantos	indefinido	X	x	X
88. toda	indefinido	X	x	
89. todas	indefinido	X	x	
90. todo	indefinido	X	x	X
91. todos	indefinido	X	x	
92. un	indefinido			
93. una	indefinido		x	x
94. unas	indefinido		x	x
95. unos	indefinido		x	x
96. varias	indefinido			
97. varios	indefinido			
98. cuánta	interrogativo		x	
99. cuántas	interrogativo		x	
100. cuánto	interrogativo		x	
101. cuántos	interrogativo		x	
102. qué	interrogativo		x	
103. cuya	posesivo			
104. cuyas	posesivo			
105. cuyo	posesivo			X
106. cuyos	posesivo			X
107. mi	posesivo			
108. mis	posesivo			
109. nuestra	posesivo	X		
110. nuestras	posesivo	X		
111. nuestro	posesivo	X		
112. nuestros	posesivo	X		
113. su	posesivo		x	
114. sus	posesivo		x	

Unidad léxica	Funciones			
	Determinante	Adjetivo	Pronombre	Nombre
115. vuestra	posesivo	X		
116. vuestras	posesivo	X		
117. vuestro	posesivo	X		
118. vuestros	posesivo	X		

Anexo B: Características de documentos usados en los experimentos

Archivo	Palabra	Adj	Adv	Det	Nom	Verb	Pron	Inter	Conj	Prep	Punt	Abrev	Num	Fecha	Desc	UL
a1	2152	174	122	333	510	370	152	2	186	303	382	0	0	7	0	2541
a2	194	20	7	36	55	22	7	0	7	40	22	0	2	0	0	218
a4	204	21	17	35	45	25	16	0	16	29	18	0	0	0	0	222
a10a	912	102	59	152	199	135	71	0	53	141	171	0	1	0	0	1084
a10b	908	74	58	133	203	172	84	0	78	106	180	0	0	0	0	1088
a11a	812	47	65	114	192	150	76	3	55	110	187	0	0	0	0	999
a11b	645	66	42	111	138	107	45	1	50	85	93	0	0	0	0	738
a12	1355	101	82	234	310	187	122	1	112	206	140	0	0	0	0	1495
a13a	984	78	54	164	238	158	84	2	53	153	163	0	2	0	0	1149
a13b	918	58	49	136	220	160	79	1	66	149	181	0	1	0	0	1100
a13c	471	40	35	68	106	81	38	1	26	76	69	0	1	0	0	541
a14	824	58	40	140	198	136	52	0	68	132	134	0	2	1	0	961
a15a	982	70	53	157	272	154	62	3	71	140	186	0	2	7	0	1177
a15b	552	40	27	83	118	102	45	0	45	92	62	0	3	3	0	620
a15c	496	50	30	77	111	76	33	0	34	85	101	0	0	0	0	597
a18	717	39	42	119	150	121	57	0	63	126	74	0	3	0	0	794
a19	95	8	6	17	23	14	8	0	6	13	11	0	0	0	0	106
a20	501	37	25	85	127	68	28	0	43	88	65	0	0	0	0	566
a21a	1361	159	93	210	314	198	74	0	90	223	216	0	11	1	0	1589
a21b	610	66	41	84	142	92	42	0	45	98	85	0	6	4	0	705
a21c	390	43	24	66	89	54	26	0	24	64	53	0	0	5	0	448
a22a	770	50	44	119	157	160	70	2	57	111	130	0	0	0	0	900
a22b	297	26	16	42	63	58	20	0	30	42	36	0	0	0	0	333
a23a	997	123	59	160	229	131	70	0	65	160	130	0	0	0	0	1127
a23b	1229	111	92	202	255	201	98	0	84	186	163	0	0	0	0	1392
a24	2094	134	84	429	545	243	171	0	161	327	338	0	1	1	0	2434
a25a	1688	150	58	319	418	234	89	4	99	317	230	0	2	13	0	1933
a25b	664	35	38	122	159	108	51	0	42	109	84	0	0	0	0	748
a26a	965	81	53	158	216	165	77	0	76	139	121	0	0	0	0	1086
a26b	869	107	28	194	224	96	38	0	60	122	121	0	0	0	0	990
a26c	845	77	59	129	184	151	62	0	68	115	104	0	0	1	0	950
a27	547	37	26	100	132	75	48	0	41	88	57	0	1	2	0	607
a28a	870	60	43	140	213	146	57	0	73	138	157	0	1	1	0	1029
a28b	821	62	43	134	219	116	51	1	53	142	151	0	1	0	0	973
a28c	242	21	14	44	59	35	12	0	17	40	30	0	0	2	0	274
a29	419	40	36	51	89	67	49	0	43	44	71	0	0	0	0	490
a30a	895	44	75	125	196	152	97	12	79	115	301	0	0	0	0	1196
a30b	161	5	12	23	34	35	13	5	15	19	55	0	0	0	0	216
Suma	30456	2514	1751	5045	7152	4755	2274	38	2254	4673	4872	0	40	48	0	35416
Promedio	801.5	66.2	46.1	132.8	188.2	125.1	59.8	1.0	59.3	123.0	128.2	0.0	1.1	1.3	0.0	932.0
Desvstd	467.3	41.2	25.5	85.7	115.7	70.1	36.1	2.2	36.5	74.1	84.4	0.0	2.1	2.7	0.0	545.7

Anexo C: Ejemplo del texto usado para el experimento

6. Cuando escribo esto la Madre_Coraje peruana acaba de ser reventada por los senderistas.
7. Veo su foto en los periódicos: una mujer joven, atractiva, probablemente zamba, esto_es, mestiza de negra e india; oscura de color, en_fin, como son oscuros todos los habitantes de las villas limeñas, arrabales de miseria en donde se hacinan cientos de miles de personas.
8. Son, en su mayoría, indígenas que bajaron de los Andes huyendo del hambre, del atraso y la tuberculosis; quisieron llegar a la ciudad, pero quedaron varados en las afueras, a una decena de kilómetros, en los sórdidos arenales que rodean Lima, en donde plantaron sus chabolas, precarios tenderetes de cartón y cajones astillados.
9. Aspiraban a más: a mejorar su situación, a ser felices; pero les atrapó la miseria suburbana y subhumana, la ferocidad y la violencia de los arrabales.
10. Alguno consigue escapar de allí, muy de_tarde_en_tarde; pero para la inmensa mayoría no hay ni salida ni retorno.
11. En los alrededores de Lima malviven entre chabolas varios millones de personas; sólo en Villa_El_Salvador, el suburbio de la Madre_Coraje, hay 300.000 habitantes.
12. Hace 14 años visité uno de estos asentamientos, y creo recordar que fue precisamente Villa_El_Salvador.
13. Me acompañaba un muchacho que había logrado la proeza de escaparse del barrio; su familia seguía viviendo allí, pero él había conseguido un trabajo y una cama en la ciudad.
14. Yo quería conocer aquello, hacer un reportaje, y él se ofreció a servirme de guía.
15. Fuimos hasta allí, tras un largo trayecto de traqueteantes autobuses, para almorzar con su hermana.
16. La villa había nacido como una excrescencia cancerosa, desordenadamente, allí donde termina el asfalto y la esperanza.
17. Imaginen un pueblo de chabolas de varios cientos de miles de personas: caminas y caminas por los mugrientos arenales y la miseria resulta inacabable, inabarcable.
18. Era media mañana y había bastante gente en las calles, esto_es, en las veredas sin urbanizar que habían quedado abiertas entre las chozas.
19. Mi guía iba evitando, precisamente, los lugares más poblados: avanzábamos por las callejas solitarias, huyendo de la gente, porque te podían rajar en_mitad_de una muchedumbre sin que nadie hiciera nada.
20. La policía no se atrevía a entrar en las villas: eran lugares sin ley, territorios prohibidos.
21. Llegamos al_fin a la casa de la hermana: un tenderete de latas y cartones de apenas tres metros por tres metros, con un infernillo de gas, una silla de enea, un plástico y una manta en un rincón del suelo, sobre la sucia arena.

22. Allí vivían, no sé cómo, la hermana de mi amigo, su marido y cuatro niños.
23. Comimos un arroz con pollo muy sabroso.
24. A mí la invitada de honor, me hicieron sentar en la única silla.
25. Pienso ahora en María_Elena_Moyano, la Madre_Coraje, y recuerdo aquel almuerzo conmovedor, aquel poblado espeluznante.
26. El liderazgo de María_Elena nació de aquella miseria y de una increíble voluntad de superación.
27. De la generosidad, de la inteligencia, del tesón.
28. Consiguió, me dicen, organizar el suburbio, y sacarlo de la violencia y el salvajismo.
29. Construyó un marco de dignidad en el que reconocerse: luchó por otorgar una dimensión humana a unas vidas embrutecidas por la miseria.
30. Por todo esto le dieron el premio Príncipe_de_Asturias_de_la_Paz de 1987: porque logró ser una persona, aunque nada en su entorno se lo permitiese.
31. Y por todo esto ha sido ahora asesinada.
32. Porque los fanáticos de Sendero_Luminoso no pueden admitir que haya seres libres.
33. La verdadera heroicidad no es un acto único: el soldado que se inmola para salvar a sus compañeros, el hombre que entra en un edificio en llamas a rescatar a un niño, son sin_duda personas admirables porque supieron responder a un instante de gloria, de generosidad y de exigencia.
34. Pero ese momento heroico, me parece, es más una especie de fiebre que un talante.
35. Por_el_contrario, la verdadera heroicidad se construye calladamente, día a día, sobreponiéndose una y otra vez a circunstancias dolorosas y extremas.
36. Como hizo María_Elena.
37. Es heroico levantarse todas las mañanas a luchar contra la desesperación y la incultura.
38. Y es heroico sentir miedo ante las amenazas senderistas y seguir actuando año tras año, sin_embargo, como si el propio miedo no existiese.
39. Pero, al_fin, el terror atrapó a María_Elena; fueron a buscarla, y duele imaginar el instante de pánico que debió de sentir ante sus asesinos.
40. ¿Sirve de algo esa muerte salvaje y absurda?
41. ¿Es cierta esa versión cristiana y consoladora que asegura que los sacrificios no son inútiles?
42. Probablemente no; probablemente, en la historia concreta de los suburbios limeños la desaparición de María_Elena sea tan sólo una catástrofe.
43. Pero sí creo que la vida de Moyano, su entereza hasta el final y su coraje, forma parte del legado de los humanos, del inconsciente colectivo, de la sustancia común que todos somos.

44. Y aun cuando la olvidemos, como hemos olvidado a todos los demás héroes anónimos, es gracias_a ella, y a gentes como ella, que la humanidad puede perseverar en el sueño de la felicidad y la razón.

45. Porque ellos nos demuestran que el bien también existe.

Anexo D: Ejemplo del archivo de entrada etiquetado

Cuando cuando CS
escribo escribir VMIP1S0
esto este PD0NS000
la el DA0FS0
Madre_Coraje madre_coraje NP00000
peruana peruano AQ0FS0
acaba acabar VMIP3S0
de de SPS00
ser ser VSN0000
reventada reventar VMP00SF
por por SPS00
los el DA0MP0
senderistas senderista NCCP000
. . Fp
Veo ver VMIP1S0
su su DP3CS0
foto foto NCFS000
en en SPS00
los el DA0MP0
periódicos periódico NCMP000
: : Fd
una uno DI0FS0
mujer mujer NCFS000
joven joven AQ0CS0
, , Fc
atractiva atractivo AQ0FS0
, , Fc
probablemente probablemente RG
zamba zambo AQ0FS0
, , Fc
esto_es esto_es CC
, , Fc
mestiza mestizo AQ0FS0
de de SPS00
negra negro AQ0FS0
e e CC
india indio AQ0FS0
; ; Fx
oscura oscuro AQ0FS0
de de SPS00
color color NCMS000
, , Fc
en_fin en_fin RG
, , Fc
como como CS
son ser VSIP3P0
oscuros oscuro AQ0MP0
todos todo DI0MP0
los el DA0MP0
habitantes habitante NCCP000
de de SPS00
las el DA0FP0
villas villa NCFP000
limeñas limeño AQ0FP0
, , Fc
arrabales arrabal NCMP000
de de SPS00
miseria miseria NCFS000
en en SPS00
donde donde PR000000
se él P0300000
hacinan hacinar VMIP3P0
cientos ciento PN0CP000
de de SPS00
miles mil PN0CP000
de de SPS00
personas persona NCFP000
. . Fp
Son ser VSIP3P0
, , Fc
en en SPS00
su su DP3CS0
mayoría mayoría NCFS000
, , Fc
indígenas indígena NCCP000
que que PR0CN000
bajaron bajar VMIS3P0
de de SPS00
los el DA0MP0
Andes andes NP00000
huyendo huir VMG0000
del del SPCMS
hambre hambre NCFS000
, , Fc

del del SPCMS
atraso atraso NCMS000
y y CC
la el DA0FS0
tuberculosis tuberculosis NCFN000
; ; Fx
quisieron querer VMIS3P0
llegar llegar VMN0000
a a SPS00
la el DA0FS0
ciudad ciudad NCFS000
, , Fc
pero pero CC
quedaron quedar VMIS3P0
varados varado AQ0MPP
en en SPS00
las el DA0FP0
afueras afueras NCFP000
, , Fc
a a SPS00
una uno DI0FS0
decena decena NCFS000
de de SPS00
kilómetros kilómetro NCMP000
, , Fc
en en SPS00
los el DA0MP0
sórdidos sórdido AQ0MP0
arenales arenal NCMP000
que que PR0CN000
rodean rodear VMIP3P0
Lima lima NP00000
, , Fc
en en SPS00
donde donde PR000000
plantaron plantar VMIS3P0
sus su DP3CP0
chabolas chabola NCFP000
, , Fc
precarios precario AQ0MP0
tenderetes tenderete NCMP000
de de SPS00
cartón cartón NCMS000
y y CC
cajones cajón NCMP000
astillados astillado AQ0MPP
. . Fp
Aspiraban aspirar VMII3P0
a a SPS00
más más RG
: : Fd
a a SPS00
mejorar mejorar VMN0000
su su DP3CS0
situación situación NCFS000
, , Fc
a a SPS00
ser ser VSN0000
felices felice AQ0CP0
; ; Fx
pero pero CC
les él PP3CPD00
atrapó atrapar VMIS3S0
la el DA0FS0
miseria miseria NCFS000
suburbana suburbano AQ0FS0
y y CC
subhumana subhumano AQ0FS0
, , Fc
la el DA0FS0
ferocidad ferocidad NCFS000
y y CC
la el DA0FS0
violencia violencia NCFS000
de de SPS00
los el DA0MP0
arrabales arrabal NCMP000
. . Fp
Alguno alguno PI0MS000
consigue conseguir VMIP3S0
escapar escapar VMN0000
de de SPS00
allí allí RG
, , Fc
muy mucho RG
de_tarde_en_tarde de_tarde_en_tarde RG
; ; Fx
pero pero CC
para para SPS00
la el DA0FS0
inmensa inmenso AQ0FS0
mayoría mayoría NCFS000
no no RN
hay haber VAIP3S0
ni ni CC
salida salida NCFS000
ni ni CC
retorno retorno NCMS000
. . Fp
En en SPS00
los el DA0MP0
alrededores alrededor NCMP000
de de SPS00
Lima lima NP00000
malviven malvivir VMIP3P0
entre entre SPS00
chabolas chabola NCFP000
varios varios DI0MP0
millones millón PN0CP000
de de SPS00

personas persona NCFP000
; ; Fx
sólo sólo RG
en en SPS00
Villa_El_Salvador villa_el_salvador NP00000
, , Fc
el el DA0MS0
suburbio suburbio NCMS000
de de SPS00
la el DA0FS0
Madre_Coraje madre_coraje NP00000
, , Fc
hay haber VAIP3S0
300.000 300.000 Z
habitantes habitante NCCP000
. . Fp
Hace hacer VMIP3S0
14 14 Z
años año NCMP000
visité visitar VMIS1S0
uno uno PI0MS000
de de SPS00
estos este DD0MP0
asentamientos asentamiento NCMP000
, , Fc
y y CC
creo creer VMIP1S0
recordar recordar VMN0000
que que CS
fue ser VSIS3S0
precisamente precisamente RG
Villa_El_Salvador villa_el_salvador NP00000
. . Fp
Me yo PP1CS000
acompañaba acompañar VMII3S0
un uno DI0MS0
muchacho muchacho NCMS000
que que PR0CN000
había haber VAII3S0
logrado lograr VMP00SM
la el DA0FS0
proeza proeza NCFS000
de de SPS00
escaparse escapar VMN0000
del del SPCMS
barrio barrio NCMS000
; ; Fx
su su DP3CS0
familia familia NCFS000
seguía seguir VMII3S0
viviendo vivir VMG0000
allí allí RG
, , Fc
pero pero CC
él él PP3MS000
había haber VAII3S0
conseguido conseguir VMP00SM
un uno DI0MS0
trabajo trabajo NCMS000
y y CC
una uno DI0FS0
cama cama NCFS000
en en SPS00
la el DA0FS0
ciudad ciudad NCFS000
. . Fp
Yo yo PP1CSN00
quería querer VMII1S0
conocer conocer VMN0000
aquello aquel PD0NS000
, , Fc
hacer hacer VMN0000
un uno DI0MS0
reportaje reportaje NCMS000
, , Fc
y y CC
él él PP3MS000
se él P0300000
ofreció ofrecer VMIS3S0
a a SPS00
servirme servir VMN0000
de de SPS00
guía guía NCMS000
. . Fp
Fuimos ir VMIS1P0
hasta hasta SPS00
allí allí RG
, , Fc
tras tras SPS00
un uno DI0MS0
largo largo AQ0MS0
trayecto trayecto NCMS000
de de SPS00
traqueteantes traqueteante AQ0CP0
autobuses autobús NCMP000
, , Fc
para para SPS00
almorzar almorzar VMN0000
con con SPS00
su su DP3CS0
hermana hermana NCFS000
. . Fp
La el DA0FS0
villa villa NCFS000
había haber VAII3S0
nacido nacer VMP00SM
como como CS
una uno DI0FS0
excrescencia excrescencia NCFS000
cancerosa canceroso AQ0FS0

, , Fc
desordenadamente desordenadamente RG
, , Fc
allí allí RG
donde donde PR000000
termina terminar VMIP3S0
el el DA0MS0
asfalto asfalto NCMS000
y y CC
la el DA0FS0
esperanza esperanza NCFS000
. . Fp
Imaginen imaginar VMM03P0
un uno DI0MS0
pueblo pueblo NCMS000
de de SPS00
chabolas chabola NCFP000
de de SPS00
varios varios DI0MP0
cientos ciento PN0CP000
de de SPS00
miles mil PN0CP000
de de SPS00
personas persona NCFP000
: : Fd
caminas caminar VMIP2S0
y y CC
caminas caminar VMIP2S0
por por SPS00
los el DA0MP0
mugrientos mugriento AQ0MP0
arenales arenal NCMP000
y y CC
la el DA0FS0
miseria miseria NCFS000
resulta resultar VMIP3S0
inacabable inacabable AQ0CS0
, , Fc
inabarcable inabarcable AQ0CS0
. . Fp
Era ser VSII3S0
media medio DN0FS0
mañana mañana NCFS000
y y CC
había haber VAII3S0
bastante bastante DI0CS0
gente gente NCFS000
en en SPS00
las el DA0FP0
calles calle NCFP000
, , Fc
esto_es esto_es CC
, , Fc
en en SPS00
las el DA0FP0
veredas vereda NCFP000
sin sin SPS00
urbanizar urbanizar VMN0000
que que PR0CN000
habían haber VAII3P0
quedado quedar VMP00SM
abiertas abierto AQ0FPP
entre entre SPS00
las el DA0FP0
chozas choza NCFP000
. . Fp
Mi mi DP1CSS
guía guía NCFS000
iba ir VMII1S0
evitando evitar VMG0000
, , Fc
precisamente precisamente RG
, , Fc
los el DA0MP0
lugares lugar NCMP000
más más RG
poblados poblado AQ0MPP
: : Fd
avanzábamos avanzar VMII1P0
por por SPS00
las el DA0FP0
callejas calleja NCFP000
solitarias solitario AQ0FP0
, , Fc
huyendo huir VMG0000
de de SPS00
la el DA0FS0
gente gente NCFS000
, , Fc
porque porque CS
te tú PP2CS000
podían poder VMII3P0
rajar rajar VMN0000
en_mitad_de en_mitad_de SPS00
una uno DI0FS0
muchedumbre muchedumbre NCFS000
sin sin SPS00
que que CS
nadie nadie PI0CS000
hiciera hacer VMSI3S0
nada nada PI0CS000
. . Fp
La el DA0FS0
policía policía NCCS000
no no RN
se él P0300000
atrevía atreverse VMII3S0
a a SPS00
entrar entrar VMN0000
en en SPS00

las el DA0FP0
villas villa NCFP000
: : Fd
eran ser VSII3P0
lugares lugar NCMP000
sin sin SPS00
ley ley NCFS000
, , Fc
territorios territorio NCMP000
prohibidos prohibido AQ0MPP
. . Fp
Llegamos llegar VMIP1P0
al_fin al_fin RG
a a SPS00
la el DA0FS0
casa casa NCFS000
de de SPS00
la el DA0FS0
hermana hermana NCFS000
: : Fd
un uno DI0MS0
tenderete tenderete NCMS000
de de SPS00
latas lata NCFP000
y y CC
cartones cartón NCMP000
de de SPS00
apenas apenas RG
tres tres DN0CP0
metros metro NCMP000
por por SPS00
tres tres DN0CP0
metros metro NCMP000
, , Fc
con con SPS00
un uno DI0MS0
infernillo infernillo NCMS000
de de SPS00
gas gas NCMS000
, , Fc
una uno DI0FS0
silla silla NCFS000
de de SPS00
enea enea NCFS000
, , Fc
un uno DI0MS0
plástico plástico NCMS000
y y CC
una uno DI0FS0
manta manta NCFS000
en en SPS00
un uno DI0MS0
rincón rincón NCMS000
del del SPCMS
suelo suelo NCMS000
, , Fc
sobre sobre SPS00
la el DA0FS0
sucia sucio AQ0FS0
arena arena NCFS000
. . Fp
Allí allí RG
vivían vivir VMII3P0
, , Fc
no no RN
sé saber VMIP1S0
cómo cómo PT000000
, , Fc
la el DA0FS0
hermana hermana NCFS000
de de SPS00
mi mi DP1CSS
amigo amigo NCMS000
, , Fc
su su DP3CS0
marido marido NCMS000
y y CC
cuatro cuatro DN0CP0
niños niño NCMP000
. . Fp
Comimos comer VMIS1P0
un uno DI0MS0
arroz arroz NCMS000
con con SPS00
pollo pollo NCMS000
muy mucho RG
sabroso sabroso AQ0MS0
. . Fp
A a SPS00
mí yo PP1CSO00
la el DA0FS0
invitada invitada NCFS000
de de SPS00
honor honor NCMS000
, , Fc
me yo PP1CS000
hicieron hacer VMIS3P0
sentar sentar VMN0000
en en SPS00
la el DA0FS0
única único AQ0FS0
silla silla NCFS000
. . Fp
Pienso pensar VMIP1S0
ahora ahora RG
en en SPS00
María_Elena_Moyano maría_elena_moyano NP00000
, , Fc
la el DA0FS0
Madre_Coraje madre_coraje NP00000

, , Fc
y y CC
recuerdo recordar VMIP1S0
aquel aquel DD0MS0
almuerzo almuerzo NCMS000
conmovedor conmovedor AQ0MS0
, , Fc
aquel aquel DD0MS0
poblado poblado NCMS000
espeluznante espeluznante AQ0CS0
. . Fp
El el DA0MS0
liderazgo liderazgo NCMS000
de de SPS00
María_Elena maría_elena NP00000
nació nacer VMIS3S0
de de SPS00
aquella aquel DD0FS0
miseria miseria NCFS000
y y CC
de de SPS00
una uno DI0FS0
increíble increíble AQ0CS0
voluntad voluntad NCFS000
de de SPS00
superación superación NCFS000
. . Fp
De de SPS00
la el DA0FS0
generosidad generosidad NCFS000
, , Fc
de de SPS00
la el DA0FS0
inteligencia inteligencia NCFS000
, , Fc
del del SPCMS
tesón tesón NCMS000
. . Fp
Consiguió conseguir VMIS3S0
, , Fc
me yo PP1CS000
dicen decir VMIP3P0
, , Fc
organizar organizar VMN0000
el el DA0MS0
suburbio suburbio NCMS000
, , Fc
y y CC
sacarlo sacar VMN0000
de de SPS00
la el DA0FS0
violencia violencia NCFS000
y y CC
el el DA0MS0
salvajismo salvajismo NCMS000
. . Fp
Construyó construir VMIS3S0
un uno DI0MS0
marco marco NCMS000
de de SPS00
dignidad dignidad NCFS000
en en SPS00
el el DA0MS0
que que PR0CN000
reconocerse reconocer VMN0000
: : Fd
luchó luchar VMIS3S0
por por SPS00
otorgar otorgar VMN0000
una uno DI0FS0
dimensión dimensión NCFS000
humana humano AQ0FS0
a a SPS00
unas uno DI0FP0
vidas vida NCFP000
embrutecidas embrutecer AQ0FPP
por por SPS00
la el DA0FS0
miseria miseria NCFS000
. . Fp
Por por SPS00
todo todo DI0MS0
esto este PD0NS000
le él PP3CSD00
dieron dar VMIS3P0
el el DA0MS0
premio premio NCMS000
Príncipe_de_Asturias_de_la_Paz
príncipe_de_asturias_de_la_paz NP00000
de de SPS00
1987 [??:??/??/1987:??.??] W
: : Fd
porque porque CS
logró lograr VMIS3S0
ser ser VSN0000
una uno DI0FS0
persona persona NCFS000
, , Fc
aunque aunque CS
nada nada PI0CS000
en en SPS00
su su DP3CS0
entorno entorno NCMS000
se él PP3CN000
lo él PP3MSA00
permitiese permitir VMSI3S0
. . Fp
Y y CC
por por SPS00
todo todo DI0MS0

esto este PD0NS000
ha haber VAIP3S0
sido ser VSP00SM
ahora ahora RG
asesinada asesinar VMP00SF
. . Fp
Porque porque CS
los el DA0MP0
fanáticos fanático NCMP000
de de SPS00
Sendero_Luminoso sendero_luminoso NP00000
no no RN
pueden poder VMIP3P0
admitir admitir VMN0000
que que CS
haya haber VASP3S0
seres ser NCMP000
libres libre AQ0CP0
. . Fp
La el DA0FS0
verdadera verdadero AQ0FS0
heroicidad heroicidad NCFS000
no no RN
es ser VSIP3S0
un uno DI0MS0
acto acto NCMS000
único único AQ0MS0
: : Fd
el el DA0MS0
soldado soldado NCMS000
que que PR0CN000
se él PP3CN000
inmola inmolar VMIP3S0
para para SPS00
salvar salvar VMN0000
a a SPS00
sus su DP3CP0
compañeros compañero NCMP000
, , Fc
el el DA0MS0
hombre hombre NCMS000
que que PR0CN000
entra entrar VMIP3S0
en en SPS00
un uno DI0MS0
edificio edificio NCMS000
en en SPS00
llamas llama NCFP000
a a SPS00
rescatar rescatar VMN0000
a a SPS00
un uno DI0MS0
niño niño NCMS000
, , Fc
son ser VSIP3P0
sin_duda sin_duda RG
personas persona NCFP000
admirables admirable AQ0CP0
porque porque CS
supieron saber VMIS3P0
responder responder VMN0000
a a SPS00
un uno DI0MS0
instante instante NCMS000
de de SPS00
gloria gloria NCFS000
, , Fc
de de SPS00
generosidad generosidad NCFS000
y y CC
de de SPS00
exigencia exigencia NCFS000
. . Fp
Pero pero CC
ese ese DD0MS0
momento momento NCMS000
heroico heroico AQ0MS0
, , Fc
me yo PP1CS000
parece parecer VMIP3S0
, , Fc
es ser VSIP3S0
más más RG
una uno DI0FS0
especie especie NCFS000
de de SPS00
fiebre fiebre NCFS000
que que CS
un uno DI0MS0
talante talante NCMS000
. . Fp
Por_el_contrario por_el_contrario RG
, , Fc
la el DA0FS0
verdadera verdadero AQ0FS0
heroicidad heroicidad NCFS000
se se P0000000
construye construir VMIP3S0
calladamente calladamente RG
, , Fc
día día NCMS000
a a SPS00
día día NCMS000
, , Fc
sobreponiéndose sobreponer VMG0000
una uno DI0FS0
y y CC
otra otro DI0FS0
vez vez NCFS000
a a SPS00

circunstancias circunstancia NCFP000
dolorosas doloroso AQ0FP0
y y CC
extremas extremo AQ0FP0
. . Fp
Como como CS
hizo hacer VMIS3S0
María_Elena maría_elena NP00000
. . Fp
Es ser VSIP3S0
heroico heroico AQ0MS0
levantarse levantar VMN0000
todas todo DI0FP0
las el DA0FP0
mañanas mañana NCFP000
a a SPS00
luchar luchar VMN0000
contra contra SPS00
la el DA0FS0
desesperación desesperación NCFS000
y y CC
la el DA0FS0
incultura incultura NCFS000
. . Fp
Y y CC
es ser VSIP3S0
heroico heroico AQ0MS0
sentir sentir VMN0000
miedo miedo NCMS000
ante ante SPS00
las el DA0FP0
amenazas amenaza NCFP000
senderistas senderista AQ0CP0
y y CC
seguir seguir VMN0000
actuando actuar VMG0000
año año NCMS000
tras tras SPS00
año año NCMS000
, , Fc
sin_embargo sin_embargo CC
, , Fc
como como CS
si si CS
el el DA0MS0
propio propio AQ0MS0
miedo miedo NCMS000
no no RN
existiese existir VMSI3S0
. . Fp
Pero pero CC
, , Fc
al_fin al_fin RG
, , Fc
el el DA0MS0
terror terror NCMS000
atrapó atrapar VMIS3S0
a a SPS00
María_Elena maría_elena NP00000
; ; Fx
fueron ser VSIS3P0
a a SPS00
buscarla buscar VMN0000
, , Fc
y y CC
duele doler VMIP3S0
imaginar imaginar VMN0000
el el DA0MS0
instante instante NCMS000
de de SPS00
pánico pánico NCMS000
que que PR0CN000
debió deber VMIS3S0
de de SPS00
sentir sentir VMN0000
ante ante SPS00
sus su DP3CP0
asesinos asesino NCMP000
. . Fp
¿ ¿ Fia
Sirve servir VMIP3S0
de de SPS00
algo algo PI0CS000
esa ese DD0FS0
muerte muerte NCFS000
salvaje salvaje AQ0CS0
y y CC
absurda absurdo AQ0FS0
? ? Fit
¿ ¿ Fia
Es ser VSIP3S0
cierta cierto AQ0FS0
esa ese DD0FS0
versión versión NCFS000
cristiana cristiano AQ0FS0
y y CC
consoladora consolador AQ0FS0
que que PR0CN000
asegura asegurar VMIP3S0
que que CS
los el DA0MP0
sacrificios sacrificio NCMP000
no no RN
son ser VSIP3P0
inútiles inútil AQ0CP0
? ? Fit
Probablemente probablemente RG
no no RN
; ; Fx
probablemente probablemente RG

, , Fc
en en SPS00
la el DA0FS0
historia historia NCFS000
concreta concreto AQ0FS0
de de SPS00
los el DA0MP0
suburbios suburbio NCMP000
limeños limeño AQ0MP0
la el DA0FS0
desaparición desaparición NCFS000
de de SPS00
María_Elena maría_elena NP00000
sea ser VSSP3S0
tan tanto RG
sólo sólo RG
una uno DI0FS0
catástrofe catástrofe NCFS000
. . Fp
Pero pero CC
sí sí RG
creo creer VMIP1S0
que que CS
la el DA0FS0
vida vida NCFS000
de de SPS00
Moyano moyano NP00000
, , Fc
su su DP3CS0
entereza entereza NCFS000
hasta hasta SPS00
el el DA0MS0
final final NCMS000
y y CC
su su DP3CS0
coraje coraje NCMS000
, , Fc
forma formar VMIP3S0
parte parte NCFS000
del del SPCMS
legado legado NCMS000
de de SPS00
los el DA0MP0
humanos humano NCMP000
, , Fc
del del SPCMS
inconsciente inconsciente NCMS000
colectivo colectivo AQ0MS0
, , Fc
de de SPS00
la el DA0FS0
sustancia sustancia NCFS000
común común AQ0CS0
que que PR0CN000
todos todo PI0MP000
somos ser VSIP1P0
. . Fp
Y y CC
aun aun RG
cuando cuando CS
la él PP3FSA00
olvidemos olvidar VMSP1P0
, , Fc
como como CS
hemos haber VAIP1P0
olvidado olvidar VMP00SM
a a SPS00
todos todo DI0MP0
los el DA0MP0
demás demás DI0CP0
héroes héroe NCMP000
anónimos anónimo AQ0MP0
, , Fc
es ser VSIP3S0
gracias_a gracias_a SPS00
ella él PP3FS000
, , Fc
y y CC
a a SPS00
gentes gente NCFP000
como como CS
ella él PP3FS000
, , Fc
que que CS
la el DA0FS0
humanidad humanidad NCFS000
puede poder VMIP3S0
perseverar perseverar VMN0000
en en SPS00
el el DA0MS0
sueño sueño NCMS000
de de SPS00
la el DA0FS0
felicidad felicidad NCFS000
y y CC
la el DA0FS0
razón razón NCFS000
. . Fp
Porque porque CS
ellos él PP3MP000
nos yo PP1CP000
demuestran demostrar VMIP3P0
que que CS
el el DA0MS0
bien bien NCMS000
también también RG
existe existir VMIP3S0
. . Fp

Anexo E: Ejemplo de la salida del programa

A continuación se muestra la impresión de salida en la detección de correferencia y anáfora indirecta donde los números indican:

0 = Nombre no precedido por determinante (no es expresión referencial)
1 = Nombre propio **sin** correferencia
2 = Nombre propio **con** correferencia
3 = Nombre común **sin** correferencia
4 = Nombre común **con** correferencia
6 = Nombre común **con** anáfora indirecta

Marcado de nombres propios

2 Madre_Coraje
2 Lima
2 Villa_El_Salvador
2 María_Elena
2 María_Elena
2 Moyano
1 Andes
2 Villa_El_Salvador
2 María_Elena_Moyano
1 Príncipe_de_Asturias_de_la_Paz
2 María_Elena
2 Lima
2 Madre_Coraje
2 Madre_Coraje
1 Sendero_Luminoso
2 María_Elena

Correferencias encontradas entre nombres comunes

un largo **trayecto** ↔ servirme de **guía**
Imaginen un **pueblo** ↔ La **villa** había
había bastante **gente** ↔ de **personas**
una **muchedumbre** ↔ de la **gente**
mí la **invitada** ↔ de mi **amigo**
del **tesón** ↔ increíble **voluntad**
dieron el **premio** ↔ de la **violencia**
el **hombre** que ↔ haya **seres** libres
Pero ese **momento** ↔ **instante** de gloria
más una **especie** ↔ **exigencia**
el propio **miedo** ↔ heroico sentir **miedo**
el **terror** ↔ el propio **miedo**
una **catástrofe** ↔ **versión** cristiana
Hasta el **final** ↔ una **catástrofe**
Su **coraje**, ↔ una **catástrofe**
de la **sustancia** ↔ de los **humanos**
la **humanidad** ↔ **gentes** como ella

Marcado de nombres comunes

3	senderistas	3	Foto	3	periódicos	3	Mujer	0	color
3	Habitantes	3	Villas	0	arrabales	0	miseria	0	personas
3	Mayoría	0	Indígenas	3	hambre	3	atraso	3	tuberculosis
3	Ciudad	3	Afueras	3	decena	0	kilómetros	3	arenales
3	Chabolas	0	Tenderetes	0	cartón	0	cajones	3	situación
3	Miseria	3	Ferocidad	3	violencia	3	arrabales	3	mayoría
0	salida	0	Retorno	3	alrededores	0	chabolas	3	personas
3	suburbio	0	Habitantes	0	años	3	asentamientos	3	muchacho
3	proeza	3	Barrio	3	familia	3	trabajo	3	cama
3	ciudad	3	Reportaje	4	guía	4	trayecto	0	autobuses
3	hermana	4	Villa	3	excrescencia	3	asfalto	3	esperanza
4	pueblo	0	Chabolas	4	personas	3	arenales	3	miseria
3	mañana	4	Gente	3	calles	3	veredas	3	chozas
3	Guía	3	Lugares	3	callejas	4	gente	4	muchedumbre
3	policía	3	Villas	0	lugares	0	ley	0	territorios
3	Casa	3	Hermana	3	tenderete	0	latas	0	cartones
3	metros	3	Metros	3	infernillo	0	gas	3	silla
0	Enea	3	Plástico	3	manta	3	rincón	3	suelo
3	Arena	3	Hermana	4	amigo	3	marido	3	niños
3	Arroz	0	Pollo	4	invitada	0	honor	3	silla
3	almuerzo	3	Poblado	3	liderazgo	3	miseria	4	voluntad
0	superación	3	Generosidad	3	inteligencia	4	tesón	3	suburbio
4	violencia	3	Salvajismo	3	marco	0	dignidad	3	dimensión
3	vidas	3	Miseria	4	premio	3	persona	3	entorno
3	fanáticos	4	Seres	3	heroicidad	3	acto	3	soldado
3	compañeros	4	Hombre	3	edificio	0	llamas	3	niño
0	personas	4	Instante	0	gloria	0	generosidad	4	exigencia
4	momento	4	Especie	0	fiebre	3	talante	3	heroicidad
0	día	0	Día	3	vez	0	circunstancias	3	mañanas
3	desesperación	3	Incultura	4	miedo	3	amenazas	0	año
0	año	4	Miedo	4	terror	3	instante	0	pánico
3	asesinos	3	Muerte	4	versión	3	sacrificios	3	historia
3	suburbios	3	desaparición	4	catástrofe	3	vida	3	entereza
4	final	4	Coraje	0	parte	3	legado	4	humanos
3	inconsciente	4	Sustancia	3	héroes	4	gentes	4	humanidad
3	sueño	3	Felicidad	3	razón	3	bien		

Anáforas indirectas encontradas entre nombres

Nº UL	Unidades léxicas	Rel.	Unidades léxicas	Nº UL
23	una mujer	↔	Madre_Coraje peruana	5
73	su mayoría	↔	personas,	67
102	las afueras	↔	ciudad,	95
122	sus chabolas	↔	arenales que	113
149	la miseria	↔	tenderetes de	125
195	de personas	↔	mayoría no	176
218	estos asentamientos	↔	suburbio de	202
231	un muchacho	↔	Villa_El_Salvador .	226
243	su familia	↔	muchacho que	231
286	largo trayecto	↔	guía.	277
295	su hermana	↔	guía.	277
319	un pueblo	↔	villa había	298
328	de personas	↔	villa había	298
336	mugrientos arenales	↔	chabolas de	321
339	la miseria	↔	chabolas de	321
351	bastante gente	↔	personas:	328
372	Mi guía	↔	gente en	351
392	la gente	↔	guía iba	372
416	las villas	↔	muchedumbre sin	400
430	la casa	↔	villas :	416
436	un tenderete	↔	casa de	430
487	mi amigo	↔	hermana :	433
529	aquel almuerzo	↔	Comimos un	495
845	los sacrificios	↔	muerte salvaje	828

Estadísticas

Nombres propios			**Nombres comunes**				**Otras**	
Total	Corref	No corref	Total	Corref	AnaInd	No corref	Pron	Verb
16	10	6	179	31	24	95	53	137

Anexo F: Fuentes de Programas

Sólo se presentan los programas fundamentales que ilustran los algoritmos

```
/*
    Archivo de encabezado clic.h para incluir en programas para el
    procesamiento de archivos etiquetados de LEXESP.

    Desarrollado por: Raúl Morales Carrasco        20020505
      Modificado por: Raúl Morales Carrasco        20020509
*/
#include <exception>    // Para manipular errores en new
#define nl '\n'
#define MAXBUFFER 1024
#define REN 500
#define COL 50

extern char buffer[];

/* Se define la clase de Unidades léxicas (tokens) */
struct Inf_lex {
    char *lex;       // Unidad léxica
    char *lema;      // Lema
    char *cat;       // Clave de codificación de Categoría, tipo, etc
    int   ind;       // indicador de tipo de correferencia
    int ancla;       // identificador único para enlace
    int enlace;      // enlace a la referencia "ancla" anterior
};

struct Dnodo {       // El elemento de la lista
    Inf_lex *Elem;   // apunta a la información del nodo
    Dnodo *Der;      // apunta al siguiente nodo a la derecha
    Dnodo *Izq;      // apunta al siguiente nodo a la izquierda
};
```

```
class Dlista {              // Lista de objetos nodo
   Dnodo *Dactual;          // apunta al nodo actual de la lista
   Dnodo *Dinicial;         // apunta al nodo inicial de la lista
   Dnodo *Dfinal;           // apunta al ultimo nodo insertado
   int Dnodos;              // Numero de Nodos en la lista
public:
   Dlista();                                    // constructor
   ~Dlista();                                   // destructor
   void DInserta(Inf_lex *Nuevo);               // agrega un elemento a la lista
   Dnodo * Inicio(){ return Dinicial;};         // Obtiene el nodo inicial de la lista
   Dnodo * Fin(){ return Dfinal;};              // Obtiene el nodo final de la lista
   Dnodo * Actual(){ return Dactual;};          // Obtiene el nodo actual de la lista
   void Dlistado();                             // Listado de elementos
};

/*
      Funciones para manipulación de estructura Inf_lex
*/
int verifica (char * clave);
void mostrar(Inf_lex *);
int semejanza(char * cadena1, char *cadena2);
int coref ( char * archivo, char * cadena1, char * cadena2);
int anaf_ind ( char * archivo, char * cadena1, char * cadena2 );
int cadenas ( char *cadena , const char* archivo, char cad[REN][COL]);
```

```
/*  Programa principal para el procesamiento de
     archivos etiquetados de Clic-TALP.

    Se corre así:

        anaind archivo_entrada bandera contador

    Donde:
            Archivo = archivo etiquetado a procesar
            Bandera = Categoría base para búsqueda hacia atrás
            Contador = Límite de búsqueda hacia atrás

    Desarrollado por: Raúl Morales Carrasco        20020205
      Modificado por: Raúl Morales Carrasco        20020509
*/
#include <iostream.h>
#include <string.h>
#include <fstream.h>

  Dnodo *derecha;
  Dnodo *izquierda;
  Dnodo *prev_izq;
  Dnodo *prev_der;

  Inf_lex *Dat1;
  int Ancla = 1;
  char arreglo[REN][COL];
  char buffer[MAXBUFFER];
  Dlista texto;         // Se declara texto como una variable de lista

int main ( int argc, char* argv[])
{
    char *p1, *p2, *p3;
    int indice, linea = 0, temp;
    int k, l, existe, contador;
    int Tprop, Cprop, Nprop, Tcom, Ccom, Acom, Ncom, Nverb, Npron;
    ifstream f_in;
    if ( argc < 4 ) {
        cerr << " Se debe correr asi:" << nl;
        cerr << " anaind archivo_entrada bandera contador" << endl;
        return(1);
    }
    else {
        cerr << " Corriendo con archivo " << argv[1] << "  "
             << argv[2] << "  " << argv[3] << endl;
    }
```

```
    contador = atoi(argv[3]);           // Contador de búsqueda hacia atrás
    f_in.open(argv[1]);
    if (!f_in ){                // error: no se pudo abrir el archivo de entrada
        cerr << " No se pudo abrir el archivo" << argv[1] << endl;
        return(2);
    }
/*
    Se lee archivo de entrada hasta el fin de archivo
*/
    while ( f_in.getline(buffer,MAXBUFFER, nl) ) {
        linea++;
        p1 = strtok(buffer, " "); // Se obtiene el token o lex
        if( (p2 = strtok(NULL, " ")) == NULL) { // lema
            cerr << " Linea " << linea << " sin lema " <<  p1 << nl;
            continue;
        }
        if( (p3 = strtok(NULL, " ")) == NULL) { // categoria
            cerr << " Linea " << linea << " sin categoria "
            <<  p1 << "  " << p2 << nl;
            continue;
        }
        indice = verifica ( p3 );
        if ( indice > 0 ) {          // Si la clave es incorrecta no se toma
            cerr << " Clave MAL " <<  indice  << "  " << p1
                 << "  " <<  p2 << "  " << p3 << nl;
        }
        else {          // Si es clave correcta se almacena en memoria
            try {           // Se prueban excepciones al pedir memoria
              Dat1 = new Inf_lex;
            }
            catch (std::bad_alloc) {  // verifica el error de new
            cerr << " No hay Memoria suficiente Inf_lex " << endl;
            exit(-1);
        }
         Dat1->lex = new char[strlen(p1)+1];
         strcpy(Dat1->lex, p1);
         Dat1->lema = new char[strlen(p2)+1];
         strcpy(Dat1->lema, p2);
         Dat1->cat = new char[strlen(p3)+1];
         strcpy(Dat1->cat, p3);
          Dat1->ind = 0;
          Dat1->ancla = Ancla++;
          Dat1->enlace = 0;
          texto.DInserta(Dat1);   // Se inserta en la lista
        }
    }
```

```
f_in.close ();
// Almacenado el archivo en Memoria se aplican los algoritmos
//
// Se marca correferencia de nombres propios
cerr << " Marcado de correferencia de Nombres propios " << endl;
derecha = texto.Inicio();
while( derecha != NULL)  {      // de inicio hacia final
    if( (derecha->Elem->cat[0] == 'N') &&    // Nombre
        (derecha->Elem->cat[1] == 'P') &&    // Propio
        (derecha->Elem->ind    ==  0 ) )  {      // sin marcar
        // se inicializa variable arreglo
        for ( k = 0; k < REN; k++ )
            for ( l = 0; l < COL; l++) arreglo[k][l] = 0;
        derecha->Elem->ind = 1;                  // Se marca la primera ocurrencia
        derecha->Elem->enlace = 0;
        existe = cadenas ( derecha->Elem->lex, "dic\\sinompro.txt", arreglo);
        // se verifica correferencia entre nombres propios
        izquierda = texto.Fin();
        while( izquierda != derecha->Der )  {
         if( (izquierda->Elem->cat[0] == 'N') &&     // nombre propio
               (izquierda->Elem->cat[1] == 'P') &&  // sin marcar
               (izquierda->Elem->ind    ==  0 ) )  {
               if ( existe ) {
               // Se compara lex con cadenas
               for ( k = 0; k <= existe; k++ ){
                        temp = strcmp(arreglo[k], izquierda->Elem->lex);
                        if ( temp == 0) break;
               }
            }
            else {   // o directo con lex
                temp = strcmp(derecha->Elem->lex, izquierda->Elem->lex);
            }
         // si se encuentra se marcan como correferencia
         if ( temp == 0){
            izquierda->Elem->ind = 2;
            izquierda->Elem->enlace = derecha->Elem->ancla;
             cout << izquierda->Elem->lex << "\t"
                    << izquierda->Elem->ancla << " -> "
                    << derecha->Elem->lex << "\t"
                    << derecha->Elem->ancla << endl;
         }
        }
        izquierda = izquierda->Izq; // apunta al siguiente nodo
        }
    }
    derecha = derecha->Der;   // apunta al siguiente nodo
```

```
    };
//          Se identifican las expresiones referenciales marcando
//          los nombres comunes precedidos por determinantes
//
    derecha = texto.Inicio();
    cerr << " Marcado de Nombres comunes" << endl;
    while( derecha != NULL)  {      // de izquierda a derecha
       temp = strcmp(derecha->Elem->cat, "SPCMS");
       // localiza determinativo o preposición contraída (al, del)
       if( (derecha->Elem->cat[0] == 'D') || (temp == 0) ) {
          while( derecha != NULL )  {                 // se busca nombre común
           if( (derecha->Elem->cat[0] == 'N') &&
               (derecha->Elem->cat[1] == 'C') )  {
                 derecha->Elem->ind = 3;              // y se marca
           break;
         }
         // el ciclo se interrumpe si hay ...
         else if( derecha->Elem->cat[0]== 'V') break;      // Verbo
         else if( derecha->Elem->cat[0]== 'F') break;      // Puntuación
         else if( derecha->Elem->cat[0]== 'C') break;      // Conjunción
              else derecha = derecha->Der;                 // de otra forma continúa
       }
       }
       derecha = derecha->Der;   // apunta al siguiente nodo
    };
    // Se localiza posición después de la primera oración completa sea
    // simple, compuesta o compleja (hasta primer punto y seguido)
    derecha = texto.Inicio();
    while (derecha != NULL )  {                   // de inicio hacia final
       if( derecha->Elem->cat[0] == 'F') break;
       else derecha = derecha->Der;               // apunta al siguiente nodo
    };
    //
    // Se verifica relación de sinonímia con el diccionario
    // para cada componente léxico marcado
    //
    cout << "Búsqueda de coreferencias nombres comunes " << endl;
    prev_der = derecha;
    while( derecha != texto.Fin() )  {    // hasta el fin de lista
       if( (derecha->Elem->cat[0] == 'N') &&       // nombre común
           (derecha->Elem->cat[1] == 'C') &&       // marcado
           (derecha->Elem->ind    ==  3 ) )  {
           // Se localiza posición antes de esta oración (del punto y seguido)
           izquierda = derecha->Izq;
           while (izquierda != NULL )  {         // de derecha a izquierda
            if( izquierda->Elem->cat[0]== 'F') break;
```

```
            else izquierda = izquierda->Izq;
          };
        // Se recorre de derecha a izquierda hasta encontrar relaciones
        indice = 0;
        prev_izq = izquierda;
         while ( izquierda != NULL && indice < contador )  {
            if (  izquierda->Elem->cat[0] == argv[2][0]) indice++;
            if (  izquierda->Elem->cat[0] == 'N' &&
                  izquierda->Elem->cat[1] == 'C') {
            // se verifica correferencia
             k = coref(argv[1], derecha->Elem->lema, izquierda->Elem->lema);
            if(  k == 1 ) { // se marcan los coreferentes
              derecha->Elem->ind = 4;
              derecha->Elem->enlace = izquierda->Elem->ancla;
              indice = contador;
              cout
              << derecha->Elem->ancla << " "
              << prev_der->Elem->lex << " "
              << derecha->Elem->lex << " --> "
              << izquierda->Elem->lex << " "
              << prev_izq->Elem->lex << " "
              << izquierda->Elem->ancla
              << endl;
              break;
            }
            }
           prev_izq = izquierda;
           izquierda = izquierda->Izq; // apunta al siguiente nodo
           }  // termina while de derecha a izquierda
      }  // fin de if marcado
        prev_der = derecha;
        derecha = derecha->Der;   // apunta al siguiente nodo
    }  // termina while correferencia
/*
     Resolución de anáfora indirecta
*/
// Se localiza posición después de la primera oración completa sea
// simple, compuesta o compleja (hasta primer punto y seguido)
    derecha = texto.Inicio();
    while (derecha != NULL )  {        // de inicio hacia final
        if( derecha->Elem->cat[0] == 'F') break;
        else derecha = derecha->Der; // apunta al siguiente nodo
    };
// Se verifica relación con nombres comunes del diccionario
// para cada componente léxico marcado
    cout << endl << "Búsqueda de anáfora indirecta " << endl;
```

```
   prev_der = derecha;
    while( derecha != texto.Fin() )  {                // hasta el fin de lista
        if( (derecha->Elem->cat[0] == 'N') &&    // nombre comun
         (derecha->Elem->cat[1] == 'C') &&       // marcado
         (derecha->Elem->ind    ==  3 ) )  {     //  sin correferencia
        // Se localiza posición antes de esta oración (del punto y seguido)
          izquierda = derecha->Izq;
          while (izquierda != NULL )  {
              if( izquierda->Elem->cat[0]== 'F') break;
              else izquierda = izquierda->Izq;
          };
        // Se recorre de derecha a izquierda hasta encontrar relaciones
        indice = 0;
        prev_izq = izquierda;
          while ( izquierda != NULL && indice < contador )  {
              if (  izquierda->Elem->cat[0] == argv[2][0]) indice++;
              if (  izquierda->Elem->cat[0] == 'N' ||
                     (izquierda->Elem->cat[0] == 'V' &&
                       izquierda->Elem->cat[1] == 'M' ) ) {
              // se verifica relacion anaforica
              k = anaf_ind(argv[1],derecha->Elem->lema, izquierda->Elem->lema);
              if(  k == 1 ) { // se marcan los coreferentes por anaind
                 derecha->Elem->ind = 6;
                 derecha->Elem->enlace = izquierda->Elem->ancla;
                 indice = contador;
                 cout
                 << derecha->Elem->ancla << " "
                 << prev_der->Elem->lex << " "
                 << derecha->Elem->lex << " --> "
                 << izquierda->Elem->lex << " "
                 << prev_izq->Elem->lex << " "
                 << izquierda->Elem->ancla
                 << endl;
                 break;
              }
              }
            prev_izq = izquierda;
            izquierda = izquierda->Izq; // apunta al siguiente nodo
            }  // termina while de derecha a izquierda
      }  // fin de if marcado
      prev_der = derecha;
         derecha = derecha->Der;   // apunta al siguiente nodo
     }  // termina while anafora indirecta
//
// Se imprimen nombres y cuentan ocurrencias de cada fenómeno
//
```

```
    derecha = texto.Inicio();
    cout << nl << "Impresion de marcado de nombres  " << endl;
   Tcom = Ccom = Ncom = Acom = Nverb = Npron = 0;
   Tprop = Cprop = Nprop = 0;
    while ( derecha != NULL )  { // de izquierda a derecha
        if ( derecha->Elem->cat[0] == 'N' ) {
          if ( derecha->Elem->cat[1] == 'P')   {
          if (derecha->Elem->ind == 2 ) Cprop++;
          else if (derecha->Elem->ind == 1 )Nprop++;
          Tprop++;
        } else {
          if (derecha->Elem->ind == 4 ) Ccom++;
          else if (derecha->Elem->ind == 6 )Acom++;
          else if (derecha->Elem->ind == 3 )Ncom++;
          else Tcom++;
        }
              cout
           << derecha->Elem->ind << " "
           << derecha->Elem->lex << " "
           << derecha->Elem->ancla;
          if ( derecha->Elem->enlace )
          cout
           << " -> "
           << derecha->Elem->enlace;
           cout << endl;
      }
      else if ( derecha->Elem->cat[0] == 'V' )Nverb++;
      else if ( derecha->Elem->cat[0] == 'P' )Npron++;
      derecha = derecha->Der;  // apunta al siguiente nodo
    }
   cout << endl;
//
//          Se imprime estadística del proceso
//
cout << "\tPROPIOS\t|COMUNES\t\t\t\t|OTROS" << endl;
cout << "Total\tCorref\tNoCorr\tTotal\tCorref\tAnaInd\tNoCorr\tNmark\tpron\tverb"
     << endl;
cout << Tprop << "\t" << Cprop << "\t" << Nprop
     << "\t" << (Tcom+Ccom+Acom+Ncom)  << "\t" << Ccom  << "\t"
     << Acom  << "\t" << Ncom  << "\t"
     << Tcom  << "\t" << Npron << "\t" << Nverb << endl;
    cerr << " Terminado OK" << endl;
    return(0);
}
```

```
/* Función para buscar la posible "semejanza" conceptual por medio de
    relaciones entre sinónimos
    p2 es el posible referente       y       p3 es el posible referido
    Desarrollado por: Raúl Morales Carrasco         20020604
     Modificado por: Raúl Morales Carrasco         20020604
*/
#include <iostream.h>
#include <fstream.h>
#include <string.h>
#include "clic.h"
/*
     Calcula semejanza aproximada entre dos palabras
*/
int semejanza(char *p2, char* p3){
  float porciento = 0.0, lonp;
  int lon1, lon2, i, conta=0;
  lon1 = strlen(p2);
  lon2 = strlen(p3);
  // promedio de longitudes de cadena
  lonp = ((float)lon1 + (float)lon2) / 2.0;
  // se toma la menor longitud de cadena para comparar
  if ( lon2 < lon1) lon1 = lon2;
  for ( i = 0; i < lon1; i++) {
    if ( p2[i] == p3[i]) conta++;
    else break;
  }
  if ( conta )porciento = (float)conta/lonp;
  if ( porciento > 0.90 ) return(1);
  else return (0);
}
/*
       Función para verificar la correferencia entre sinónimos
       p1 es el posible referente   y   p2 es el posible referido

    Desarrollado por: Raúl Morales Carrasco         20020604
     Modificado por: Raúl Morales Carrasco         20020604
*/
int coref ( char * p1, char * p2 )
{
  FILE* fout;
  char cad1[REN][COL];
  char cad2[REN][COL];
  int i = 0, j = 0, k, l, semeja = 0;
  // se inicializan variables de cadenas
  for ( k = 0; k < REN; k++ )
    for ( l = 0; l < COL; l++) {
```

```
        cad1[k][l] = 0;
        cad2[k][l] = 0;
      }
   // se espera que el diccionario sea único
   i = cadenas ( p1, "sinomcom", cad1 );
   j = cadenas ( p2, "sinomcom", cad2 );
   if ( (i*j) == 0 ) return (semeja);            // si falta alguna no hay que comparar
   // Se comparan aproximadamente todas las cadenas de p1 y las de p2
   for ( k = 0; k <= i; k++ ) {
      for ( l = 0; l <= j; l++) {
        semeja = semejanza(cad1[k], cad2[l]);
        if ( semeja ) return (semeja);
      }
   }
   return(semeja);
}
/*
       Obtiene arreglo de sinónimos del diccionario
        para poder verificar la correferencia
*/
int cadenas ( char * p1, const char* archivo, char cad1[REN][COL])
{
   ifstream f_in;
   char * p3;
   int i = 0, j;
    f_in.open( archivo );         // se abre diccionario
    if (!f_in ){
        cerr << " No se pudo abrir el diccionario" << endl;
        return(i);
    }
   // desde el inicio de archivo
    while(f_in.getline(buffer,MAXBUFFER, nl)) {
      p3 = strtok(buffer, " ");              // Se obtiene entrada en diccionario
     j = strcmp(p1, p3);
     if ( j == 0) {                          // Es igual. se almacenan cadenas
        strcpy (cad1[i], p3);
          while((p3 = strtok(NULL, " "))!= NULL) {
          strcpy (cad1[i], p3);
          i++;
        }
     }
   } // fin while lectura archivo
   f_in.close ();
   return(i);
}
```

Anexo G: Resultados para determinar el tamaño de ventana

En este anexo se presentan las tablas de resultados que permitieron analizar y determinar el tamaño de la ventana de búsqueda hacia atrás. En las columnas se utilizaron las siguientes abreviaturas:

ban = tipo de **ban**dera a evaluar
vent = tamaño de **vent**ana utilizado
cd = **c**orreferencia **d**irecta
ci = **c**orreferencia **i**ndirecta
ai = **a**náfora **i**ndirecta
det = **det**ectados (suma de cd, ci, y ai)
ndet = **n**o **det**ectados (nuevas referencias o falla del método)
nr = **n**o **r**eferenciados (nombres no precedidos por determinante)
com = total de nombres **com**unes marcados en el texto.
CP = **C**orreferencia entre nombres **P**ropios
NCP = **N**o **C**orreferencia entre nombres **P**ropios
Prop = total de nombres **Prop**ios marcados enel texto
Verb = total de **Verb**os marcados enel texto
Pron = total de **Pron**ombres marcados enel texto

Resultados del archivo Contra la guerra													
		Nombres comunes							Nom. Propios				
ban	**vent**	**cd**	**ci**	**ai**	**det**	**ndet**	**nr**	**com**	**CP**	**NCP**	**Prop**	**Verb**	**Pron**
F	2	2	0	3	5	25	5	35	0	2	2	44	12
F	3	2	1	3	6	24	5	35	0	2	2	44	12
F	4	2	1	3	6	24	5	35	0	2	2	44	12
F	5	4	4	3	11	19	5	35	0	2	2	44	12
F	6	5	4	2	11	19	5	35	0	2	2	44	12
F	7	5	4	2	11	19	5	35	0	2	2	44	12
F	8	6	5	2	13	17	5	35	0	2	2	44	12
F	9	6	5	2	13	17	5	35	0	2	2	44	12
F	10	6	5	2	13	17	5	35	0	2	2	44	12
F	11	6	5	3	14	16	5	35	0	2	2	44	12
F	12	6	5	3	14	16	5	35	0	2	2	44	12
F	13	6	5	3	14	16	5	35	0	2	2	44	12
F	14	6	5	3	14	16	5	35	0	2	2	44	12
F	15	6	5	3	14	16	5	35	0	2	2	44	12
F	16	6	5	3	14	16	5	35	0	2	2	44	12
F	17	6	5	3	14	16	5	35	0	2	2	44	12
F	18	6	6	3	15	15	5	35	0	2	2	44	12
F	19	6	6	3	15	15	5	35	0	2	2	44	12
F	20	6	6	3	15	15	5	35	0	2	2	44	12

Resultados del archivo Contra la guerra													
	Nombres comunes								Nom. Propios				
ban	vent	cd	ci	ai	det	ndet	nr	com	CP	NCP	Prop	Verb	Pron
F	20	6	6	3	15	15	5	35	0	2	2	44	12
F	21	6	6	3	15	15	5	35	0	2	2	44	12
F	22	6	7	3	16	14	5	35	0	2	2	44	12
F	23	6	7	3	16	14	5	35	0	2	2	44	12
F	24	6	7	3	16	14	5	35	0	2	2	44	12
F	25	6	7	3	16	14	5	35	0	2	2	44	12
F	26	6	7	3	16	14	5	35	0	2	2	44	12
F	27	6	7	3	16	14	5	35	0	2	2	44	12
F	28	6	7	3	16	14	5	35	0	2	2	44	12
F	29	6	7	3	16	14	5	35	0	2	2	44	12
F	30	6	7	3	16	14	5	35	0	2	2	44	12
N	2	1	1	3	5	25	5	35	0	2	2	44	12
N	3	2	1	3	6	24	5	35	0	2	2	44	12
N	4	2	3	3	8	22	5	35	0	2	2	44	12
N	5	3	4	3	10	20	5	35	0	2	2	44	12
N	6	4	4	2	10	20	5	35	0	2	2	44	12
N	7	5	4	2	11	19	5	35	0	2	2	44	12
N	8	5	4	2	11	19	5	35	0	2	2	44	12
N	9	5	4	2	11	19	5	35	0	2	2	44	12
N	10	6	4	3	13	17	5	35	0	2	2	44	12
N	11	6	5	3	14	16	5	35	0	2	2	44	12
N	12	6	5	3	14	16	5	35	0	2	2	44	12
N	13	6	5	3	14	16	5	35	0	2	2	44	12
N	14	6	5	3	14	16	5	35	0	2	2	44	12
N	15	6	5	3	14	16	5	35	0	2	2	44	12
N	16	6	5	3	14	16	5	35	0	2	2	44	12
N	17	6	5	3	14	16	5	35	0	2	2	44	12
N	18	6	6	3	15	15	5	35	0	2	2	44	12
N	19	6	6	3	15	15	5	35	0	2	2	44	12
N	20	6	6	3	15	15	5	35	0	2	2	44	12
N	21	6	6	3	15	15	5	35	0	2	2	44	12
N	22	6	6	3	15	15	5	35	0	2	2	44	12
N	23	6	7	3	16	14	5	35	0	2	2	44	12
N	24	6	7	3	16	14	5	35	0	2	2	44	12
N	25	6	7	3	16	14	5	35	0	2	2	44	12
N	26	6	7	3	16	14	5	35	0	2	2	44	12
N	27	6	7	3	16	14	5	35	0	2	2	44	12
N	28	6	7	3	16	14	5	35	0	2	2	44	12
N	29	6	7	3	16	14	5	35	0	2	2	44	12
N	30	6	7	3	16	14	5	35	0	2	2	44	12
P	2	2	3	3	8	22	5	35	0	2	2	44	12
P	3	3	3	2	8	22	5	35	0	2	2	44	12

Resultados del archivo Contra la guerra													
	Nombres comunes								Nom. Propios				
ban	vent	cd	ci	ai	det	ndet	nr	com	CP	NCP	Prop	Verb	Pron
P	4	4	4	3	11	19	5	35	0	2	2	44	12
P	5	5	6	3	14	16	5	35	0	2	2	44	12
P	6	6	7	3	16	14	5	35	0	2	2	44	12
P	7	6	7	3	16	14	5	35	0	2	2	44	12
P	8	6	7	3	16	14	5	35	0	2	2	44	12
P	9	6	7	3	16	14	5	35	0	2	2	44	12
P	10	6	7	3	16	14	5	35	0	2	2	44	12
P	11	6	7	3	16	14	5	35	0	2	2	44	12
P	12	6	7	3	16	14	5	35	0	2	2	44	12
P	13	6	7	3	16	14	5	35	0	2	2	44	12
P	14	6	7	3	16	14	5	35	0	2	2	44	12
P	15	6	7	3	16	14	5	35	0	2	2	44	12
P	16	6	7	3	16	14	5	35	0	2	2	44	12
P	17	6	7	3	16	14	5	35	0	2	2	44	12
P	18	6	7	3	16	14	5	35	0	2	2	44	12
P	19	6	7	3	16	14	5	35	0	2	2	44	12
P	20	6	7	3	16	14	5	35	0	2	2	44	12
P	21	6	7	3	16	14	5	35	0	2	2	44	12
P	22	6	7	3	16	14	5	35	0	2	2	44	12
P	23	6	7	3	16	14	5	35	0	2	2	44	12
P	24	6	7	3	16	14	5	35	0	2	2	44	12
P	25	6	7	3	16	14	5	35	0	2	2	44	12
P	26	6	7	3	16	14	5	35	0	2	2	44	12
P	27	6	7	3	16	14	5	35	0	2	2	44	12
P	28	6	7	3	16	14	5	35	0	2	2	44	12
P	29	6	7	3	16	14	5	35	0	2	2	44	12
P	30	6	7	3	16	14	5	35	0	2	2	44	12
V	2	2	0	1	3	27	5	35	0	2	2	44	12
V	3	2	3	3	8	22	5	35	0	2	2	44	12
V	4	2	3	3	8	22	5	35	0	2	2	44	12
V	5	2	3	3	8	22	5	35	0	2	2	44	12
V	6	4	4	3	11	19	5	35	0	2	2	44	12
V	7	4	4	3	11	19	5	35	0	2	2	44	12
V	8	4	4	3	11	19	5	35	0	2	2	44	12
V	9	5	4	2	11	19	5	35	0	2	2	44	12
V	10	5	4	2	11	19	5	35	0	2	2	44	12
V	11	6	4	2	12	18	5	35	0	2	2	44	12
V	12	6	4	2	12	18	5	35	0	2	2	44	12
V	13	6	4	2	12	18	5	35	0	2	2	44	12
V	14	6	5	2	13	17	5	35	0	2	2	44	12
V	15	6	5	2	13	17	5	35	0	2	2	44	12
V	16	6	5	2	13	17	5	35	0	2	2	44	12

Resultados del archivo Contra la guerra													
		Nombres comunes							Nom. Propios				
ban	**vent**	**cd**	**ci**	**ai**	**det**	**ndet**	**nr**	**com**	**CP**	**NCP**	**Prop**	**Verb**	**Pron**
V	17	6	5	2	13	17	5	35	0	2	2	44	12
V	18	6	5	3	14	16	5	35	0	2	2	44	12
V	19	6	5	3	14	16	5	35	0	2	2	44	12
V	20	6	5	3	14	16	5	35	0	2	2	44	12
V	21	6	5	3	14	16	5	35	0	2	2	44	12
V	22	6	5	3	14	16	5	35	0	2	2	44	12
V	23	6	5	3	14	16	5	35	0	2	2	44	12
V	24	6	5	3	14	16	5	35	0	2	2	44	12
V	25	6	5	3	14	16	5	35	0	2	2	44	12
V	26	6	6	3	15	15	5	35	0	2	2	44	12
V	27	6	6	3	15	15	5	35	0	2	2	44	12
V	28	6	7	3	16	14	5	35	0	2	2	44	12
V	29	6	7	3	16	14	5	35	0	2	2	44	12
V	30	6	7	3	16	14	5	35	0	2	2	44	12

Resultados del archivo Instituto Oriente													
		Nombres comunes							Nom. Propios				
ban	vent	cd	ci	ai	cor	det	nr	com	CP	NCP	Prop	Verb	Pron
F	2	3	1	2	6	45	21	72	8	34	42	32	9
F	3	3	3	3	9	42	21	72	8	34	42	32	9
F	4	3	3	4	10	41	21	72	8	34	42	32	9
F	5	7	4	5	16	35	21	72	8	34	42	32	9
F	6	8	4	5	17	34	21	72	8	34	42	32	9
F	7	8	4	5	17	34	21	72	8	34	42	32	9
F	8	8	4	6	18	33	21	72	8	34	42	32	9
F	9	9	5	6	20	31	21	72	8	34	42	32	9
F	10	9	9	4	22	29	21	72	8	34	42	32	9
F	11	9	10	4	23	28	21	72	8	34	42	32	9
F	12	9	11	3	23	28	21	72	8	34	42	32	9
F	13	10	11	3	24	27	21	72	8	34	42	32	9
F	14	10	11	3	24	27	21	72	8	34	42	32	9
F	15	11	11	3	25	26	21	72	8	34	42	32	9
F	16	11	11	3	25	26	21	72	8	34	42	32	9
F	17	11	11	3	25	26	21	72	8	34	42	32	9
F	18	11	11	3	25	26	21	72	8	34	42	32	9
F	19	11	11	3	25	26	21	72	8	34	42	32	9
F	20	12	13	3	28	23	21	72	8	34	42	32	9
F	20	12	13	3	28	23	21	72	8	34	42	32	9
F	21	12	13	3	28	23	21	72	8	34	42	32	9
F	22	12	13	3	28	23	21	72	8	34	42	32	9
F	23	12	13	3	28	23	21	72	8	34	42	32	9
F	24	12	13	3	28	23	21	72	8	34	42	32	9
F	25	12	13	3	28	23	21	72	8	34	42	32	9
F	26	12	13	3	28	23	21	72	8	34	42	32	9
F	27	12	13	3	28	23	21	72	8	34	42	32	9
F	28	13	13	3	29	22	21	72	8	34	42	32	9
F	29	13	13	3	29	22	21	72	8	34	42	32	9
F	30	13	13	3	29	22	21	72	8	34	42	32	9
N	2	3	0	1	4	47	21	72	8	34	42	32	9
N	3	4	1	1	6	45	21	72	8	34	42	32	9
N	4	4	2	1	7	44	21	72	8	34	42	32	9
N	5	5	2	4	11	40	21	72	8	34	42	32	9
N	6	7	2	4	13	38	21	72	8	34	42	32	9
N	7	7	2	5	14	37	21	72	8	34	42	32	9
N	8	7	2	7	16	35	21	72	8	34	42	32	9
N	9	7	2	7	16	35	21	72	8	34	42	32	9
N	10	7	3	7	17	34	21	72	8	34	42	32	9
N	11	7	3	7	17	34	21	72	8	34	42	32	9
N	12	7	5	7	19	32	21	72	8	34	42	32	9

Resultados del archivo Instituto Oriente													
		Nombres comunes							Nom. Propios				
ban	**vent**	**cd**	**ci**	**ai**	**cor**	**det**	**nr**	**com**	**CP**	**NCP**	**Prop**	**Verb**	**Pron**
N	13	7	6	6	19	32	21	72	8	34	42	32	9
N	14	7	9	3	19	32	21	72	8	34	42	32	9
N	15	7	9	3	19	32	21	72	8	34	42	32	9
N	16	7	9	3	19	32	21	72	8	34	42	32	9
N	17	8	9	3	20	31	21	72	8	34	42	32	9
N	18	8	10	3	21	30	21	72	8	34	42	32	9
N	19	8	10	3	21	30	21	72	8	34	42	32	9
N	20	10	10	4	24	27	21	72	8	34	42	32	9
N	21	11	11	3	25	26	21	72	8	34	42	32	9
N	22	11	12	3	26	25	21	72	8	34	42	32	9
N	23	12	12	3	27	24	21	72	8	34	42	32	9
N	24	12	12	3	27	24	21	72	8	34	42	32	9
N	25	12	13	3	28	23	21	72	8	34	42	32	9
N	26	12	13	3	28	23	21	72	8	34	42	32	9
N	27	12	13	3	28	23	21	72	8	34	42	32	9
N	28	12	13	3	28	23	21	72	8	34	42	32	9
N	29	12	13	3	28	23	21	72	8	34	42	32	9
N	30	12	13	3	28	23	21	72	8	34	42	32	9
P	2	3	1	2	6	45	21	72	8	34	42	32	9
P	3	3	3	3	9	42	21	72	8	34	42	32	9
P	4	7	4	6	17	34	21	72	8	34	42	32	9
P	5	7	4	6	17	34	21	72	8	34	42	32	9
P	6	9	5	5	19	32	21	72	8	34	42	32	9
P	7	9	9	5	23	28	21	72	8	34	42	32	9
P	8	9	11	3	23	28	21	72	8	34	42	32	9
P	9	9	11	3	23	28	21	72	8	34	42	32	9
P	10	10	11	3	24	27	21	72	8	34	42	32	9
P	11	11	11	3	25	26	21	72	8	34	42	32	9
P	12	11	11	3	25	26	21	72	8	34	42	32	9
P	13	12	13	3	28	23	21	72	8	34	42	32	9
P	14	12	13	3	28	23	21	72	8	34	42	32	9
P	15	12	13	3	28	23	21	72	8	34	42	32	9
P	16	12	13	3	28	23	21	72	8	34	42	32	9
P	17	12	13	3	28	23	21	72	8	34	42	32	9
P	18	13	13	3	29	22	21	72	8	34	42	32	9
P	19	13	14	3	30	21	21	72	8	34	42	32	9
P	20	13	14	3	30	21	21	72	8	34	42	32	9
P	21	13	14	3	30	21	21	72	8	34	42	32	9
P	22	13	14	3	30	21	21	72	8	34	42	32	9
P	23	13	14	3	30	21	21	72	8	34	42	32	9
P	24	13	14	3	30	21	21	72	8	34	42	32	9
P	25	13	14	3	30	21	21	72	8	34	42	32	9

Resultados del archivo Instituto Oriente													
		Nombres comunes							Nom. Propios				
ban	**vent**	**cd**	**ci**	**ai**	**cor**	**det**	**nr**	**com**	**CP**	**NCP**	**Prop**	**Verb**	**Pron**
P	26	13	14	3	30	21	21	72	8	34	42	32	9
P	27	13	14	3	30	21	21	72	8	34	42	32	9
P	28	13	14	3	30	21	21	72	8	34	42	32	9
P	29	13	14	3	30	21	21	72	8	34	42	32	9
P	30	13	14	3	30	21	21	72	8	34	42	32	9
V	2	5	1	2	8	43	21	72	8	34	42	32	9
V	3	8	4	3	15	36	21	72	8	34	42	32	9
V	4	8	7	4	19	32	21	72	8	34	42	32	9
V	5	8	9	4	21	30	21	72	8	34	42	32	9
V	6	8	9	5	22	29	21	72	8	34	42	32	9
V	7	10	10	5	25	26	21	72	8	34	42	32	9
V	8	10	10	5	25	26	21	72	8	34	42	32	9
V	9	11	10	5	26	25	21	72	8	34	42	32	9
V	10	11	12	3	26	25	21	72	8	34	42	32	9
V	11	11	13	3	27	24	21	72	8	34	42	32	9
V	12	12	13	3	28	23	21	72	8	34	42	32	9
V	13	12	13	3	28	23	21	72	8	34	42	32	9
V	14	12	13	3	28	23	21	72	8	34	42	32	9
V	15	12	13	3	28	23	21	72	8	34	42	32	9
V	16	12	13	3	28	23	21	72	8	34	42	32	9
V	17	13	13	3	29	22	21	72	8	34	42	32	9
V	18	13	13	3	29	22	21	72	8	34	42	32	9
V	19	13	13	3	29	22	21	72	8	34	42	32	9
V	20	13	13	3	29	22	21	72	8	34	42	32	9
V	21	13	13	3	29	22	21	72	8	34	42	32	9
V	22	13	14	3	30	21	21	72	8	34	42	32	9
V	23	13	14	3	30	21	21	72	8	34	42	32	9
V	24	13	14	3	30	21	21	72	8	34	42	32	9
V	25	13	14	3	30	21	21	72	8	34	42	32	9
V	26	13	14	3	30	21	21	72	8	34	42	32	9
V	27	13	14	3	30	21	21	72	8	34	42	32	9
V	28	13	14	3	30	21	21	72	8	34	42	32	9
V	29	13	14	3	30	21	21	72	8	34	42	32	9
V	30	13	14	3	30	21	21	72	8	34	42	32	9

Resultados del archivo El Cerebro													
	Nombres comunes								Nom. Propios				
ban	vent	cd	ci	ai	cor	det	nr	com	CP	NCP	Prop	Verb	Pron
F	2	3	6	1	10	77	33	120	0	4	4	92	48
F	3	5	8	6	19	68	33	120	0	4	4	92	48
F	4	8	13	12	33	54	33	120	0	4	4	92	48
F	5	8	14	12	34	53	33	120	0	4	4	92	48
F	6	10	15	10	35	52	33	120	0	4	4	92	48
F	7	10	16	12	38	49	33	120	0	4	4	92	48
F	8	10	16	17	43	44	33	120	0	4	4	92	48
F	9	10	17	18	45	42	33	120	0	4	4	92	48
F	10	10	18	19	47	40	33	120	0	4	4	92	48
F	11	10	19	18	47	40	33	120	0	4	4	92	48
F	12	10	19	23	52	35	33	120	0	4	4	92	48
F	13	11	23	22	56	31	33	120	0	4	4	92	48
F	14	11	24	21	56	31	33	120	0	4	4	92	48
F	15	11	24	21	56	31	33	120	0	4	4	92	48
F	16	11	26	19	56	31	33	120	0	4	4	92	48
F	17	11	27	19	57	30	33	120	0	4	4	92	48
F	18	11	27	19	57	30	33	120	0	4	4	92	48
F	19	11	27	19	57	30	33	120	0	4	4	92	48
F	20	11	27	20	58	29	33	120	0	4	4	92	48
F	20	11	27	20	58	29	33	120	0	4	4	92	48
F	21	11	28	20	59	28	33	120	0	4	4	92	48
F	22	12	28	19	59	28	33	120	0	4	4	92	48
F	23	12	28	21	61	26	33	120	0	4	4	92	48
F	24	12	28	21	61	26	33	120	0	4	4	92	48
F	25	12	28	22	62	25	33	120	0	4	4	92	48
F	26	12	28	23	63	24	33	120	0	4	4	92	48
F	27	12	28	23	63	24	33	120	0	4	4	92	48
F	28	12	28	23	63	24	33	120	0	4	4	92	48
F	29	12	28	23	63	24	33	120	0	4	4	92	48
F	30	12	28	23	63	24	33	120	0	4	4	92	48
N	2	4	10	3	17	70	33	120	0	4	4	92	48
N	3	5	10	7	22	65	33	120	0	4	4	92	48
N	4	6	10	12	28	59	33	120	0	4	4	92	48
N	5	6	11	16	33	54	33	120	0	4	4	92	48
N	6	7	15	15	37	50	33	120	0	4	4	92	48
N	7	7	16	17	40	47	33	120	0	4	4	92	48
N	8	8	16	19	43	44	33	120	0	4	4	92	48
N	9	8	16	22	46	41	33	120	0	4	4	92	48
N	10	10	20	22	52	35	33	120	0	4	4	92	48
N	11	11	21	21	53	34	33	120	0	4	4	92	48
N	12	11	21	21	53	34	33	120	0	4	4	92	48

Resultados del archivo El Cerebro													
	Nombres comunes								Nom. Propios				
ban	**vent**	**cd**	**ci**	**ai**	**cor**	**det**	**nr**	**com**	**CP**	**NCP**	**Prop**	**Verb**	**Pron**
N	13	11	21	21	53	34	33	120	0	4	4	92	48
N	14	11	23	20	54	33	33	120	0	4	4	92	48
N	15	11	23	21	55	32	33	120	0	4	4	92	48
N	16	11	25	20	56	31	33	120	0	4	4	92	48
N	17	11	25	20	56	31	33	120	0	4	4	92	48
N	18	11	26	20	57	30	33	120	0	4	4	92	48
N	19	11	26	20	57	30	33	120	0	4	4	92	48
N	20	11	26	20	57	30	33	120	0	4	4	92	48
N	21	11	26	20	57	30	33	120	0	4	4	92	48
N	22	11	26	21	58	29	33	120	0	4	4	92	48
N	23	11	26	23	60	27	33	120	0	4	4	92	48
N	24	11	26	23	60	27	33	120	0	4	4	92	48
N	25	11	27	23	61	26	33	120	0	4	4	92	48
N	26	12	28	21	61	26	33	120	0	4	4	92	48
N	27	12	28	21	61	26	33	120	0	4	4	92	48
N	28	12	28	23	63	24	33	120	0	4	4	92	48
N	29	12	28	23	63	24	33	120	0	4	4	92	48
N	30	12	28	23	63	24	33	120	0	4	4	92	48
P	2	8	13	9	30	57	33	120	0	4	4	92	48
P	3	10	15	16	41	46	33	120	0	4	4	92	48
P	4	10	19	20	49	38	33	120	0	4	4	92	48
P	5	11	22	20	53	34	33	120	0	4	4	92	48
P	6	11	26	20	57	30	33	120	0	4	4	92	48
P	7	11	28	20	59	28	33	120	0	4	4	92	48
P	8	12	28	23	63	24	33	120	0	4	4	92	48
P	9	12	28	23	63	24	33	120	0	4	4	92	48
P	10	12	29	22	63	24	33	120	0	4	4	92	48
P	11	12	31	21	64	23	33	120	0	4	4	92	48
P	12	12	31	22	65	22	33	120	0	4	4	92	48
P	13	12	34	19	65	22	33	120	0	4	4	92	48
P	14	12	36	18	66	21	33	120	0	4	4	92	48
P	15	12	36	18	66	21	33	120	0	4	4	92	48
P	16	12	36	18	66	21	33	120	0	4	4	92	48
P	17	12	36	18	66	21	33	120	0	4	4	92	48
P	18	12	36	18	66	21	33	120	0	4	4	92	48
P	19	12	36	18	66	21	33	120	0	4	4	92	48
P	20	12	36	18	66	21	33	120	0	4	4	92	48
P	21	12	36	18	66	21	33	120	0	4	4	92	48
P	22	12	36	18	66	21	33	120	0	4	4	92	48
P	23	12	36	18	66	21	33	120	0	4	4	92	48
P	24	12	36	18	66	21	33	120	0	4	4	92	48
P	25	12	36	18	66	21	33	120	0	4	4	92	48

Resultados del archivo El Cerebro													
	Nombres comunes								Nom. Propios				
ban	**vent**	**cd**	**ci**	**ai**	**cor**	**det**	**nr**	**com**	**CP**	**NCP**	**Prop**	**Verb**	**Pron**
P	26	12	36	18	66	21	33	120	0	4	4	92	48
P	27	12	36	18	66	21	33	120	0	4	4	92	48
P	28	12	36	18	66	21	33	120	0	4	4	92	48
P	29	12	36	18	66	21	33	120	0	4	4	92	48
P	30	12	36	18	66	21	33	120	0	4	4	92	48
V	2	6	9	6	21	66	33	120	0	4	4	92	48
V	3	6	10	7	23	64	33	120	0	4	4	92	48
V	4	7	14	10	31	56	33	120	0	4	4	92	48
V	5	7	15	12	34	53	33	120	0	4	4	92	48
V	6	8	17	13	38	49	33	120	0	4	4	92	48
V	7	9	18	14	41	46	33	120	0	4	4	92	48
V	8	10	18	15	43	44	33	120	0	4	4	92	48
V	9	11	19	19	49	38	33	120	0	4	4	92	48
V	10	11	19	19	49	38	33	120	0	4	4	92	48
V	11	11	22	19	52	35	33	120	0	4	4	92	48
V	12	11	23	19	53	34	33	120	0	4	4	92	48
V	13	11	24	19	54	33	33	120	0	4	4	92	48
V	14	11	24	21	56	31	33	120	0	4	4	92	48
V	15	11	24	21	56	31	33	120	0	4	4	92	48
V	16	11	26	19	56	31	33	120	0	4	4	92	48
V	17	11	27	19	57	30	33	120	0	4	4	92	48
V	18	11	28	21	60	27	33	120	0	4	4	92	48
V	19	11	28	22	61	26	33	120	0	4	4	92	48
V	20	11	28	23	62	25	33	120	0	4	4	92	48
V	21	12	28	22	62	25	33	120	0	4	4	92	48
V	22	12	28	22	62	25	33	120	0	4	4	92	48
V	23	12	28	23	63	24	33	120	0	4	4	92	48
V	24	12	28	23	63	24	33	120	0	4	4	92	48
V	25	12	28	23	63	24	33	120	0	4	4	92	48
V	26	12	28	23	63	24	33	120	0	4	4	92	48
V	27	12	28	23	63	24	33	120	0	4	4	92	48
V	28	12	28	23	63	24	33	120	0	4	4	92	48
V	29	12	28	23	63	24	33	120	0	4	4	92	48
V	30	12	29	22	63	24	33	120	0	4	4	92	48

Anexo H: Archivos de texto libre utilizados para evaluación

En este anexo se presentan los listados de los archivos de texto libre en formato texto de windows, tal como fueron utilizados para su evaluación.

El primero es un archivo de correo recibido por e-mail titulado “Contra la guerra”

Contra la guerra.
En todo el mundo 10000000, dicen, hemos desfilado contra la guerra.
Participé durante un trecho en la de Barcelona.
Era impresionante.
Dicen que éramos 1300000 personas clamando por la paz.
El problema es por qué no sucede lo mismo ante cualquier guerra o pisoteo de los derechos humanos.
No se trata sólo de esta guerra, sino de muchas que se están librando por todo el mundo.
Además creo que debería quedar claro que el marchar por la paz no significa que estemos a favor de Sadam, pues "gracias" a él su pueblo se está hundiendo.
¿Vivimos realmente en un mundo de buenos y malos?
¿Están los malos legitimados para permitir que, en su nombre, mueran inocentes?
¿Está tan claro lo que es el bien y el mal?
Hemos elaborado los derechos humanos y parece que no sirven para nada, pues en todas partes se siguen atropellando.
Quizá llegue el momento en que muchos millones nos manifestemos en contra de este atropello, que a menudo sucede a nuestro lado.
No hace falta alejarse mucho para observarlo.
Como dice la sabiduría popular "es fácil ver la paja en el ojo del vecino, sin darse cuenta de la viga en nuestro propio ojo".
Un abrazo a todos.
Maite.

El segundo es un archivo de correo recibido por e-mail del Instituto Oriente en Puebla, Pue. (mis hijos estudian en esta institución).

Instituto Oriente.
Puebla, Pue.
Febrero de 2003.
Comparto contigo los puntos de vista de los jesuitas sobre la propuesta de guerra del Presidente de los Estados Unidos de América, contra Irak.

José Amado Fernández Ruiz S.J.

PROVINCIA MEXICANA DE LA COMPAÑÍA DE JESUS.

"LA GUERRA ES UNA DERROTA DE LA HUMANIDAD".

AL PUEBLO Y AL GOBIERNO DE MÉXICO:

Fieles a nuestro compromiso de ser "servidores de la misión de Cristo", queremos reafirmar nuestra oposición a la guerra y nuestra determinación de trabajar a favor de una paz anclada firmemente en la justicia.
Deseamos exponer las razones principales que se oponen a una guerra contra Irak.

La 'doctrina' de la guerra preventiva no está de acuerdo con la doctrina y el derecho de la ONU, ni es moralmente sostenible.
(1) La aplicación de esta doctrina abriría las puertas a una guerra infinita, a 'una guerra sin fin'.
(2) En lugar de traer una paz duradera, una guerra contra Irak aumentaría las tensiones entre musulmanes, cristianos y judíos.
(3) Los masivos gastos militares están en contraste con el interés por promover el desarrollo sostenible para todos.
(4) Que los líderes de unos pocos países industrializados hayan tomado decisiones que afectan a todos los pueblos, afecta gravemente el derecho internacional y debilita los organismos multilaterales tan pacientemente construidos por la humanidad entera.
(5) La experiencia nos ha enseñado que los pobres son siempre las víctimas principales de la violencia y de la guerra.

Estas son las razones por las cuales nuestros esfuerzos a favor de la paz adquieren una apremiante urgencia y por las cuales respaldamos los esfuerzos de las organizaciones sociales en contra de la guerra y la actual política del Gobierno de México en el seno del Consejo de Seguridad de la ONU.
El Papa Juan Pablo II proclamó enfáticamente que la "guerra nunca es una simple fatalidad, es siempre una derrota de la humanidad".

Provincia Mexicana de la Compañía de Jesús
Juan Luis Orozco Hernández, S.J., Provincial

El tercero es un archivo de correo recibido por e-mail del Club de efectividad que normalmente envía aríticulos de reflexión, motivación y superación personal; su URL es: **http://www.efectividad.net/club**

La luz del cerebro.
Comparar el cerebro con una galaxia es, sin duda, una analogía modesta.
Todos los seres humanos llevamos con nosotros cerca de 1600 gramos de tejidos sin darle mucha importancia y, lo que es más, cada cerebro humano es capaz de realizar más interconexiones perfectamente configuradas que átomos hay en el universo.
Si este número teórico pudiera escribirse, sería un "1" seguido de cerca de diez millones y medio de kilómetros de ceros.

"El cerebro humano es un telar encantado donde millones de deslumbrantes lanzaderas tejen un dibujo que se desvanece en el aire, un dibujo siempre rebosante de un significado nunca perdurable. Es como si la Vía Láctea ejecutara una danza cósmica."
- Sir Charles Sherrington -

El cerebro humano no es, físicamente, muy sorprendente.
Aquellos que lo han visto, no lo describen como una visión particularmente extraordinaria.
Sin embargo, su potencial es fascinante, pero a pesar de ello, lo hacemos funcionar muy por debajo de su capacidad, incluso a veces lo maltratamos.

Es comprensible que un concertista de piano, o un artesano, valoren sus manos por encima de todo; que un pintor aprecie sus ojos; que un atleta se preocupe de sus piernas... Pero las manos son inútiles sin el cerebro, tanto como el piano sin ejecutante.
El potencial del cerebro ha sido largamente subestimado, precisamente por su omnipresencia.
El cerebro está presente en todo lo que hacemos, en cada cosa que nos sucede.
Es quizás por ser una "constante", que lo anulamos de la ecuación y así, notamos lo particular de cada experiencia, pasando por alto aquello sin lo cuál nada es posible para nosotros.

En cada cabeza se esconde una central eléctrica formidable, un órgano compacto y eficiente cuya capacidad -cuanto más sabemos de él- parecería expandirse hacia el infinito.
"Si la complejidad del cerebro pudiera de alguna manera transformarse en algo visible, de forma tal que se manifestara más claramente a nuestros sentidos, el mundo biológico sería un campo de luz comparado con el mundo físico. El sol, con sus grandes erupciones, se apagaría hasta quedar reducido a la pálida simplicidad de un ramo de rosas; una lombriz sería un faro; un perro, una ciudad de luz, y los seres humanos parecerían soles resplandecientes de complejidad, enceguecedores estallidos de sentido mutuo atravesando la triste noche del mundo físico que los separa. Podríamos herirnos los ojos mutuamente. Mirad las cabezas aureoladas de vuestros raros y complejos semejantes. ¿No es así?"
- John Rader Platt -

Cada cosa que hacemos y experimentamos, desde jugar al tenis hasta pagar las cuentas, tiene como base este complejo sistema bio-eléctrico.
Bien mirado, no es tan desconcertante como parece.
Sabemos que los ojos no ven por sí mismos, son simplemente lentes.
Los oídos no oyen por sí mismos; son -por así decirlo- micrófonos.

Cuando miramos un partido de fútbol por televisión, no vemos a los jugadores, sino representaciones electrónicas de ellos en la pantalla.
Entre el gato vivo que usted mira y la imagen del gato que se forma en su cerebro, hay una serie de procesos neuro-fisiológicos que separan la imagen que usted recibe, de la misma manera que una serie de procesos electrónicos separan el partido de fútbol, ahí donde se está jugando, y lo lleva a su pantalla.

Nuestro cerebro es, casi literalmente, todo.
Podemos ofrecerle más y, como contrapartida, él también nos dará más.
Es nuestra arma secreta y silenciosa.
Si comenzamos a usar más su poder, veremos una luz que herirá y asombrará a nuestros ojos.

Anexo I: Resultados de corridas con texto libre

En este anexo se presentan los listados, de los tres tipos de archivo, de resultados obtenidos por el programa, en una corrida con tipo de bandera “punto” y tamaño de ventana nueve para el archivo reportado “Contra la guerra”. Se presentan como los genera el programa; en ellos se utilizan la siguiente simbología y tipos de abreviatura:

(NNN) = número de unidad léxica (token) consecutiva en el archivo
***** = marca la unidad léxica enlazada
<- abr - tipo de enlace donde abr puede ser
cd = **c**orreferencia **d**irecta
ci = **c**orreferencia **i**ndirecta
ai = **a**náfora **i**ndirecta

El primer archivo es un listado de unidades léxicas consecutivas marcando la ocurrencia o no de correferencia directa e indirecta, y la de anáfora indirecta; de nombre res\contra.cad

```
Archivo f_tnt\contra.tts

(    1)   Contra                          (   27)   .
(    2)   la                              (   28)   Era
(    3) * guerra  <-cd-(17) guerra        (   29)   impresionante
(    4)   .                               (   30)   .
(    5)   En                              (   31)   Dicen
(    6)   todo                            (   32)   que
(    7)   el                              (   33)   éramos
(    8) * mundo  <-cd-(78) mundo          (   34)   1300000
(    9)   10000000                        (   35)   personas
(   10)   ,                               (   36)   clamando
(   11)   dicen                           (   37)   por
(   12)   ,                               (   38)   la
(   13)   hemos                           (   39) * paz  <-cd-(91) paz
(   14)   desfilado                       (   40)   .
(   15)   contra                          (   41)   El
(   16)   la                              (   42) * problema  <-ci-(151) mal
(   17) * guerra  <-ci-(42) problema      (   43)   es
(   18)   .                               (   44)   por
(   19)   Participé                       (   45)   qué
(   20)   durante                         (   46)   no
(   21)   un                              (   47)   sucede
(   22)   trecho                          (   48)   lo
(   23)   en                              (   49)   mismo
(   24)   la                              (   50)   ante
(   25)   de                              (   51)   cualquier
(   26)   Barcelona                       (   52) * guerra  <-cd-(66) guerra
```

```
(   53)   o                                  (  108)   pueblo
(   54)   pisoteo                            (  109)   se
(   55)   de                                 (  110)   está
(   56)   los                                (  111)   hundiendo
(   57)   derechos                           (  112)   .
(   58)   humanos                            (  113)   ¿
(   59)   .                                  (  114)   Vivimos
(   60)   No                                 (  115)   realmente
(   61)   se                                 (  116)   en
(   62)   trata                              (  117)   un
(   63)   sólo                               (  118)   mundo
(   64)   de                                 (  119)   de
(   65)   esta                               (  120)   buenos
(   66)   guerra                            (  121)   y
(   67)   ,                                  (  122)   malos
(   68)   sino                               (  123)   ?
(   69)   de                                 (  124)   ¿
(   70)   muchas                            (  125)   Están
(   71)   que                                (  126)   los
(   72)   se                                 (  127)   malos
(   73)   están                             (  128)   legitimados
(   74)   librando                          (  129)   para
(   75)   por                                (  130)   permitir
(   76)   todo                               (  131)   que
(   77)   el                                 (  132)   ,
(   78) * mundo  <-ai-(108) pueblo           (  133)   en
(   79)   .                                  (  134)   su
(   80)   Además                             (  135)   nombre
(   81)   creo                               (  136)   ,
(   82)   que                                (  137)   mueran
(   83)   debería                            (  138)   inocentes
(   84)   quedar                            (  139)   ?
(   85)   claro                             (  140)   ¿
(   86)   que                               (  141)   Está
(   87)   el                                 (  142)   tan
(   88)   marchar                           (  143)   claro
(   89)   por                                (  144)   lo
(   90)   la                                 (  145)   que
(   91)   paz                                (  146)   es
(   92)   no                                 (  147)   el
(   93)   significa                         (  148)   bien
(   94)   que                               (  149)   y
(   95)   estemos                           (  150)   el
(   96)   a                                  (  151)   mal
(   97) * favor  <-ci-(135) nombre           (  152)   ?
(   98)   de                                 (  153)   Hemos
(   99)   Sadam                              (  154)   elaborado
(  100)   ,                                  (  155)   los
(  101)   pues                               (  156)   derechos
(  102)   "                                  (  157)   humanos
(  103) * gracias  <-ci-(156)                (  158)   y
derechos                                     (  159)   parece
(  104)   "                                  (  160)   que
(  105)   a                                  (  161)   no
(  106)   él                                 (  162)   sirven
(  107)   su                                 (  163)   para
```

```
(  164)   nada                          (  203)   para
(  165)   ,                             (  204)   observarlo
(  166)   pues                          (  205)   .
(  167)   en                            (  206)   Como
(  168)   todas                         (  207)   dice
(  169) * partes  <-ai-(196) lado       (  208)   la
(  170)   se                            (  209)   sabiduría
(  171)   siguen                        (  210)   popular
(  172)   atropellando                  (  211)   "
(  173)   .                             (  212)   es
(  174)   Quizá                         (  213)   fácil
(  175)   llegue                        (  214)   ver
(  176)   el                            (  215)   la
(  177)   momento                       (  216)   paja
(  178)   en                            (  217)   en
(  179)   que                           (  218)   el
(  180)   muchos                       (  219)   ojo
(  181)   millones                     (  220)   del
(  182)   nos                           (  221)   vecino
(  183)   manifestemos                 (  222)   ,
(  184)   en                            (  223)   sin
(  185)   contra                       (  224)   darse
(  186)   de                            (  225)   cuenta
(  187)   este                        (  226)   de
(  188)   atropello                    (  227)   la
(  189)   ,                             (  228)   viga
(  190)   que                           (  229)   en
(  191)   a                             (  230)   nuestro
(  192)   menudo                       (  231)   propio
(  193)   sucede                       (  232)   ojo
(  194)   a                             (  233)   "
(  195)   nuestro                      (  234)   .
(  196)   lado                        (  235)   Un
(  197)   .                           (  236)   abrazo
(  198)   No                          (  237)   a
(  199)   hace                        (  238)   todos
(  200)   falta                       (  239)   .
(  201)   alejarse                    (  240)   Maite
(  202)   mucho                      (  241)   .
```

El segundo archivo es un resumen estadístico de la corrida con nombre estadis.txt

```
archivo          ban  vent  cd   ci   ai   cor  det nr  com  CP   NCP  Prop Verb
Pron
f_tnt\contra.tts  P   9      6    6    2    14   14 9   37   0     2      2   44
14
```

El tercer archivo es un archivo en formato htm para permitir una mejor navegación en el seguimiento de resultados de nombre res\contra.htm. En este archivo se pueden observar subíndices númericos entre parentesis que marcan las unidades léxicas (tokens); si el subíndice está subrayado indica que la unidad léxica es una coreferencia o anáfora indirecta. La

información adicional se puede obtener con un navegador y colocando el puntero del ratón sobre el elemento (ver anexo J).

Contra la guerra (3).
En todo el mundo (8)10000000 , dicen , hemos desfilado contra la guerra (17).
Participé durante un trecho (22)en la de Barcelona (26).
Era impresionante .
Dicen que éramos 1300000 personas (35)clamando por la paz (39).
El problema (42)es por qué no sucede lo mismo ante cualquier guerra (52)o pisoteo (54)de los derechos (57)humanos .
No se trata sólo de esta guerra (66), sino de muchas que se están librando por todo el mundo (78).
Además creo que debería quedar claro que el marchar por la paz (91)no significa que estemos a favor (97)de Sadam (99), pues " gracias (103)" a él su pueblo (108)se está hundiendo .
¿ Vivimos realmente en un mundo (118)de buenos (120)y malos (122)? ¿ Están los malos (127)legitimados para permitir que , en su nombre (135), mueran inocentes ? ¿ Está tan claro lo que es el bien (148)y el mal (151)? Hemos elaborado los derechos (156)humanos y parece que no sirven para nada , pues en todas partes (169)se siguen atropellando .
Quizá llegue el momento (177)en que muchos millones (181)nos manifestemos en contra de este atropello (188), que a menudo sucede a nuestro lado (196).
No hace falta alejarse mucho para observarlo .
Como dice la sabiduría (209)popular " es fácil ver la paja (216)en el ojo (219)del vecino (221), sin darse cuenta (225)de la viga (228)en nuestro propio ojo (232)" .
Un abrazo (236)a todos .
Maite .

Anexo J: Cómo correr el programa de demostración e interpretar su archivo de salida

Una versión de demostración del software desarrollado en esta tesis se anexa a la misma en forma de un CD-ROM. El programa corre bajo el sistema Windows. En el disco aparece el archivo `leame.txt` que contiene las instrucciones para su uso; aquí repetimos la parte más importante de ese archivo y también damos la información para la interpretación de los resultados de la corrida del programa.

Para utilizar la demostración es necesario hacer los siguientes pasos:

1. Copiar todos los archivos del subdirectorio `demo` del CD-ROM a un subdirectorio en el disco duro de su PC, por ejemplo, `C:\demo` (porque el sistema necesita crear y almacenar en archivos al desarrollar su proceso).

2. Abrir una ventana MS-DOS.

3. Ir al directorio donde colocó los archivos del CD, usando el comando `cd`, por ejemplo:

```
C:\> cd demo
```

4. El disco ya inclye un texto de ejemplo, `demo\f_txt\io.txt`. Para correr el programa usando este archivo como entrada, ejecutar:

```
C:\demo> demo io
```

5. Si quiere procesar su propio archivo, por ejemplo, `su_archivo.txt`:

    a. Colocarlo en el subdirectorio `f_txt`, así que, en nuestro ejemplo, aparezca como `C:\demo> demo\f_txt\su_archivo.txt.`

    b. Ejecutar:

```
C:\demo> demo su_archivo
```

nótese que el nombre aparece sin la extensión del archivo (.txt).

6. Averiguar que el programa se terminó con el mensaje “Terminado OK”.

7. Ver el archivo de salida en cualquier visualisador de Internet actualizado, de preferencia en Interner Explorer de version mayor a 5.0. El archivo de salida aparece en el directorio `res` y su nombre se comone del nombre del archivo original más un código que refleja los parámetros del programa. Para visualizar el archivo, se puede abrirlo de la ventana del visualizador o bien ejecutar desde la líne de comandos:

```
C:\demo> start res\io_F_9.htm
```

o en su caso

```
C:\demo> start res\su_archivo_F_9.htm
```

A continuación se explica cómo se interpretan los mensajes que el programa muestra durante su ejecución y más abajo, cómo se interpreta el archivo .htm de salida.

Seguimiento de corrida del programa

Al correr el programa podrá observarse la corrida en la pantalla, véase la figura 23. Los pasos indicados en la figura son los siguientes:

1. Procesando archivo. Este mensaje es de retroalimentación
2. Convierte el texto a tokens. Se transforma el archivo a una unidad léxica por línea. El resultado puede apreciarse en el archivo temp\su_archivo.tt

3 al 18. Convierte tokens a etiquetado. En la línea tres se llama al etiquetador TnT que imprime las acciones que está realizando, además de información estadística. El resultado puede apreciarse en el archivo temp\su_archivo.tts

19. Limpia comentarios de etiquetado. En este paso se elimina información de TnT como la identificación y estadística para obtener un archivo con etiquetado simple. El resultado puede apreciarse en el archivo f_tnt\su_archivo.tts
20. convierte tnt a htm. En este paso se llama al programa de detección de anáfora indirecta que despliega mensajes de acuerdo a las tareas que va realizando y termina en un

tiempo máximo de dos minutos aproximadamante (con un archivo no mayor de 45 KB). El resultado puede apreciarse en el archivo res\su_archivo_F_9.htm (que se muestra en la figura 24). Se puede apreciar que el nombre del archivo se altera con los argumentos por omisión (la bandera y el tamaño de la ventana) que utiliza el programa demo. En este caso F = signo de punto y el número 9 indica el ancho de la ventana en número de puntos a encontrar en la búsqueda hacia atrás.

21 al 27 son pasos que internamente ejecuta el algoritmo. En el renglón 21 aparece el nombre del archivo interno derivado del nombre del archivo de la entrada, junto con las opciones de corrida: la cifra según la tabla 5 y la letra y cifra según la sección 6.5. Después se dan mensajes al activarse diferentes módulos del programa. Finalmente, se da mensaje de la terminación normal del programa.

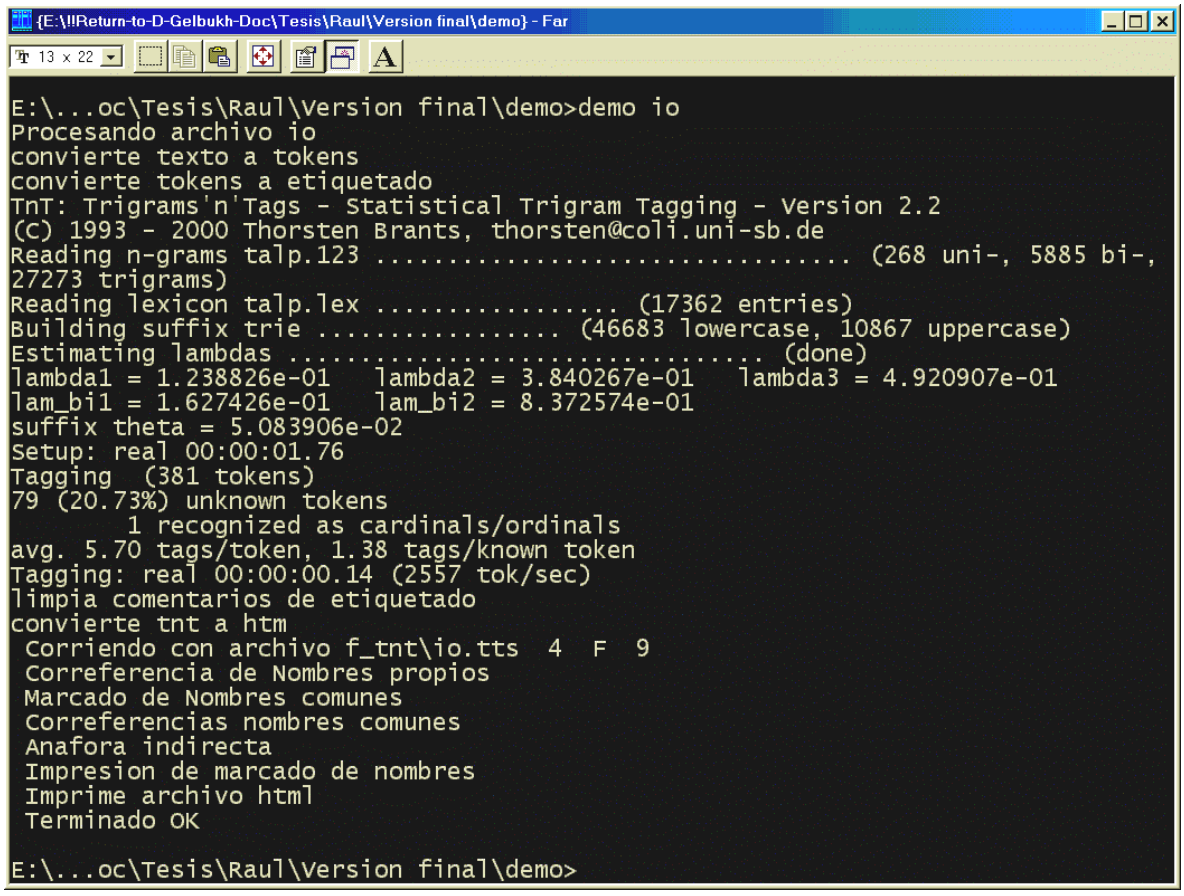


Figura 23 Corrida del programa

Interpretación del archivo de salida

El archivo de salida, mostrado en la figura 24, es del tipo de hipertexto (.htm) y muestra el texto de entrada con el siguiente marcado agregado.

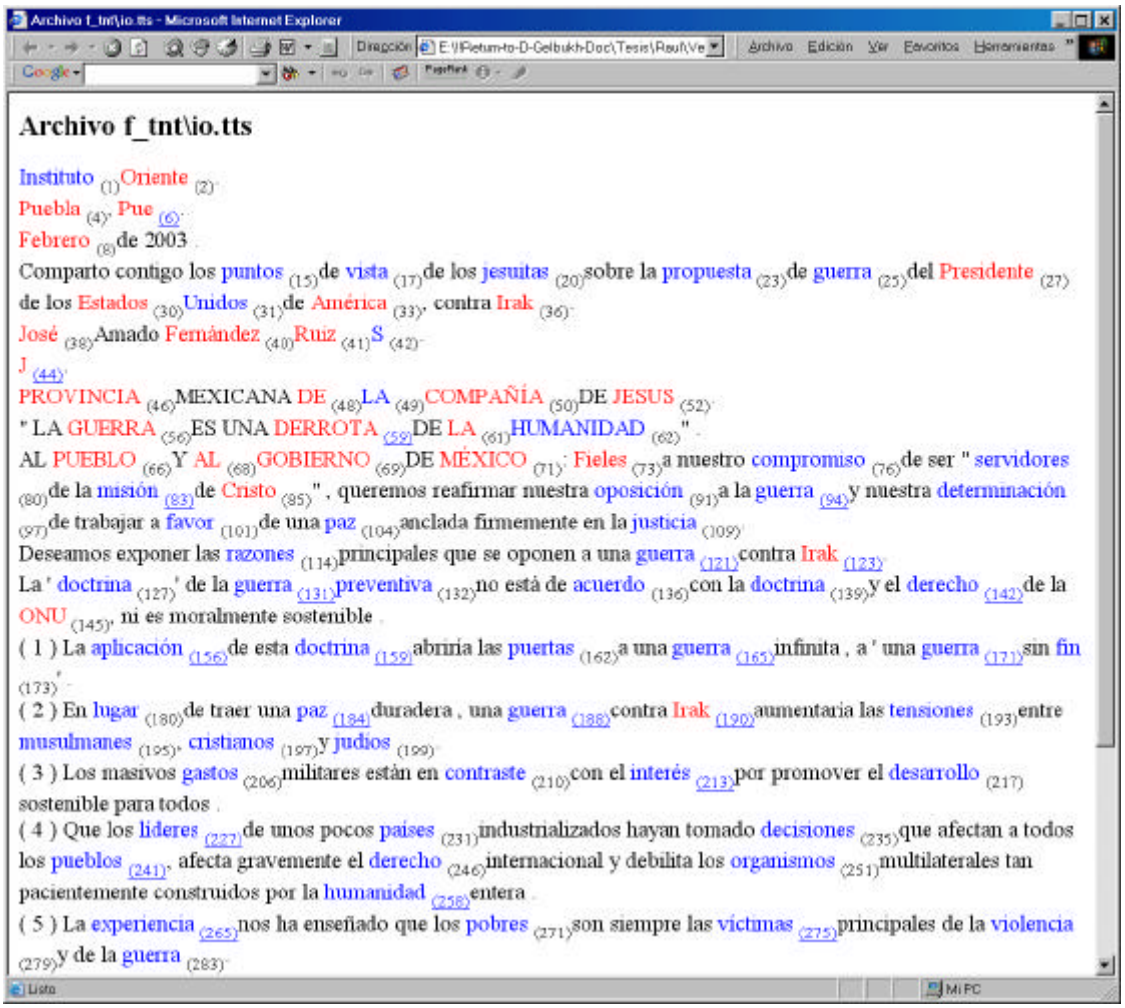


Figura 24 Archivo de salida del programa

Los nombres propios aparecen en rojo; los nombres comunes en azul, y los demás elementos léxicos. Las unidades léxicas (*tokens*) se identifican con números secuenciales que aparecen como subíndices entre paréntesis (sólo se muestran los números para los nombres propios y comunes, aunque se cuentan todas las palabras).

La aportación principal del presente trabajo es la identificación de las relaciones de correferencia y de anáfora indirecta entre palabras, las que forman cadenas de palabras una refiriéndose a otra. Estas cadenas encontradas por el programa se muestran a través de las referencias cruzadas en la pantalla de salida. Así, si el subíndice está subrayado indica que la unidad léxica es una correferencia o anáfora indirecta, según la información adicional que se puede obtener con un navegador colocando el puntero del ratón sobre la palabra; en este momento se desplegará en una ventanita amarilla la abreviatura del tipo de correferencia o anáfora indirecta (véase Anexo I) seguido del número de palabra con la que existe dicha relación. Al hacer clic con el ratón en tal palabra, el apuntador se desplazará al elemento referenciado.

De esta manera se queda demostrada la habilidad del programa a encontrar las relaciones de correferencia y anáfora indirecta en un texto libre (no preparado) en español.