Propuesta de un espacio de accesibilidad anafórica estructural para textos HTML*

Borja Navarro, Patricio Martínez-Barco y Rafael Muñoz
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Carretera de San Vicente del Raspeig - Alicante - España
Tel. 965903400 ext. 2963 Fax. 965909326
{borja, patricio, rafael}@dlsi.ua.es

Resumen En este artículo se presenta un estudio sobre el espacio de accesibilidad anafórico que puede extraerse de la estructura generada por la publicación electrónica de documentos a través del formalismo HTML. Esta propuesta se basa en la dependencia existente entre la resolución de la anáfora y la estructura del discurso. Aprovechando las etiquetas de marca proporcionadas por HTML, proponemos una especificación del espacio de accesibilidad para cada tipo de expresión anafórica. A diferencia de otras propuestas de espacios de accesibilidad, nuestra propuesta no hace una definición arbitraria basada en unidades sintácticas (oraciones) sino en unidades de la estructura del discurso. Ello permite establecer un criterio consistente basado en información lingüística independiente del corpus.

1 Introducción

El texto, considerado tanto una unidad lingüística como una unidad pragmático-comunicativa, consta siempre de una estructura interna y subyacente. Como queda indicado en otros estudios [1] [18] [7] [14], el texto no es una simple suma o conjunto de oraciones, sino que las diferentes unidades que conforman un texto (léxicas, sintagmáticas, oracionales, etc.) están en su interior interrelacionadas de tal manera que actúan como un todo, como una unidad en sí misma. Así, a partir de esta interrelación de los elementos lingüísticos que forman un texto, y, por tanto, a partir de su estructura interna, el texto se presenta en una situación comunicativa como una unidad lingüística dotada de los dos fenómenos que lo definen y lo caracterizan: la cohesión (entendida como la interrelación superficial de los elementos textuales) y la coherencia (entendida como la interrelación semántica y pragmática subyacente) .

Sin embargo, el planteamiento de que todo texto consta de una estructura interna implica, entre otros asuntos, considerar que está formado por un conjunto de unidades textuales (superiores a la unidad oración) que, desde un punto de vista estático, mantienen entre sí relaciones jerárquicas. El texto está totalmente condicionado y determinado por la situación comunicativa contextual donde se halla, dado que es una unidad del "uso" lingüístico, y no tanto del sistema lingüístico de una lengua natural [2] [14] [11]. Por ello, no se pueden establecer unidades textuales abstractas y fijas, al estilo de las unidades oracionales, sino que las unidades textuales dependen de cada texto concreto y de su situación comunicativa.

De esta forma, el estudio del texto debe partir del conocimiento previo de una tipología textual, donde se determine qué tipo de texto vamos a tratar, las situaciones contextuales donde suele producirse y, con ello, sus características específicas. En nuestro caso, vamos a centrarnos en documentos periodísticos de información general en soporte electrónico.

Por otro lado, uno de los principales mecanismos gramaticales de cohesión textual, gracias al cual se mantiene la coherencia del discurso, es la anáfora, en tanto que mecanismo lingüístico que establece una relación entre dos unidades en cualquier nivel de la estructura textual. En este sentido, resulta determinante conocer, para la correcta interpretación de un elemento anafórico, dónde se sitúa su antecedente: determinar el ámbito de resolución anafórica. Normalmente, ante la dificultad o imposibilidad de establecer desde un punto de vista teórico dónde se sitúa el antecedente de una anáfora, se tiende a realizar

* Este trabajo ha sido subvencionado por la CICYT (Comisión Interministerial de Ciencia y Tecnología) bajo el proyecto TIC2000-0664-C02-01/02.

indicaciones generales con relación a la mayor o menor cercanía del elemento anafórico con su antecedente. Sin embargo, vamos aquí a plantear y a considerar que la situación del antecedente anafórico está íntimamente relacionada con la estructura jerárquica subyacente a todo texto.

Puesto que la mayoría de los documentos periodísticos electrónicos a los que se puede acceder se publican en Internet, el lenguaje HTML (HyperText Markup Language) permite especificar una estructura del mismo indicando sus unidades textuales a través de la inclusión de etiquetas que posteriormente serán analizadas por el visor utilizado por el lector. Así, el estudio que se plantea en este artículo tiene como objetivo aprovechar las relaciones existentes entre el mecanismo de la anáfora y la estructura jerárquica generada por dicho lenguaje. De esta forma, puede ser aplicada automáticamente a cualquier sistema de resolución computacional de la anáfora, salvando la dificultad existente en otros métodos que necesitan establecer previamente dicha estructura, y que en este caso ya aparece explícita.

Así, en el siguiente punto se realizará una revisión de algunos planteamientos existentes sobre las relaciones entre anáfora y estructura del discurso, y cómo han sido llevados a la práctica en otros tipos de textos. A partir de algunos de estos planteamientos se realizará, posteriormente, una propuesta de un espacio de accesibilidad anafórica que aproveche la información estructural aportada por HTML para el establecimiento del ámbito de resolución anafórica. Este planteamiento se justificará finalmente a través de un estudio empírico sobre un conjunto de artículos periodísticos en formato electrónico del que extraeremos una serie de relaciones entre la estructura HTML de estos documentos y la resolución de la anáfora para cada tipo de anáfora tratada.

2 *La resolución de la anáfora y la estructura del discurso*

Uno de los estudios más significativos acerca de las relaciones existentes entre la estructura del discurso y la generación (y por tanto, resolución) de la anáfora es el efectuado por Fox [4]. En este estudio, Fox investiga la anáfora basándose en sus relaciones con la estructura de discurso, mediante el uso del análisis convencional que le permite definir una estructura jerárquica de dicho discurso.

Fox basa su estudio en la creencia de que existe una fuerte relación entre el flujo de información en un texto, la estructura del texto y el uso de la anáfora. De hecho es sabido que aquellos antecedentes que se encuentran en el "foco" o en "la consciencia del oyente" pueden ser pronominalizados. Así la autora destaca que cualquier tratamiento de la anáfora debe ser entendido desde la estructura jerárquica así como desde el tipo de discurso usado como fuente de entrada.

En este sentido, se distinguen básicamente dos tipos de discurso, los conversacionales producidos por más de una persona y los no conversacionales escritos, y justifica el uso de modelos de representación de discurso distintos basándose en las siguientes diferencias:

- Los discursos conversacionales son muy interactivos y los modelos que lo representan deben capturar las relaciones sociales que se mantienen entre cada una de las piezas de la estructura (básicamente relaciones entre los enunciados).
- Por otra parte, el discurso no conversacional escrito es puramente informativo, y por tanto los modelos de discurso que lo representen deberán recoger las relaciones informativas que existen entre las piezas de la estructura (en este caso, relaciones entre proposiciones).

Puesto que la teoría de Fox se fundamenta en que la anáfora está, al menos en parte, relacionada con la estructura jerárquica del discurso, es necesario realizar un análisis adecuado de esta estructura en sus unidades básicas para poder comprender estas relaciones.

Así la propuesta de Fox se centra en el uso de dos herramientas de análisis básicas, una para cada tipo de discurso. Su propuesta sobre las relaciones anafóricas en el discurso conversacional se basa en el análisis derivado del modelo de discurso para conversaciones espontáneas propuesto por Sacks *et al.* [16]. Sin embargo, para el tratamiento de las relaciones anafóricas en el discurso no conversacional (monólogos) utiliza el análisis de la estructura retórica, propuesto en la Teoría de la Estructura Retórica [8] que ha sido diseñada especialmente para el tratamiento de la prosa expositiva.

De acuerdo con los estudios de Fox, la forma básica de distribución de herramientas anafóricas en conversación se rige por la siguiente norma:

"La primera mención de un referente en una secuencia se realiza mediante un sintagma nominal completo. Después de esto, mediante el uso de un pronombre, el hablante subraya un conocimiento de que la secuencia anterior aún no ha quedado cerrada".

De esta forma, Fox plantea una serie de casos en los que la estructura conversacional según el modelo de Sacks *et al.* genera secuencias dentro de las cuales, el uso de los pronombres queda justificado para indicar precisamente que dicha secuencia aún está abierta.

Estos estudios fueron extendidos y llevados a la práctica por Palomar y Martínez-Barco en [9] y [12] a través de la definición de un *Espacio de Accesibilidad Anafórica Estructural* aplicado a la resolución de la anáfora en discurso dialogado. Tal y como demuestran estos trabajos, el hablante de una conversación utiliza la estructura jerárquica del discurso para generar relaciones entre sus unidades que definan secuencias de discurso coherentes. Pero además, por la propia definición de coherencia, el hablante necesita transmitir al oyente la idea de que esa secuencia no ha finalizado aún, para lo cual hace uso de ciertos mecanismos. Uno de esos mecanismos es la generación de anáforas. Así, si se conocen las secuencias posibles establecidas dentro del discurso, se puede conocer el espacio donde el hablante tiene un "ámbito de trabajo" para hacer referencia explícita a un elemento y después referirse a él mediante el uso de una anáfora, o lo que es lo mismo, se puede conocer el espacio de accesibilidad anafórica definido para una determinada anáfora. Este mismo espacio será, por tanto, el que use el oyente para buscar el antecedente.

Los estudios anteriores sirvieron a los autores para construir un sistema de resolución automática de la anáfora en los diálogos a través de la detección automática de un *espacio de accesibilidad anafórica estructural* para cada anáfora detectada, desde el cual se extraían los candidatos posibles a ser antecedentes en los casos de anáforas pronominales y adjetivas.

Estos espacios se basaban en la detección de secuencias de unidades que se encontraban en dos niveles distintos de la estructura jerárquica del discurso: en el nivel local, formado por secuencias de turnos de habla dentro de un *par adyacente*[1], y por secuencias de pares adyacentes consecutivos; y en el nivel global, formado por secuencias de pares adyacentes que hacen referencia a un mismo tópico de conversación.

Mediante el uso de dicho *espacio de accesibilidad anafórica estructural*, los autores demostraron conseguir mejores resultados que mediante el uso de otros espacios de accesibilidad utilizados por otros métodos de resolución de la anáfora. Este es el caso de los sistemas que usan un *espacio de accesibilidad anafórica completo*: aquellos que toman para cada anáfora todos los candidatos encontrados desde el inicio del discurso hasta el momento en el que se encuentra la anáfora. Otro tipo son aquellos que realizan sus búsquedas en un *espacio de accesibilidad anafórica basado en ventanas de oraciones* en el cual, los candidatos se toman desde una ventana deslizante de oraciones con un tamaño definido para cada tipo de anáfora (generalmente mediante el estudio de corpus). Los primeros presentaban el problema de generar demasiados candidatos para una determinada anáfora conforme avanzaba el discurso, generando no sólo un coste computacional considerable, sino también elevando exponencialmente las posibilidades de producir errores en la respuesta. Los segundos, por su parte, suplían estas deficiencias, pero por contra presentan el problema de la falta de un criterio estable para definir el tamaño de la ventana, ya que, al no obedecer a razones estructurales, podría variar dependiendo del dominio y del corpus que se considerara para su estudio.

De esta forma, la propuesta de Fox para los diálogos se ha demostrado útil, y ha sido fácilmente aplicada a la resolución de la anáfora ya que el modelo de estructura del discurso definido por Sacks *et al.* se puede adaptar a cualquiera de los modelos usados por la mayoría de los sistemas de procesamiento automático de diálogos. De hecho, la construcción de un par adyacente se basa

[1] En el modelo del discurso de Sacks *et al.* [16], tal y como lo presenta Gallardo [5] en su adaptación al español, los pares adyacentes se definen como una agrupación de turnos de habla encabezados por una turno de intervención iniciativa y su correspondiente turno de intervención reactiva. Ello conforma pares de pregunta-respuesta, acción-reacción, invitación-aceptación/declinación, etc.

en el reconocimiento de los turnos de habla, unidades identificables sin necesidad de consideraciones metalingüísticas ya que el simple cambio de hablante indica el establecimiento de un nuevo turno.

Sin embargo, la propuesta de Fox para el establecimiento de relaciones entre la anáfora y la estructura del discurso en el monólogo utiliza un modelo mucho más complejo de implementar como es el de la Teoría de la Estructura Retórica. De acuerdo con esta propuesta, se pueden establecer relaciones entre anáforas y antecedentes que se encuentran en una secuencia de unidades relacionadas por un esquema del tipo afirmación-evidencia (llamado de núcleo y satélite), así como relaciones que siguen un esquema de secuencialidad (llamado multinuclear). En el primer caso se introduce una unidad en el texto (núcleo) que posteriormente es explicada por medio de otra unidad (satélite) referida a la primera a través de diferentes tipos de relaciones (que pueden ser relaciones de evidencia, de interpretación, justificación, etc.). En el segundo caso se introduce una unidad (núcleo) que es seguida de otras unidades (otros núcleos) relacionadas por condiciones de secuencialidad (contrastes, listas, uniones, etc.).

En cualquiera de los casos, el establecimiento de un espacio de accesibilidad de forma automática pasaría por reconocer las diferentes unidades y establecer las relaciones entre ellas para así determinar qué unidades pueden compartir entre sí anáforas y antecedentes. Sin embargo, en este caso no es sólo una dificultad el establecer el rol de cada unidad (algo que necesita cuanto menos del uso de información semántica y pragmática), sino que en sí mismo ya es complejo realizar una segmentación del texto en unidades, puesto que el reconocimiento de los límites de éstas, a diferencia de los turnos de habla, sí necesita de información metalingüística.

A estas dificultades se le añade el hecho de que, si bien el modelo de discurso de Sacks *et al.* ya ha sido estudiado e implementado en la mayoría de los sistemas que procesan diálogos automáticamente, sin embargo, el uso del modelo de la Teoría de la Estructura Retórica no se ha generalizado en la práctica de los sistemas de procesamiento automático de monólogos[2].

De esta forma, se hace necesario buscar un mecanismo que nos permita crear relaciones evidentes entre la anáfora y la estructura del discurso en el monólogo, de forma que esta estructura sea fácilmente identificable (implementable), o como es el caso de nuestra propuesta, que incluso esté explícita en el propio texto que se va a tratar.

3 *Propuesta de un espacio de accesibilidad anafórica para textos etiquetados en HTML*

En este trabajo, dado que nos centramos en textos escritos por un solo productor (tipo monólogo), vamos a tomar como unidad textual formal[3] fundamental el párrafo. Son sobre todo dos razones las que justifican la elección del párrafo como unidad textual fundamental: en primer lugar, con independencia de su condición tipográfica, el párrafo constituye en sí mismo una sub-unidad textual semántica, dado que suele responder a un sub-tópico determinado [6], dentro del conjunto de tópicos y sub-tópicos que constituyen la macroestructura de un texto [18]. En segundo lugar, gracias a su condición tipográfica, el párrafo se presenta al receptor como un conjunto de oraciones dadas por el productor del texto como independientes, y es éste, precisamente, el criterio que utiliza Petöfi para determinar las unidades textuales [14] [11]. Por lo tanto, vamos a considerar el párrafo como unidad formal estructural, dado que responde, en principio, tanto a una unidad de sentido como a la intención lingüístico-textual y comunicativa del productor del texto.

En la estructura general del texto escrito, los párrafos se combinan entre sí formando unidades estructurales superiores. Con ello se forman secuencias de párrafos que pueden estar constituidas tanto por dos párrafos, que mantienen entre sí una relación de núcleo - satélite, o por más de dos [15]. Estas secuencias de párrafos no las vamos a considerar aquí, dado que no vienen especificadas en las etiquetas HTML del documento electrónico. Únicamente nos centraremos en la agrupación en parejas de párrafos, de tal manera que tomaremos como posible espacio

[2] Si bien la Teoría de la Estructura Retórica tuvo su origen en el estudio de la producción automática de textos, no existe una generalización de su aplicación a la interpretación de los mismos

[3] En este contexto usamos el término "formal" para hacer referencia a que estas unidades se podrán establecer con independencia de su significado.

de accesibilidad anafórico estructural no sólo la propia unidad donde se localiza el elemento anafórico (el propio párrafo), sino también la unidad estructural inmediatamente anterior (su párrafo anterior).

Por encima de estas unidades (el párrafo y la pareja de párrafos), la siguiente unidad estructural es, desde un punto de vista jerárquico, el epígrafe. Esta unidad se presenta determinada formalmente por aparecer bajo un subtítulo dentro del cuerpo del documento. Desde un punto de vista semántico, este subtítulo suele especificar, en principio, el sub-tópico o el marco del sub-tópico de esta unidad textual. Por esta razón se puede considerar un importante ámbito de localización del posible antecedente (o antecedentes) de los elementos anafóricos que aparezcan en el epígrafe.

Por último, la unidad superior que vamos a tomar en consideración será el texto completo, que puede estar constituido o bien por una secuencia de párrafos, o bien por una secuencia de epígrafes, según si el documento presenta o no subtítulos. Al igual que estos sub-títulos, el título general del documento también lo vamos a considerar aquí como un importante ámbito de localización del antecedente, dado que suele indicar el tópico general global del texto, o, por lo menos, el marco referencial donde se inserta ese tópico.

En conclusión, las unidades estructurales que vamos a determinar como espacio de accesibilidad anafórica son: el párrafo donde aparece el elemento anafórico, el párrafo anterior, el subtítulo del epígrafe (si existe) y el título general del documento. Cada una de estas unidades viene especificada por el lenguaje HTML, por lo que vienen dadas por el propio texto, sin necesidad de aportar ningún tipo de información estructural adicional al documento.

Además, destacaremos el caso especial de las referencias temporales, un tipo de descripción definida que suele hacer referencia en la mayoría de sus ocurrencias a la fecha del documento, tal y como se demostró en el trabajo de Saquete y Martínez-Barco [17]. De esta forma, la extensión del espacio de accesibilidad anafórica a la fecha del documento que viene especificada dentro de la propia estructura HTML, permite la resolución de dichas expresiones.

Sin embargo, cada tipo de expresión anafórica viene caracterizada por un espacio de accesibilidad propio. A continuación vamos a presentar una propuesta de la determinación del espacio de accesibilidad para cada tipo de expresión anafórica basada en una experimentación sobre treinta documentos HTML. Así, se han considerado cinco tipos de expresiones anafóricas: pronombres, anáforas adjetivas[4], descripciones definidas[5] (DD), alias[6] y expresiones temporales[7].

4 *Experimentación*

Para la experimentación de esta propuesta se ha utilizado un corpus formado por artículos de periódicos digitales de diferentes dominios (deportes, sociedad, internacional, economía, etc.). La cantidad de expresiones anafóricas que se encuentran en los 30 artículos que forman el corpus se muestran en la tabla 1. En este corpus se observa que hay un total de 577 expresiones anafóricas distribuidas de la siguiente forma: 330 de ellas son descripciones definidas (DD), 70 alias, 88 pronombres (no se ha hecho distinción entre los diferentes tipos de pronombres), 16 anáforas adjetivas y 73 expresiones temporales.

Expresión anafórica	**Cantidad**
DD	330
Alias	70
Pronombres	88
Adjetivos	16
Expresiones temporales	73
Total	**577**

Tabla 1: Distribución de las expresiones anafóricas en el corpus periodístico

Los artículos periodísticos no tienen un formato fijo. Normalmente este formato depende del periódico y del tema. En líneas generales hemos encontrado mayoritariamente tres formatos de textos diferentes, que son:

- **Formato 1**. En este primer formato se distinguen únicamente dos partes: a)

[4]Generadas por la elipsis del núcleo en un sintagma nominal con modificador de tipo adjetivo

[5]Aquellos sintagmas nominales que son introducidos por un artículo definido o por un demostrativo, como por ejemplo, la operación.

[6]Apariciones de parte de una entidad introducida previamente. Por ejemplo, el Sr. Gómez es un alias de la entidad José Gómez.

[7]Expresiones del tipo ayer, mañana, la semana siguiente.

título de la noticia y b) cuerpo de la noticia, como se muestra en el ejemplo de la figura 1.

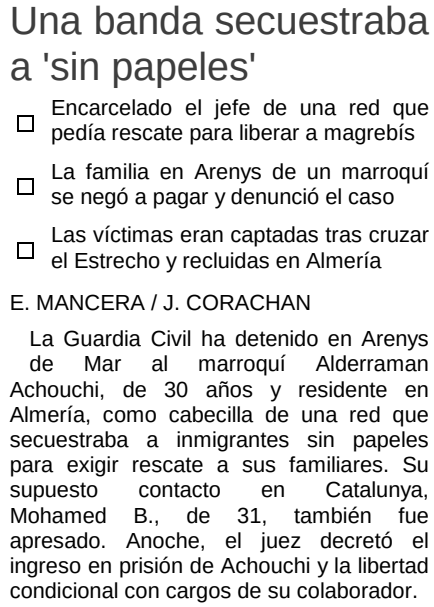
Endesa pacta la prejubilación de casi la mitad de su plantilla

A. FUENTES / F. J. DE PALACIO

Los sindicatos CCOO y UGT y la dirección de Endesa alcanzaron ayer un preacuerdo que prevé la prejubilación de unos 5.000 trabajadores de más de 50 años. Estos prejubilados se suman a los 4.000 que abandonarán la compañía en los próximos dos años, que estaban incluidos en un plan anterior y que afectaba a 8.000 empleados. El resultado es que el 48% de la plantilla actual, que es de 18.000 personas, podría dejar el grupo en los cinco años siguientes, aunque se prevé la contratación de unos 25 empleados por cada 100 prejubilados.

Figura 1: *Formato 1*

- **Formato 2**. En el segundo formato se diferencian tres partes diferentes en el texto: a) título de la noticia, b) subtítulos de la noticia y c) cuerpo de la noticia, como se muestra en el ejemplo de la figura 2.



Una banda secuestraba a 'sin papeles'

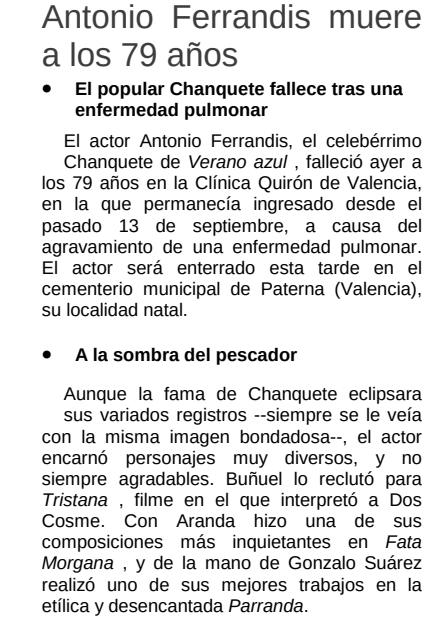
- ☐ Encarcelado el jefe de una red que pedía rescate para liberar a magrebís
- ☐ La familia en Arenys de un marroquí se negó a pagar y denunció el caso
- ☐ Las víctimas eran captadas tras cruzar el Estrecho y recluidas en Almería

E. MANCERA / J. CORACHAN

La Guardia Civil ha detenido en Arenys de Mar al marroquí Alderraman Achouchi, de 30 años y residente en Almería, como cabecilla de una red que secuestraba a inmigrantes sin papeles para exigir rescate a sus familiares. Su supuesto contacto en Catalunya, Mohamed B., de 31, también fue apresado. Anoche, el juez decretó el ingreso en prisión de Achouchi y la libertad condicional con cargos de su colaborador.

Figura 2: *Formato 2*

- **Formato 3**. En el tercer formato se diferencian las mismas partes que en el formato 2 pero cada subtítulo aparece acompañado del cuerpo de esa noticia. Como se muestra en el ejemplo de la figura 3.



Antonio Ferrandis muere a los 79 años

- **El popular Chanquete fallece tras una enfermedad pulmonar**

El actor Antonio Ferrandis, el celebérrimo Chanquete de *Verano azul* , falleció ayer a los 79 años en la Clínica Quirón de Valencia, en la que permanecía ingresado desde el pasado 13 de septiembre, a causa del agravamiento de una enfermedad pulmonar. El actor será enterrado esta tarde en el cementerio municipal de Paterna (Valencia), su localidad natal.

- **A la sombra del pescador**

Aunque la fama de Chanquete eclipsara sus variados registros --siempre se le veía con la misma imagen bondadosa--, el actor encarnó personajes muy diversos, y no siempre agradables. Buñuel lo reclutó para *Tristana* , filme en el que interpretó a Dos Cosme. Con Aranda hizo una de sus composiciones más inquietantes en *Fata Morgana* , y de la mano de Gonzalo Suárez realizó uno de sus mejores trabajos en la etílica y desencantada *Parranda*.

Figura 3: *Formato 3*

Sin embargo, pese a la evidencia de diferentes formatos, podemos concluir que el espacio de accesibilidad anafórico será el mismo independientemente del formato y dependerá únicamente del tipo de anáfora tratada.

Como se ha comentado anteriormente, la experimentación se ha centrado en estudiar la situación del antecedente de cada tipo de expresión anafórica en el texto. Así, se han diferenciado los siguientes casos:

- **T**. Antecedente en el título
- **ST**. Antecedente en el subtítulo
- **P**. Antecedente en el párrafo anterior
- **F**. Antecedente toma como base la fecha del artículo
- **O**. Antecedente no está ni en el párrafo anterior ni en el título. Normalmente varios párrafos antes.

La experimentación realizada ha proporcionado los resultados que se muestran en la tabla 2. Los pronombres realizan mayoritariamente referencias a antecedentes que están en el mismo párrafo que él (83 veces de 88), pero en ocasiones hacen referencias a entidades que están en el párrafo anterior. Estos casos ocurren cuando el pronombre está en la primera frase del párrafo y el antecedente en

la última del párrafo anterior. Las descripciones definidas hacen referencia a antecedentes que se encuentran en cualquier parte del texto: 81 de ellas a antecedentes en el mismo párrafo, 116 a antecedentes del párrafo anterior, 63 a antecedentes que se encuentran en el título, 43 a antecedentes que se encuentran en los subtítulos y 74 a antecedentes alejados más de un párrafo y que no son ni el título ni el subtítulo. En el caso de los alias ocurre lo mismo que para las DD, es decir hace referencia a cualquier parte del texto. Las anáforas adjetivas tienen un tratamiento similar al pronombre, normalmente su antecedente está en el mismo párrafo. En las expresiones temporales su antecedente más usual suele ser la fecha del artículo, aunque también se hacen referencia a antecedentes que se encuentran en cualquier parte del texto.

A partir de los resultados obtenidos se establecen las siguientes reglas para establecer el espacio de accesibilidad en cada caso:

- El espacio de accesibilidad para los pronombres y las anáforas adjetivas se establece en el mismo párrafo salvo que aparezcan en la primera frase del párrafo. En este caso se toma como espacio de búsqueda del antecedente correcto el párrafo anterior.
- El espacio de accesibilidad para las DD, alias, y expresiones temporales se establece en el párrafo actual más el anterior y los sintagmas nominales que se encuentren en el título y en el subtítulo que se corresponden con el tópico y subtópicos. Se observa de los resultados mostrados en la tabla 2 que en 89 casos el antecedente se encuentra alejado más de un párrafo. Estos casos son producidos por DD que son en la mayoría de los casos inferencias del título como por ejemplo

  La mafia rusa asesina a dos personas... El asesinato tuvo lugar...

  En este ejemplo, la DD el asesinato es una inferencia de la acción indicada por el verbo asesinar.

5 Conclusiones

La mayoría de algoritmos que se pueden encontrar en la literatura tradicional utilizan un espacio de accesibilidad basado en ventanas formadas por un número determinado de oraciones previas a la aparición de la expresión anafórica. Nuestra propuesta se fundamenta en criterios lingüísticos a partir de la utilización de la información sobre la estructura del discurso. De esta forma se puede precisar o especificar con mayor exactitud el espacio de accesibilidad anafórica. Al igual que los métodos basados en ventanas de oraciones, nuestra propuesta establece un espacio distinto para diferentes tipos de expresiones anafóricas.

Dadas las características de los textos tomados en consideración, la información sobre la estructura del discurso ha sido obtenida exclusivamente a partir de las marcas de etiquetado utilizadas por los documentos HTML.

A diferencia de otras propuestas de espacios de accesibilidad, nuestra propuesta no hace una definición arbitraria basada en unidades sintácticas (oraciones) sino en unidades de la estructura del discurso. Mientras que las primeras, al carecer de base lingüística, dependen únicamente de la información estadística extraída del estudio de algún corpus, nuestra propuesta permite establecer un criterio consistente independiente del corpus que se trate.

Actualmente, este trabajo sirve como punto de partida para el desarrollo de un sistema completo de resolución de la anáfora que utiliza la infraestructura creada por otros sistemas [3] [13] [10]. Así se pretende mejorar los resultados obtenidos eliminando las inconsistencias estructurales ocasionadas por una definición inadecuada del espacio de accesibilidad.

Referencias

[1] R. Beaugrande and W.U. Dressler. *Einführung in die Textlinguistik.* Max Niemeyer Verlag, Tübingen (Germany), 1972.

[2] E. Bernárdez. *Teoría y epistemología del texto.* Cátedra, Madrid, 1995.

[3] A. Ferrández, M. Palomar, and L. Moreno. An empirical approach to Spanish anaphora resolution. *Machine Traslation*, 14(2-3), 1999.

[4] B. Fox. *Discourse Structure and Anaphora.* Written and conversational English. Cambridge Studies in Linguistics. Cambridge University Press, Cambridge, 1987.

[5] B. Gallardo. *Análisis conversacional y pragmática del receptor.* Colección Si-

Expresión anafórica	Total	P	PA	T	ST	O	F
DD	330	81	116	63	43	74	
Alias	70	27	23	10	9	13	
Pronombres	88	83	5	0	0	0	
Adjetivos	16	13	2	0	0	1	
Expresiones temporales	73	2	1	1	2	1	65
Total	**577**	206	147	74	54	89	65

Tabla 2: Situación del antecedente de una expresión anfórica en el texto

napsis. Ediciones Episteme, S.L., Valencia, 1996.

[6] A. García and T. Albaladejo. Estructura composicional. Macroestructura. *Estudios de Lingüística de la Universidad de Alicante*, 1:127–180, 1983.

[7] M.A.K. Halliday and R. Hasan. *Cohesion in English.* Longman, London/New York, 1976.

[8] W.C. Mann and S.A. Thompson. Rethorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[9] P. Martínez-Barco. *Resolución computacional de la anáfora en diálogos: estructura del discurso y conocimiento lingüístico.* PhD thesis, Universidad de Alicante, Departamento de Lenguajes y Sistemas Informáticos, 2001.

[10] R. Muñoz, M. Palomar, and A. Ferrández. Processing of Spanish Definite Descriptions. In O. Cairo, L.E. Sucar, and F.J. Cantu, editors, *MICAI 2000: Advances in Artificial Intelligence*, volume 1793 of *Lecture Notes in Artificial Intelligence*, pages 526–537, Acapulco, México, 2000. Springer-Verlag.

[11] F.B. Navarro. Introducción a la Textología Semiótica. Bases teóricas para la consideración multimedial del texto. Universidad de Alicante, 2001. Memoria de Licenciatura inédita.

[12] M. Palomar and P. Martínez-Barco. Computational approach to anaphora resolution in Spanish dialogues. *Journal of Research in Artificial Intelligence*, 2001. Accepted to be published.

[13] M. Palomar, M. Saiz-Noeda, R. Muñoz, A. Suárez, P. Martínez-Barco, and A. Montoyo. PHORA: A NLP aystem for Spanish. In Alexander Gelbukh, editor, *Proc. of 2nd International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001)*, Lecture Notes in Artificial Intelligence, pages 128–139, Mexico City, 2001. Springer-Verlag.

[14] J.S. Petöfi. La lingua come mezzo di comunicaziones scrita: il testo. *Sistemi segnici e loro uso nella comunicaziones umana*, 3:66–107, 1990.

[15] M. Pérez. *Rutinas de la escritura. Un estudio perceptivo de la unidad párrafo.* LynX. Universidad de Valencia/University of Minnesota, Valencia, 1998.

[16] H. Sacks, E. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn taking for conversation. *Language*, 50(4):696–735, 1974.

[17] E. Saquete and P. Martínez-Barco. Grammar Specification for the Recognition of Temporal Expressions. In *Proceedings of the Machine Translation and multilingual applications in the new millennium, MT2000*, pages 21.1–21.7, Exeter, UK, 2000.

[18] T. van Dijk. *Text and context: explorations in the semantics and pragmatics of discourse.* Longman, London, 1977.