

Periodismo y web semántica: retos y propuesta de un periodismo estructurado mediante las 5 Ws

Journalism and semantic web: challenges and proposal of a structured journalism through the 5 Ws

Aingeru Genaut. *Universidad del País Vasco*

<https://orcid.org/0000-0003-3762-3570>

<https://doi.org/10.61283/r285fk77>

Recepción: 16.03.2024

Aceptación: 25.05.2024

Publicación: 30.06.24

RESUMEN

El Big Data y el Procesamiento del Lenguaje Natural abren nuevas vías de redacción, edición, clasificación, difusión y recuperación para los textos periodísticos. Sin embargo, el mundo académico le ha prestado aún una atención limitada frente a las aportaciones de otras áreas como la lingüística, la biblioteconomía y, obviamente, la informática. Sin embargo, el periodismo tiene mucho que ofrecer y que recibir de la web semántica. El presente trabajo presenta el estado de la cuestión más allá de las aportaciones del ámbito de la comunicación periodística, buscando elementos comunes entre ellas, para pasar después a exponer una forma de abordar el reto de estructurar la información periodística sin que esta resulte disruptiva o rompa con la tradición académica ni con la profesión –desde la teoría del discurso a los libros de estilo y la redacción–. Se propone acercar la web semántica a la práctica periodística utilizando las cinco Ws del periodismo como criterios de identificación y clasificación del contenido. Al utilizar las Ws del periodismo –*Qué, Quién, Dónde, Cuándo, Cómo y Por Qué*– como criterios de identificación y clasificación, el contenido quedaría organizado en torno a su significado –a su función semántica–, aprovechando las especificidades de la estructura narrativa de la noticia.

Palabras clave: web semántica, periodismo estructurado, 5 W, big data, procesamiento del lenguaje natural.

ABSTRACT

Big Data and Natural Language Processing offer new ways of writing, editing, classifying, diffusion and retrieving journalistic texts. However, the academic world has still paid limited attention to it, compared to the contributions of other areas such as linguistics, library sciences and, obviously, computer sciences. However, journalism has much to offer – and to receive – from the semantic web. The present work presents the state of the question beyond the contributions of the field of journalistic communication, looking for common elements between them, to then go on to expose a way of approaching the challenge of structuring journalistic information without it being disruptive or breaking neither with the academic tradition nor with the profession –from discourse theory to style books and writing–. It is proposed to bring the semantic web closer to journalistic practice using the five Ws of journalism as criteria for identifying and classifying content. Using the Ws of journalism –*What, Who, Where, When, How and Why*– as identification and classification criteria, the content would be organized around its meaning –its semantic function–, taking advantage of the specificities of the narrative structure of the news.

Keywords: Semantic web, structured journalism, 5W, big data, natural language processing.

1. Introducción

La idea de una red que vinculara cualquier tipo de información bajo criterios semánticos ha estado presente en Internet y el hipertexto incluso antes de que estos vieran siquiera la luz. El desarrollo del concepto del hipertexto viene ligado a la figura de Vannevar Bush, que en 1945 recoge una idea que ya había sido propuesta por el mismo autor en 1939 en un artículo titulado *Mechanization and the Record*, publicado en la revista *Fortune*. Bush idea el *Memory Extender* o *Memex*; una máquina analógica capaz de recuperar información estructurada a través de la asociación conceptual estableciendo vínculos semánticos entre los términos que componen cualquier información para acercarse así a la forma en la que se desarrolla el pensamiento humano. Casi veinte años más tarde su idea de crear una máquina que permitiera el registro y la recuperación de la información a través de la asociación conceptual influyó notablemente en tres importantes investigadores: Joseph Carl Robnett Licklider, Douglas Engelbart y Theodor Nelson. Licklider desarrollará las ideas de Bush y las acercará al mundo de la tecnología de la computación y la inteligencia artificial, pero tampoco podrá ponerlas en práctica. Douglas Engelbart publica *A conceptual Framework for the information augmentation of the man's intellect* como parte del proyecto *Online System (NLS)* de la Universidad de Stanford, para la creación de un sistema computerizado que materializara las ideas de Bush. El concepto de hipertexto, sin embargo, se lo debemos a Theodor Nelson, que en 1965 publicará *The Hypertext*, y en 1979 ideará *Xanadu*, un sistema de almacenamiento y asociación de información conceptualmente mucho más ambicioso que la actual World Wide Web, pero que tampoco verá la luz. Dos décadas más tarde, entre 1989 y 1991, Tim Berners-Lee desarrollará el *HyperText Markup Language*, pero él mismo admitirá que el HTML no permitía en sus primeras versiones organizar, enlazar y localizar la información de forma semántica, pero en ningún momento renunció a alcanzar esta meta (Berners-Lee, 2000). Para comienzos del milenio, Berners-Lee tiene claro que Internet debe evolucionar hacia la web semántica, y un año más tarde publicará junto con James Hendler y Ora Lassila un influyente artículo que sienta las bases sobre las que debería sustentarse la ordenación y la localización de la información a través del contenido (Berners-Lee *et al.*, 2001).

1. Web semántica y periodismo

1.1. El interés del mundo periodístico en la web semántica

La profesión periodística, por su parte, ha encontrado en Internet un filón inagotable de información para la documentación y la investigación periodística. El constante crecimiento de todos los tipos de Web, desde la 1.0 al incipiente 3.0 ha propiciado hecho que la red se convierta en una inagotable fuente de información que, sin la asistencia de equipos informáticos y del *Big Data*, entendida como “el conjunto de estrategias, tecnologías y sistemas para el almacenamiento, procesamiento, análisis y visualización

de conjuntos de datos complejos” (Casas *et al.*, 2019, p. 34) difícilmente podría ser empleado para elaborar contenidos noticiosos.

Es en este entorno donde aparece el periodismo de datos; un nuevo perfil de periodismo o, más bien, una nueva rutina periodística que emplea fundamentalmente técnicas estadísticas y sociológicas para encontrar, extraer y ordenar gran cantidad de datos — procesos conocidos como *Data Mining*— y, más adelante, analizar y elaborar información útil y socialmente relevante. El periodismo de datos se presenta como una gran oportunidad para el desarrollo de la profesión periodística que, además, no se limita a la minería de datos. Además de saber buscar y extraer información, tanto los redactores como los documentalistas deben ser capaces de distinguir la información relevante de aquella que no lo es. A esta labor se la conoce como curación de contenidos; una labor que precisa de unas competencias “que permiten a un profesional de la comunicación buscar, evaluar y seleccionar fuentes de información de calidad para, a continuación, difundirlas con algún tipo de valor añadido” (Codina, 2016, p. 200). Por último, el periodismo de datos también ofrece una nueva forma de redactar y editar el contenido; una nueva forma de narrar en la que la infografía, la visualización de la información y el *visual thinking* juegan un papel crucial, para el que se necesita “contar con aplicaciones adecuadas y con un personal bien entrenado” (Colle, 2013, p. 7).

Pero *Big Data* tampoco es sinónimo de web semántica. Al igual que con la Web 2.0, el hecho de que la información se encuentre disponible no significa necesariamente que pueda ser localizada, comprensible o útil. De hecho, la mayor parte de la información susceptible de ser empleada por el profesional de la información sigue estando desestructurada. La diferencia se encuentra, una vez más, en la posibilidad de que los datos tengan un significado discernible que vaya más allá de la forma en la que se presenten: una huella semántica y una relación coherente e identificable entre ellas. En otras palabras, para que el *Big Data* forme parte de la Web 3.0, el contenido debe estar estructurado y etiquetado en torno a su significado, pasando del *Big Data* al *Smart Data*.

Pese a que el número de centros documentales y de documentalistas de los medios de comunicación se han ido viendo reducidos de forma constante, los estudios empíricos dedicados a recoger la opinión de los periodistas sobre ellos demuestran que su trabajo no solo sigue estando bien valorados, sino que resulta absolutamente imprescindible para la elaboración de contenidos informativos de calidad (Marcos y Edo, 2015; García *et al.*, 2019). Además, a diferencia de lo que pudiera pensarse, su labor no consiste únicamente en satisfacer los requerimientos informativos puntuales de los periodistas; su principal labor se centra en “la gestión y actualización del fondo documental [del medio]” (García *et al.*, 2019, p. 60). Este fondo documental se alimenta tanto de fuentes internas —de la información elaborada por el propio medio— como de fuentes externas, y se actualiza sobre todo gracias a la labor de rastreo, catalogación y clasificación que diariamente llevan a cabo los documentalistas “en función de lo que marque la agenda informativa, anticipándose así a sus necesidades, y facilitando una rápida consulta de los datos por parte del periodista (Beltrán, 2015, p. 42). Sin embargo, hoy por hoy, la inclusión de

técnicas recuperación del contenido externo al medio y su inclusión en la base documental tiende a llevarse a cabo sin emplear técnicas propias de la web semántica como el etiquetado semántico; un proceso en el que “apenas participan” (García *et al.*, 2019, p. 62). Los criterios empleados para clasificar, organizar y relacionar el contenido documental varían notablemente de un medio a otro al incluir las taxonomías, los tesauros y las ontologías, entre otros métodos. Mientras que las dos primeras cuentan con una larga tradición y su tecnología está consolidada en el mundo de la documentación periodística (Szostak, 2014), las ontologías “están pendientes de desarrollos en nuestro ámbito” (García, 2016, p. 23), debido a las divergencias en su aplicación y a “la pobre concepción terminológica subyacente” (García *et al.*, 2019, pp. 56 y 57).

Los procesos de implementación de técnicas del procesamiento de lenguaje natural también se extienden a las labores de edición y redacción de textos. En este sentido, hay interesantes experiencias, tanto teóricas y metodológicas (Mateo *et al.*, 2003; Plaza y Díaz, 2011; Drndarevic y Saggion, 2012) como prácticas (Lloret y Palomar, 2011) para la simplificación de textos y la elaboración de resúmenes automáticos —*text summarization*—. Sin embargo, frente a los proyectos emprendidos por medios de comunicación estadounidenses, británicos o alemanes, la inclusión de generadores automáticos de textos en medios de comunicación españoles es meramente testimonial, y “no van más allá de iniciativas concretas en servicios *online* de información del tiempo o servicios” (Túñez-López *et al.*, 2018, p. 756; Túñez-López *et al.*, 2019).

En cuanto al uso del lenguaje natural como instrumento de edición y corrección de contenidos, al igual que en el caso anterior, algunos proyectos reales que los integran en el proceso de redacción de la pieza informativa, aunque existen pequeñas *apps* y *plug-ins* como *Retext* o *Triplify*, capaces de incluir algunas características específicas de la web 3.0, como la marcación semántica. En un intento de recuperar la credibilidad en el ejercicio del periodismo, las técnicas de tratamiento del lenguaje natural están siendo empleadas en menor o mayor medida para detectar la veracidad de la información periodística, combatir la información falsa (Conroy *et al.*, 2016; Pérez-Rosas *et al.*, 2017), la desinformación (Rodríguez, 2019), y valorar la calidad del contenido informativo (Ufarte *et al.*, 2018).

Por último, servicios como ChatGPT y, por extensión, la inteligencia artificial, tiene el potencial de transformar profundamente el periodismo, aportando eficiencia, precisión y nuevas oportunidades creativas. Al automatizar tareas rutinarias como la recopilación de datos, la verificación de hechos y la redacción de informes básicos, la IA permite a los periodistas centrarse en la investigación profunda y el análisis crítico. Además, herramientas de IA pueden analizar grandes volúmenes de datos en tiempo real, detectando tendencias y proporcionando tendencias que serían inaccesibles de otro modo. También mejora la personalización de contenido, adaptando las noticias a las preferencias individuales de los lectores. No obstante, es crucial abordar éticamente su implementación para evitar sesgos y preservar la integridad periodística (Apablaza-

Campos, 2023; González-Arias & López-García, 2023; Gutiérrez-Caneda, *et al.*, 2023; Vallés Esteve, 2023).

1.2. *Hacia el periodismo estructurado*

Sin embargo, no hay respuestas claras en lo que se refiere a la praxis periodística; a cómo debe desarrollarse su día a día, puesto que, en realidad, no hay respuestas a cómo estructurar y gestionar la documentación, la edición y contenido. Sin un sistema de etiquetado semántico adaptado al quehacer periodístico o, más concretamente, a la práctica profesional del periodista, difícilmente podrá desarrollarse el periodismo estructurado. Y en el periodismo de las principales cabeceras del país no se desarrolla una política específica de etiquetado, en el mundo académico también “son escasas las investigaciones centradas en el análisis-etiquetado humano” (García *et al.*, 2019, p. 57).

Dicho esto, ¿puede el periodismo ofrecer una perspectiva alternativa, una forma de clasificación más genérica capaz de ser aplicada a cualquier tipo de información o, cuanto menos, al contenido periodístico? Llegados a este punto podemos hacer un breve resumen de los retos a los que se enfrenta un periodismo estructurado:

- a) un profundo cambio en la identificación, clasificación y catalogación de la información en el centro documental, estableciendo una rigurosa política de etiquetado semántico;
- b) un profundo cambio en la edición de la información, integrando en el proceso el etiquetado semántico del contenido;
- c) un profundo cambio en la gestión del contenido, asumiendo que la misma información pueda tener una vida muy larga —al poder ser actualizada y alterada constantemente—, muchas vidas —controlando su difusión a través de muchos canales—, e incluso muchas resurrecciones —integrando un mismo contenido en piezas informativas futuras, sin que pierda sentido ni coherencia—;
- d) y una nueva forma de visualizar y gestionar este tipo de contenido.

Estos cambios o retos coinciden con los presentados por los profesores Freixa, Pérez-Montoro y Codina (2017), quienes tras recoger los rasgos que deberían definir al periodismo estructurado, pasan a presentar algunos ejemplos de este, para concluir que “datos codificados, capas de información e interfaz forman el conjunto de tres niveles que contiene todo sistema de periodismo estructurado” (Freixa *et al.*, 2017, p. 1088). Y si ya habíamos comentado que en el periodismo de las principales cabeceras del país no se desarrollaba una política específica de etiquetado, donde el periodista tuviera control sobre el contenido, en el mundo académico también “son escasas las investigaciones centradas en el análisis-etiquetado humano” (García *et al.*, 2019, p. 57). Pero este no es un problema exclusivo del mundo periodístico; se trata, más bien, del reflejo de un problema estructural de la catalogación semántica: la flexibilidad —necesaria, para que el periodista pueda trabajar en todo tipo de noticias— parece oponerse a la precisión —requerida, en el ejercicio del periodismo—.

1.3. Retos de la clasificación semántica

El procesamiento del lenguaje natural debe hacer frente a un altísimo grado de incertidumbre connatural al lenguaje humano. Para ello, está organizado en torno a tres grandes dimensiones:

- a) la dimensión morfosintáctica: encargada de analizar una frase e identificar su estructura y la naturaleza de los elementos de los que está compuesta;
- b) la dimensión semántica: encargada de identificar, a través de la inducción, el significado de la frase;
- c) la dimensión pragmática: encargada de identificar, a través de la inducción, el sentido del texto.

Para el análisis de la dimensión morfosintáctica, los procesadores del lenguaje natural recurren primero al *Reconocimiento de entidades nombradas*, y para proporcionar un análisis general y adecuado de estas entidades, es necesario clasificarlas correctamente según su clase semántica (persona, lugar, organización, etc.) y, resolver cada una de las menciones con respecto a la entidad nombrada a la que se refiere (Pérez y Cardoso, 2014).

La segunda dimensión tiende a resolverse a través del *Etiquetado de roles semánticos*, y consiste en reconocer los argumentos semánticos de los predicados —verbales y nominales— principales de cada sintagma sintáctico. Este proceso se ha visto complementado con el *Etiquetado de roles semánticos implícitos*, que busca identificar aquellos roles semánticos que ocurren más allá del contexto sintáctico local de su predicado. El problema de la sinonimia intenta ser resuelto mediante la *Resolución de correferencias* y la *Extracción terminológica*, cotejando los resultados obtenidos anteriormente en todo el texto, identificando y creando un índice de términos asociado a cada entidad. Así, se despejan las duplicidades y se elabora, además, una taxonomía para posteriores consultas (Ratinov y Roth 2009; Passos *et al.*, 2014).

La tercera dimensión quedaría, por último, englobada dentro del *Proceso de extracción de relaciones*, mediante la detección e identificación de los posibles vínculos que pudieran existir entre los roles semánticos identificados en la etapa anterior (Galván, 2017).

A medida que nos alejamos de la primera dimensión, el grado de incertidumbre se hace cada vez más extenso, y su resolución, más y más compleja —aunque hoy día los sistemas de procesamiento del lenguaje natural han logrado un gran éxito en la resolución de la primera dimensión, y un respetable grado de aciertos en la segunda—. En cuanto a la tercera, a medida que la información analizada se vuelve más genérica y el contenido no queda circunscrito a un tema específico, el nivel de éxito disminuye notablemente. De hecho, el método más exitoso para mantener un alto grado de eficiencia frente al análisis de información genérica consiste en limitar al máximo el objeto de análisis; los estudios centrados exclusivamente en las emociones o sentimientos presentes en textos genéricos representan un buen ejemplo (Barba *et al.*, 2016; Baviera, 2017; Criado y Villodre, 2018;

Córdoba *et al.*, 2021). Con el procesamiento del lenguaje natural volvemos a encontrarnos, una vez más, con el mismo problema al que nos enfrentábamos con el etiquetado semántico: a medida que la información se aleja de un tema o área determinada, la creación de las ontologías y la identificación de las taxonomías se vuelve más compleja y menos efectiva.

La dificultad fundamental de clasificación del contenido radica en el hecho de que “el significado preciso de las palabras [que lo forman] se debe definir fuera de él (Sánchez y Gil, 2007, p. 552), y esta debe ser obtenida a través de tesauros, taxonomías, folksomías, u ontologías. Los tesauros, por sí mismos, no pueden desvincularse del vocabulario que los conforman (Martínez y Alvite, 2014); hace que sea difícil relacionar, y más aún utilizar, dos o más tesauros creados de forma independiente. Pero para que un ordenador llegue a interactuar sobre el contenido basándose de su significado, debe dar un paso más y aproximarse a los *Lenguajes de Ontologías Web*. Estos recursos “añaden valor a los tesauros tradicionales a través de una semántica más profunda [...] implicar niveles más profundos de jerarquía, unas enriquecidas relaciones entre clases y conceptos, así como la capacidad de formular reglas de inferencia, etc.” (García, 2004, p. 89). Sin embargo, el empleo de los recursos disponibles a través de los datos enlazados sigue siendo considerablemente escaso (Agerri, *et al.*, 2019).

La evidente carencia en la aplicación de estándares para el etiquetado del contenido desarrollados a lo largo de muchos años no puede ser atribuida a la desidia. Aunque la técnica esté desarrollada, definir los criterios para identificar —para nombrar— y relacionar dichas etiquetas resulta extremadamente complejo. No es casualidad que las áreas en las que más eficientemente se han desarrollado estas taxonomías u ontologías sean aquellas en las que el contenido, por muy extenso que sea, adquiere una forma más o menos definida: la biblioteconomía y la documentación —obviamente, puesto que la taxonomía ha formado siempre parte de su ámbito profesional y académico— (Pedraza *et al.*, 2007; Gomis, 2015; Fermoso-García *et al.*, 2018; Arakaki *et al.*, 2019); la medicina —a través del procesamiento de los historiales médicos— (Mungra y Calziani, 2013; Pérez *et al.*, 2014); o la jurisprudencia —mediante el análisis y la organización de los sumarios y las sentencias— (Alvite, 2012; Sancho *et al.*, 2012).

2. Propuesta para un periodismo estructurado

2.1. *Estructurando la información desestructurada: el método periodístico*

Las fases que comprenden la elaboración de un texto periodístico pueden llegar a ser considerablemente complejas, pero tanto la metodología docente como la experiencia adquirida en el ámbito profesional recomiendan ordenar las ideas y exponerlas partiendo de una serie de principios muy simples. Consideramos que tanto la metodología docente seguida desde el mundo académico como la experiencia adquirida en la práctica periodística pueden ayudar a identificar, organizar y establecer una serie de criterios para la elaboración de criterios de clasificación y relación susceptibles de ser empleadas casi

en cualquier tipo de contenido informativo. Uno de los primeros retos a los que debe enfrentarse cualquier estudiante de periodismo —y cualquier periodista durante el ejercicio de su profesión— se encuentra en el mejor modo de comunicar cualquier cosa de forma clara, precisa y concreta, y una de las primeras lecciones que aprende consiste en intentar organizar mentalmente todo tipo de información noticiosa en torno a seis preguntas fundamentales: qué ha sucedido, quién —o quiénes— están involucrados, cuándo ha sucedido, dónde ha sucedido, por qué ha sucedido, y cómo ha sucedido.

Las más que conocidas cinco W's (*What, Who, When, Where* y *Why*) —aunque siempre han sido seis preguntas, si bien la última empieza por H: *How*— no nacen del mundo periodístico, y serán aplicadas por retóricos como Quintiliano o Cicerón como un método de estructurar adecuadamente un discurso. En cualquier caso, la técnica de las cinco Ws ha formado parte de los pilares de la redacción periodística desde sus inicios, y tanto ellas como sus variaciones, precisiones y ampliaciones como el *Según*, el *Para qué* y el *Cuánto* (Grijelmo, 1997, pp. 30 y ss.) se encuentran implícita o explícitamente en todos los manuales de redacción periodística, desde los clásicos (Dovifat, 1960; Warren, 1975; Martín, 1987; Casasús y Núñez, 1991; Martínez, 1993; Grijelmo, 1997; Fontuberta, 2003), hasta los más actuales (Parratt, 2017), sin solución de continuidad. No es casualidad que Mar de Fontuberta entienda estas preguntas como la forma de tener “el mundo en seis respuestas”, antes de precisar qué tipo de información pueden contener (Fontuberta, 2003, pp. 74 y 75):

- Qué: implica los acontecimientos, las acciones e ideas de las que va a informar la noticia.
- Quiénes: son los protagonistas, sus antagonistas, y, en definitiva, todos aquellos personajes que aparecen en la noticia.
- Cuándo: sitúa la acción en un tiempo concreto, señala su inicio, su duración y su final.
- Dónde: delimita el espacio del desarrollo de ellos hechos.
- Por qué: explica al receptor las razones que han motivado el acontecimiento, sus antecedentes, etc. Además, introduce en muchos casos elementos de valoración que superan la simple descripción de los acontecimientos.
- Cómo: describe las circunstancias y las modalidades que han revestido los hechos.

Sin embargo, la formación del periodista y el desarrollo de sus competencias profesionales alrededor de estas preguntas tienden a madurar a través de la formación conceptual, confiando en su capacidad de aprehensión y su razonamiento abstracto para darles forma. En otras palabras, aunque la construcción de las piezas informativas se sustenta en torno a estas preguntas, la forma en la que se proyectan —pasan de un concepto a un texto —entendiendo el texto como cualquier producto de la actividad lingüística—, no está parametrizada y se confía en que la experiencia y la habilidad del periodista —en que su *saber hacer*; su *oficio*—, sean capaces de llevar a cabo esta traslación con éxito. Pero si se pretende llevar a cabo esta traslación utilizando como mediador —o, en casos extremos, como generador y administrador— un ordenador, el proceso no puede depender del pensamiento abstracto. Se debe crear un puente; un

sistema de relaciones que conecte aquello que una máquina es capaz de entender, y lo que, en nuestro caso, un periodista entiende cuando construye una pieza informativa.

2.2 La estructura fundamental de la información: 5Ws+1

De todas las Ws que podamos encontrar en los manuales de redacción periodística, podemos distinguir dos grupos claramente diferenciados: las simples y las relacionales. Cuando hablamos de las simples nos referimos al *Quién*, al *Cuándo*, al *Dónde*, al *Cuánto* y al *Qué*.

Las cuatro primeras poseen, desde el punto de vista léxico, una forma muy marcada en el texto, y las aplicaciones del procesamiento del lenguaje natural las tienen considerablemente identificadas. Gracias sobre todo al uso de recursos taxonómicos relativamente cerrados, el problema de la sinonimia está bastante resuelto —recordemos que las taxonomías son más eficientes cuanto más limitado sea el ámbito en el que actúen—, y las marcas de tiempo, lugar y cantidad entran dentro de este campo. En lo que respecta a los nombres de personas o instituciones, el empleo de acrónimos, abreviaturas, cargos o alias complican considerablemente la labor de las taxonomías, y se tiende a recurrir a sistemas más complejos. Sin embargo, como hemos descrito anteriormente, los extractores de entidades son cada vez más eficientes en la labor de identificación y marcación semántica de este tipo de elementos. Sin embargo, hay un lugar al que estos extractores de entidades no llegan. Aunque tradicionalmente solo se habla de un *Quién*, esta W encierra dos elementos fundamentales: *Quién (Who)*, y *A quién (Whom)*. Los extractores de entidades no son capaces de diferenciarlos, pero los asignadores de roles semánticos sí.

Esta propuesta sigue la senda de aquellos autores que defienden la pertinencia del periodismo estructurado, y que buscan “cambiar los procedimientos tradicionales [...] por otros que permitan codificar con mayor precisión los datos propios de los acontecimientos, como el lugar, la fecha, los personajes, etc., permitirá mejorar la gestión, la búsqueda y la vinculación entre informaciones” (Freixa *et al.*, 2017, p. 1076). Sin embargo, a partir de este momento, nuestra propuesta tomará un camino alternativo para llevarlo a término.

Siguiendo con las W elementales, entendemos el *Qué* como una W mixta: puede formar parte de los elementos simples o de los compuestos. El *Qué* simple representa, sobre todo en periodismo, la pregunta nuclear del acto de comunicación, y son pocas las excepciones en las que esta pregunta descansa fuera del verbo. Las noticias hablan de hechos, de acontecimientos, y el verbo es su punto de inicio, ocupando el puesto central en la estructura fundamental de las Ws. Merece la pena señalar que hace tiempo que la biblioteconomía descubrió el valor del verbo como elemento contextualizador, en detrimento del tradicional sustantivo (García, 2004). Sin embargo, los extractores de entidades suelen ignorarlo porque, paradójicamente, al estar diseñados para asignar una huella o marca semántica que pueda aplicarse en cualquier circunstancia, renuncian al elemento que los contextualiza dentro de un texto. De ahí que los extractores de entidades

presenten las marcas de lugares, personas, instituciones, etc., pero sin ninguna mención a la relación que existía entre ellas -al verbo que las unía y les daba sentido semántico-.

Nuestra propuesta parte de la premisa de que una narración puede y debe ser desestructurada sin perder el contexto —las relaciones de dependencia— de y entre los elementos que lo forman. Para ello, es imperativo que en ningún momento se llegue a romper la relación entre los elementos que conforman su estructura fundamental, y el elemento que las mantiene unidas —que les dota de sentido— es, precisamente, el *Qué*: el verbo. En todo caso, contamos una vez más con los analizadores morfosintácticos, que no solo tienen en cuenta el verbo, sino que, obviamente, lo consideran el punto de partida de todo análisis sintáctico.

2.3 Los niveles de las estructuras: las *W's* relacionales

A diferencia de las *W's* elementales, las preguntas *Cómo*, *Por qué* y, en ocasiones, el *Qué*, no tienen una estructura propia. De hecho, defendemos que en realidad se trata de una o varias estructuras fundamentales relacionadas, y el *Cómo*, el *Por qué*, el *Para qué*, y algunos *Qué*, no son más que las relaciones que se establecen entre ellas, y que dan sentido semántico a la pieza periodística. En este sentido, resulta muy sugerente observar que un planteamiento de este tipo se ajusta perfectamente a la aplicación del Big Data y a la representación de la información a través de los grafos, puesto que la relación entre los elementos que clasifican es tan importante como los elementos mismos —de hecho, no pueden existir sin ambas dimensiones—. Obsérvese que los grafos poseen tres tipos de información interrelacionados (Casas *et al.*, 2019, pp. 245 y ss.):

- a) los nodos o vértices, que usualmente almacenan la información material o real — que en nuestro caso se correspondería con la información de la pieza periodística;
- b) las aristas, que establece la relación entre los nodos —que debería corresponderse con la relación sintáctica/semántica entre los nodos, como veremos más adelante—; y
- c) las etiquetas, que resumen y facilitan la lectura y la recuperación de la información —que se definirían con las *W's* del periodismo—

Frente a las estructuras fundamentales, las *Ws* relacionales varían en cada caso, y un mismo *Qué* fundamental puede tener una estructura relacional completamente distinta. Convertir algo tan complejo como la comunicación no estructurada en una serie de reglas medibles, objetivables y clasificables no es tarea sencilla. De hecho, el número niveles de estructuras relacionales, desde la más simple —ubicada a nivel sintagmático— a la más compleja —a la altura de la pieza informativa— puede llegar a ser inabarcable si no se establecen reglas estrictas que determinen estas relaciones y sus niveles de dependencia, aunque el mero hecho de identificarlas y trasladar estas reglas a un soporte informático ya sería un logro considerable. Afortunadamente, siempre han existido recursos formales —físicos, identificables y medibles— que pueden facilitarnos la tarea; el contenido resaltado en el encabezado, el titular —antetítulo y subtítulo—, el lead o entradilla, o los ladillos pueden ser tomados como referencia para llevar a cabo una primera clasificación

de la estructura del texto; y si se desea llevar a cabo un análisis más profundo, tomar el párrafo como referencia es también un magnífico recurso (Vera, 2012). Esto no es fruto de la casualidad; responde al hecho de que la información periodística posee una estructura y una serie de reglas específicas.

2.4. *La estructura del discurso periodístico*

El análisis de la estructura que subyace en la información periodística es un campo de estudio de larga tradición en el mundo académico, aunque las aportaciones teóricas más interesantes son considerablemente antiguas, y hunden sus raíces en la semiótica, la lingüística y el análisis del discurso. En este sentido, cabe mencionar especialmente a Teun Van Dijk; y aunque su principal objeto de estudio se basa en la pragmática —en el contexto del texto, más allá del mismo— (Van Dijk, 1980), siempre ha sostenido que la noticia periodística es un tipo específico de discurso (Van Dijk, 1990), y que además de su contexto, la dimensión puramente textual de la noticia responde a una serie de estructuras y reglas específicas, ubicadas en el texto mismo, frente a otro tipo de información. Dicho autor defiende que el discurso periodístico:

- a) se organiza en torno a microestructuras narrativas: las proposiciones; “los constructos de significado más pequeños e independientes de del lenguaje y el pensamiento” (Van Dijk, 1990, p. 54) presentes en el texto, —y que en nuestro esquema podrían equipararse a las estructuras fundamentales de las 5W’s—, están conectados entre sí, y se suelen presentar de forma secuencial “para explicar los significados del discurso [...] expresadas en distintas oraciones” (Van Dijk, 1990, p. 93) —la proximidad de las estructuras fundamentales es, por lo tanto, importante—;
- b) por encima de ellas se encuentran las macroestructuras narrativas: formadas por macroproposiciones, este tipo de subtemas del discurso periodístico forman la estructura jerárquica de la narración, la dividen en distintos niveles, y se encuentran sometidas a una serie de macrorreglas, que determinan las relaciones entre ellas. En el caso del discurso periodístico, estas macrorreglas, denominadas por el autor reglas de especificación, “la información abstracta, de alto nivel, se especifica de tal modo que, para los acontecimientos o acciones totales, se apliquen descripciones detalladas a la identidad y las características de los participantes, las condiciones, los componentes y las consecuencias de la acción, el momento, el lugar o modo de los acontecimientos y diferentes tipos de circunstancias” (Van Dijk, 1990, p. 71);
- c) una noticia periodística puede estructurarse temáticamente en múltiples niveles, desde el más bajo —una proposición— al más alto —el tema de la pieza informativa—, pasando por niveles intermedios de macroestructuras narrativas como los antecedentes, las consecuencias, o las causas del tema (Van Dijk, 1990, pp. 68 y ss.) —que se corresponden con las W’s compuestas como el *Por Qué*, el *Para Qué*, el *Qué* o el *Cómo*—; y

- d) que en el discurso periodístico, muchas de estas macroestructuras tienden a tener una forma y una ubicación determinada más o menos estable. Así, los ladillos, los párrafos, el lead o el titular pueden ayudar a estructurar el contenido y a identificar su estructura narrativa (Van Dijk, 1990, pp. 61 y ss.).

Obviamente, podemos encontrar investigaciones más recientes acerca de la estructura narrativa del discurso público —y la información periodística forma parte de él— (Rodríguez, 2013; Mas, 2016a), pero en los últimos años no existen muchos trabajos de este tipo relacionados directamente con el periodismo, salvo algunas contadas excepciones (Bueno, 2000; García, 2014; Hernando, 2015; Mas, 2016b).

3. Premisas para una forma de periodismo estructurado

En resumen, dadas las características propias de los textos periodísticos, estos pueden ser estructurados semánticamente tomando como punto de partida las W's del periodismo, y que dicha estructura puede ser trasladada al mundo digital con un alto grado de automatización gracias a las tecnologías del *Big Data* y del *Procesamiento del Lenguaje Natural*.

Para ello se parte de las siguientes premisas:

- a) las W's fundamentales forman parte de las proposiciones de los textos periodísticos;
- b) las W's fundamentales que forman parte de una proposición se relacionan con otras proposiciones a través de las W's compuestas;
- c) las W's compuestas estructuran semánticamente todo el texto, desde la secuencia de proposiciones al tema principal de la noticia, a distintos niveles;
- d) la razón por la que resulta tan difícil determinar la ubicación de los W's compuestos radica en que pueden no existir sintácticamente en el texto como tales; pueden ser una construcción de carácter puramente semántico —abstracto—. Sin embargo, el discurso periodístico posee una serie de reglas bastante específicas —ubicación, orden, titulación, entradillas...— de tipo sintáctico y físico que actúan como huellas o guías para identificar las W's compuestas y jerarquizarlas.

En cuanto a la integración de dichas premisas con el *Big Data* y el *Procesamiento del Lenguaje Natural*:

- a) las W's fundamentales pueden obtenerse a través de los Procesadores del Lenguaje Natural porque se corresponden con las categorías gramaticales que dichos programas son capaces de extraer —poseen una huella sintáctica inequívoca—;
- b) las W's compuestas pueden obtenerse relacionando los grupos de W's fundamentales —las proposiciones—. Dicha relación se obtiene de los

- Procesadores de Lenguaje Natural, puesto que se corresponden con la categoría léxica y sintagmática de los analizadores sintácticos;
- c) el establecimiento de niveles semánticos entre las proposiciones puede obtenerse gracias a las aplicaciones *Big Data*. Estos son capaces de analizar gran cantidad de textos para buscar patrones y huellas de la jerarquía semántica de un texto periodístico;
 - d) enlazar los distintos niveles semánticos mediante grafos permitiría, además, visualizar —e incluso, seleccionar, mover, o editar— la estructura narrativa del texto periodístico.

El primer paso de esta propuesta consiste en demostrar que, efectivamente, las W's simples poseen una clara y distintiva huella semántica. Si, además, se busca automatizar el proceso de etiquetado, esta huella debe ser identificada y aislada a través de los analizadores semánticos. De todos los posibles géneros periodísticos, secciones o temas periodísticos susceptibles de ser analizados, hemos elegido las piezas periodísticas relativas a sucesos. Tradicionalmente, los sucesos cubren una serie de temas fácilmente identificables como catástrofes naturales, accidentes, delincuencia y crímenes (Quesada, 2007: 18 y ss.) y, a excepción de la cobertura de tribunales —área tradicionalmente asociada a sucesos—, todas poseen las siguientes características:

- a) son acontecimientos imprevistos, y su contexto —la información que se encuentra más allá del evento— es limitado;
- b) son, debido a su naturaleza, los textos periodísticos que mejor se ajustan a la estructura de la pirámide invertida; y
- c) son, ante todo, piezas periodísticas descriptivas, con una estructura narrativa relativamente simple.

Para ello, se han analizado 369 piezas, obtenidas de la página *web* del diario El País, a lo largo del año 2021, y se ha llevado a cabo un proceso de tabulación a partir de los textos obtenidos. Esta tabulación se corresponde con una tabulación sintáctica que busca relacionar cada una de las W's con su huella sintáctica, y si bien se encuentra aún en fase de desarrollo, podemos adelantar que las únicas Ws que los etiquetadores sintácticos no llegan a cubrir con garantías se corresponden con los *A quién* —si bien la inclusión de la preposición en la ecuación tiende a solucionar en gran medida este problema, cuando precede a un complemento directo o indirecto—, y el *Cuándo* —sobre todo si se emplean etiquetadores multilingüaje, puesto que carecen de índices lo suficientemente exhaustivos cuando se enfrentan a sustantivos propios de lugar poco conocidos—.

Referencias bibliográficas

- Agerri, R.; Artola, X.; Beloki, Z.; Rigau, G.; Soroa, A. (2015). Big data for Natural Language Processing: A streaming approach. *Knowledge-Based Systems*, 79, 36-42. <http://dx.doi.org/10.1016/j.knsys.2014.11.007>
- Alvite, M.L. (2012). El uso de vocabularios controlados en los sistemas de información jurídica: evolución y tendencias actuales de representación. *Scire. Representación y Organización del Conocimiento*, 18(1): 29-39. <https://www.iberid.eu/ojs/index.php/scire/article/view/3973>
- Arakaki, F.; Coneglian, C.; Santos, P.; Santarém, J. (2019). Disponibilidad de metadatos en linked data para repositorios digitales. *Anales de Documentación*, 22(2). <http://dx.doi.org/10.6018/analesdoc.32744>
- Barba, C.; García-Nieto, J.; Navas-Delgado, I.; Aldana-Montes, J.F. (2016). A fine grain sentiment analysis with semantics in Tweets. *International Journal of Interactive Multimedia and Artificial Intelligence*, 3(6): 22-28. <http://dx.doi.org/10.9781/ijimai.2016.363>
- Baviera, T. (2017). Técnicas para el análisis de sentimiento en Twitter: Aprendizaje automático supervisado y SentiStrength. *Dígitos*, 3: 33-50. <http://dx.doi.org/10.7203/rd.v1i3.74>
- Beltrán, P. (2015). La documentación informativa y la recuperación de la información escrita. Nuevas competencias para el ciberespacio. *Index.comunicación*, 5(3): 29-57. <http://hdl.handle.net/10115/15458>
- Berners-Lee, T. (2000). *Tejiendo la red*. Siglo XXI.
- Berners-Lee, T.; Hendler J., y Lassila, O. (2001). The semantic web. *Scientific American*, mayo 2001, 29-37. https://www.sop.inria.fr/acacia/cours/essi2006/scientific%20american_%20feature%20article_%20the%20semantic%20web_%20may%202001.pdf
- Bueno, M.R. (2000). Estructura textual, macroestructura semántica y superestructura formal de la noticia. *Estudios sobre el Mensaje Periodístico*, 6: 239-258. <https://revistas.ucm.es/index.php/ESMP/article/download/ESMP0000110239A/0/0>
- Casas, J.; Nin, J; Julbe, F. (2019). Big Data: Análisis de datos en entornos masivos. UOC.
- Casasús, J.M.; Núñez Ladevéze, L. (1991) *Estilo y géneros periodísticos*. Ariel.
- Codina, L. (2016). Tres dimensiones del periodismo computacional. Intersecciones con las ciencias de la documentación. *Anuario ThinkEPI*, 10: 200-202. <http://dx.doi.org/10.3145/thinkepi.2016.41>
- Colle, R. (2013). Prensa y 'Big Data': El desafío de la acumulación y análisis de datos. *Revista Mediterránea de Comunicación*, 4 (1): 1-8. <https://doi.org/10.14198/MEDCOM2013.4.1.13>
- Conroy, N.; Rubin, V.; Chen, Y. (2016). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1): 1-4. <https://doi.org/10.1002/pa2.2015.145052010082>

- Córdoba-Cabús, A; Hidalgo-Arjona, M; López-Martín, Á. (2021). Cobertura de diarios en Twitter de las elecciones autonómicas de Madrid 2021. Procesamiento de lenguaje natural y algoritmos de aprendizaje automático. *Profesional de la Información*, 30 (6). <https://doi.org/10.3145/epi.2021.nov.11>
- Criado, J.I.; Villodre, J. (2018). Comunicando datos masivos del sector público local en redes sociales. Análisis de sentimiento en Twitter. *El Profesional de la Información*, 27(3): 614-623. <https://doi.org/10.3145/epi.2018.may.14>
- Dovifat, E. (1960). *Periodismo. Vols. 1 y 2*. Uthera.
- Drndarevic, B.; Saggion, H. (2012). Reducing text complexity through automatic lexical simplification: an empirical study for spanish. *Procesamiento del Lenguaje Natural*, 49: 13-20. <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4555>
- Fermoso-García, A.; Manzano-García, M.I.; Mateos-Sánchez, M.; Hernández, C. (2018). Sistema de modelado semántico para la catalogación, clasificación, consulta y publicación en abierto de información bibliográfica. *El Profesional de la Información*, 27(2): 410-418. <https://doi.org/10.3145/epi.2018.mar.20>
- Fontuberta, M. de (2003) La noticia. Pistas para percibir el mundo. Paidós.
- Freixa, P; Pérez-Montoro, M.; Codina. L. (2017). Interacción y divulgación de datos en el periodismo estructurado. *El Profesional de la Información*, 26(6): 1076-1090. <https://doi.org/10.3145/epi.2017.nov.07>
- García, A. (2004). Instrumentos de representación del conocimiento: tesauros versus ontologías. *Anales de Documentación*, 7: 79-95. <https://revistas.um.es/analesdoc/article/view/1691>
- García, A. (2014). Análisis documental de noticias de prensa en sistemas de información factual. *Revista Española de Documentación Científica*, 37(2):e046. <http://dx.doi.org/10.3989/redc.2014.2.1094>
- García, A. (2016). Organización del conocimiento para la documentación en periodismo. Situación y prospectiva. *Scire. Representación y Organización del Conocimiento*, 22(2): 21-28. <https://dialnet.unirioja.es/servlet/articulo?codigo=5713295>
- García, A.; Rodríguez, D.; Catalina, B. (2019). Estudio sobre la indización/etiquetado y los lenguajes documentales en cinco diarios españoles. *Scire. Representación y Organización del Conocimiento*, 25(1): 55-64. <https://ojs.iberid.eu/index.php/scire/article/view/4579>
- Gomis, S. (2015). Publicar datos abiertos enlazados (LOD) en entornos bibliotecarios mediante el gestor de contenidos semántico Ximdex. *Cuadernos de Gestión de Información*, 5(1): 1-17. <https://revistas.um.es/gesinfo/article/view/232181>
- González-Arias, C.; López-García, X. (2023). ChatGPT: Stream of opinion in five newspapers in the first 100 days since its launch. *Profesional De La información*, 32(5). <https://doi.org/10.3145/epi.2023.sep.24>
- Grijelmo, A. (1997). *El estilo del periodista*. Taurus.
- Gutiérrez-Caneda, B; Vázquez-Herrero, J.; López-García, X. (2023). AI application in journalism: ChatGPT and the uses and risks of an emergent technology. *Profesional De*

- La información, 32(5). <https://doi.org/10.3145/epi.2023.sep.14>
- Hernando, A. (2015). El discurso periodístico en el reportaje de prensa. *Signa: Revista de la Asociación Española de Semiótica*, 24: 375-392
<http://www.cervantesvirtual.com/descargapdf/el-discurso-periodistico-en-el-reportaje-de-prensa/>
- Lloret, E.; Palomar, M. (2011). COMPENDIUM: Una herramienta de generación de resúmenes modular. *Procesamiento de Lenguaje Natural*, 47:107-115.
<http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/955>
- Marcos, J.C.; Edo, C. (2015). Análisis de la nueva perspectiva de la documentación periodística en los medios de comunicación españoles. *Revista General de Información y Documentación*, 25(2): 389-423.
http://dx.doi.org/10.5209/rev_RGID.2015.v25.n2.51235
- Martín, G. (1987). *Géneros periodísticos*. Paraninfo.
- Martínez Albertos, J.L. (1991). *Curso general de redacción periodística*. Paraninfo.
- Martínez, M.M.; Alvite, M.L. (2014). Propuesta metodológica de evaluación de gestores de tesauros compatibles con la web semántica. *Anales de Documentación*, 17(1): 1-18.
<https://doi.org/10.6018/analesdoc.17.1.186271>
- Mas, L. (2016a). Estructura formal, textual y oral del discurso público. *Estudios sobre el Mensaje Periodístico*, 22(1): 445-461. https://doi.org/10.5209/rev_esmp.2016.v22.n1.52606
- Mas, L. (2016b) Caracterización prosódica del inicio, desarrollo y final de la noticia. *Revista Española de Lingüística Aplicada*, 29(1): 212-237. DOI:
<https://doi.org/10.1075/resla.29.1.09mas>
- Mateo, P.L.; González, J.C.; Villena, J.; Martínez, J.L. (2003). Un sistema para resumen automático de textos en castellano. *Procesamiento del Lenguaje Natural*, 31: 29-36.
<http://hdl.handle.net/10045/1486>
- Mungra, P.; Calziani, T. (2013). Lexicographic studies in medicine: Academic World List for clinical case histories. *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE)*, 25: 39-62.
<https://dialnet.unirioja.es/servlet/articulo?codigo=4276659>
- Parratt, S. (2017). Manual práctico de redacción periodística. Síntesis.
- Passos, A.; Kumar, V.; McCallum, A. (2014). Lexicon Infused Phrase Embeddings for Named Entity Resolution. *Association for Computational Linguistics*: 78-86.
<https://doi.org/10.3115/v1/W14-1609>
- Pedraza, R.; Codina, L.; Rovira, C. (2007). Web semántica y ontologías en el procesamiento de la información documental. *El Profesional de la Información*, 16(6): 569-579.
<https://doi.org/10.3145/epi.2007.nov.04>
- Pérez, A.; Casillas, A.; Gojenola, K.; Oroz, M.; Aguirre, N.; Amillano, E. (2014). The aid of machine learning to overcome the classification of real health discharge reports written in Spanish. *Procesamiento del Lenguaje Natural*, 53: 77-84.

<http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5048>

- Pérez, M.A.; Cardoso, A.C. (2014). Técnicas de extracción de entidades con nombre. *Inteligencia artificial*, 17(53): 3-12. <https://www.redalyc.org/pdf/925/92530455002.pdf>
- Pérez-Rosas, V.; Kleinberg, B.; Mihalcena, R. (2018). Automatic Detection of Fake News. *Proceedings of the 27th International Conference on Computational Linguistics*, 3391-3401. <https://www.aclweb.org/anthology/C18-1287.pdf>
- Plaza, L.; Díaz, A. (2011). Using semantic graphs and word sense disambiguation techniques to improve text summarization. *Procesamiento del Lenguaje Natural*, 47: 97-105. <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/956>
- Quesada, M. (2007). *Periodismo de sucesos*. Síntesis.
- Ratinov, L.; Roth, D. (2009). Design Challenges and Misconceptions in Named Entity Recognition. Association for Computational Linguistics. Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009): 147-155. <https://www.aclweb.org/anthology/W09-1119.pdf>
- Rodríguez, L. (2019). Desinformación: retos profesionales para el sector de la comunicación. *El Profesional de la Información*, 28(3). <https://doi.org/10.3145/epi.2019.may.06>
- Rodríguez, T. (2013). Perspectivas en el análisis de la relación entre gramática y discurso: estructura periférica y conexiones entre oraciones. *Círculo de Lingüística Aplicada a la Comunicación*, 55, 3. https://doi.org/10.5209/rev_clac.2013.v55.43263
- Sánchez, R.; Gil, B. (2007). Lenguajes documentales y ontologías. *El Profesional de la Información*, 16(6): 551-560. <https://doi.org/10.3145/epi.2007.nov.02>
- Sancho, A.; Fernández, C.; Boulat, P. (2012). La búsqueda de información jurídica: de los tesauros a la inteligencia artificial. *Scire: Representación y Organización del Conocimiento*, 18(1): 73-83. <https://www.iberid.eu/ojs/index.php/scire/article/view/3942>
- Szostak, R. (2014). Advances in Classification Research Online 2013. Classification, Ontology, and the Semantic Web. *Advances In Classification Research Online*, 24(1): 30-37. <http://dx.doi.org/10.7152/acro.v24i1.14674>
- Túñez-López, M.; Toural-Bran, C.; Cacheiro-Requiejio, S. (2018). Uso de bots y algoritmos para automatizar la redacción de noticias: percepción y actitudes de los periodistas en España. *El Profesional de la Información*, 27(4): 750-758. <https://doi.org/10.3145/epi.2018.jul.04>
- Túñez-López, M.; Toural-Bran, C.; Valdiviezo-Abad, C. (2019). Automatización, bots y algoritmos en la redacción de noticias. Impacto y calidad del periodismo artificial. *Revista Latina de Comunicación Social*, 74(12): 1411-1433. <http://dx.doi.org/10.4185/RLCS-2018-1391>
- Ufarte, M.J.; Peralta, L.; Murcia-Verdú, F.J. (2018). Fact checking: un nuevo desafío del periodismo. *El Profesional de la Información*, 27(4): 733-741
- Van Dijk, T. (1980). *Texto y contexto*. Cátedra.
- Van Dijk, T. (1990). La noticia como discurso. Comprensión, estructura y producción de la

información. Paidós.

Vera, A. (2012). El párrafo como unidad discursiva. *Estudios de lingüística*, 26: 343-358.
<http://hdl.handle.net/10045/28733>

Warren, C. (1975). Géneros periodísticos informativos. ATE.