

## CAPITULO

## 4

CORPUS

## Capítulo 4º

### CORPUS

#### 4.1 CARACTERÍSTICAS DEL UNIVERSO DE REFERENCIA

El universo de referencia lo constituyen 747 registros informáticos provenientes del vaciado de periódicos y revistas españoles correspondientes al año 1992. Este universo ha sido extraído de la base de datos MEDIACOM del Centro de Documentación para la investigación en Comunicación y Culturas -HISPACOM- adscrito al Departamento de Comunicación de la Universidad de Sevilla, miembro de la Red latinoamericana de centros de documentación en Comunicación COMNET-AL (1).

Para la creación de MEDIACOM, se propuso como proyecto piloto el análisis documental (esto es, extracción de datos hemerográficos, resumen y descriptores) de varios medios impresos españoles, de acuerdo a sus tiradas respectivas en el ámbito nacional y autonómico andaluz, sumados a las revistas de información general, también de mayor tirada en 1992 según la OJD. Los medios seleccionados fueron los siguientes:

<b>nacional</b>	<b>autonómico</b>	<b>revistas</b>
El País	D16 Andalucía	Cambio 16
ABC	Correo de Andalucía	Tiempo
El Mundo	Diario Sur. Málaga	Epoca

Un equipo de once personas con conocimientos periodísticos y documentales procedió, bajo los mismos criterios formales, a la selección de textos en función de la temática solicitada y de las restricciones impuestas. Como regla general debía seleccionarse todo el material publicado que cubriera los siguientes ámbitos: libro, prensa, radio, televisión, video, cine, es decir, cualesquiera modos y medios de comunicación social, excluyendo la información sobre artes y espectáculos, reseñas y críticas concretas, información de menos de cuarenta líneas o dos párrafos a menos que presentase suficiencia

(estadística, necrológica, declaración...), futuros, noticias de tono electoralista sobre los media y redundancias.

El número potencial de ejemplares para leer y analizar era de unas 2350 unidades, muchas de las cuales no contenían datos relevantes. Por tanto, la dispersión media de registros por fuente está en torno a un registro por cada tres unidades leídas, descontada la redundancia de la misma noticia en varias fuentes, sobre todo, de la información proveniente de agencias.

Este resultado demuestra el alto nivel de selección practicado. De hecho, la base de datos MEDIACOM está diseñada para dar servicio no sólo a periodistas y estudiantes de Periodismo sino también a investigadores de la Comunicación que deseen basar sus reflexiones en fuentes hemerográficas. Debo señalar, en ese sentido, la importancia dada recientemente por el equipo Cultura de la Universidad de Colima, dirigido por el profesor mexicano Jorge González, a las diferentes posiciones de los medios en relación al fenómeno sociocultural de las telenovelas, objeto que fue tema de un curso de doctorado de la Universidad Complutense en 1991-1992 (2).

Por otro lado, se observa un cierto tipo de noticias que "frivoliza" el corpus: estrenos de programas de TV, disputas entre gestores de empresas informativas, guerra de cadenas de TV, etc. Como política selectiva de la base de datos hubo de ser sacrificada la precisión en beneficio de registros de obligada presencia.

En el universo global que tomamos para esta investigación, sobre todo con vistas a su crecimiento controlado y reglado, se observan ciertas variables:

- fuentes: los registros proceden de nueve fuentes distintas.
- género y estilo: los registros se dispersan entre información y opinión: editoriales, entrevistas, reportajes, crónicas, etc.
- autores: personales, institucionales y textos anónimos.
- extensión: la base de datos unifica la extensión de los originales mediante resúmenes.

Ninguna de estas variables afecta a la temática de un modo lo suficientemente notorio como para tenerlas en cuenta a la hora de crear la muestra, puesto que el universo original de referencia ha sido formateado en registros uniformes.

#### 4.2 ELEMENTOS DEL CORPUS: LOS REGISTROS

Registro es la unidad de información en una base de datos convencional (como resulta ser MEDIACOM). Un registro se compone de campos y, en el campo, se transcriben datos independientes: fecha, autor, fuente, resumen, etc. En MEDIACOM se han creado hojas de trabajo, a modo de cuestionario, con campos en los que el analista debe incluir los datos solicitados a la hora de leer la noticia. Ya que muchos de los campos en una hoja de trabajo no son visionados ni imprimidos como respuesta a una consulta, es decir, no pertenecen al usuario sino al equipo de trabajo (controles de análisis, corrección, digitación, etc.) sólo interesan, a efectos de esta investigación, aquéllos que aportan información sobre el texto analizado y no en la forma que adoptan en la hoja de trabajo sino en función del bloque al que se adscriben en el formato de salida:

**1º bloque.** Encabezamiento: contiene el nombre del autor principal (invirtiendo apellidos, nombre) o la primera palabra del título, siempre que no sea artículo, en caso de autor anónimo.

**2º bloque.** Área del título: contiene el título principal seguido de subtítulos. Separado por /, se encuentra la mención de responsabilidad (autores, caso de haberlos) y, tras un punto, el nombre del medio y fecha de publicación entre paréntesis. Finalmente, la colación viene representada por la letra "p" (página) seguida del número. (norma ISBD adaptada)

**3º bloque.** Resumen: texto libre que aclara y/o amplía el sentido del título. Ayuda al usuario a decidir la necesidad o interés de consultar el original y sitúa el texto en un relato más genérico.

**4º bloque.** Descriptores: el contenido del original es desmontado en sustantivos y sintagmas nominales a efectos de recuperar registros mediante múltiples combinaciones (ángulos de búsqueda).

Puede deducirse, tras una rápida observación del corpus, que la información proporcionada por el periodista en el título puede ser insuficiente para el investigador y que los datos aportados por los documentalistas en resúmenes y descriptores son parciales y presentan desequilibrios de un registro a otro. Por ello, parece lógico optar por la vía de formalizar los datos de los registros en otras unidades que garanticen su manipulación científica.

En definitiva, nos interesan los temas y su tratamiento no en calidad de temas sino de enciclopedismo o dispersión temática dentro de un corpus de noticias que guardan cierta afinidad, del mismo modo que nos interesará, más adelante, el comportamiento funcional del concepto dentro del texto, por encima de su significación real. Para ese objetivo, será necesario usar un lenguaje artificial que nos permita una mayor operatividad.

##### 4.2.1 Tipología textual

Antes de someter los títulos a modificación por el procedimiento de enriquecimiento, es necesario reconocer los tipos textuales que manejamos a fin de ampararnos, nuevamente, en la explicitación parcial como garantía de los resultados finales. En este caso, tomamos del campo del título, el encabezamiento principal. Haciendo un sondeo aleatorio, sobre el universo de referencia, obtenemos los siguientes tipos:

- Desde la perspectiva de la construcción:
  - oraciones atributivas: A es B.
  - oraciones impersonales: se hace...
  - oraciones intransitivas: X muere.
  - oraciones subordinadas introducidas por "que": A dice que B...
  - oraciones transitivas con interrupción verbal: X hace.
  - oraciones simples en cualquier orden (hipérbaton).

oraciones con verbo tácito: X, en TV.

oraciones en pasiva: X es acusado por Y.

oraciones coordinadas: cambia X y renueva Y.

oraciones yuxtapuestas: prensa verdadera, prensa falsa.

encabezamiento por participio: condenado por...

sintagmas nominales: periodismo inquietante nombre de personaje seguido de frase lapidaria.

- Desde la perspectiva semántica:
  - parte por el todo: X dice Y en inauguración.
  - metáfora: traje nuevo para Time.
  - título de obra: sensación de vivir.
  - nombre de personaje-noticia: Erice.
  - neologismos: zapateadores.
  - juegos de palabras: el silencio de los oscars.
  - vaguedad: erase una vez...
  - frase de personaje: ha sido un año terrible.
  - juicio: freudiano

Obsérvese en estos ejemplos de una muestra inacabable, puesto que su matriz es retórica y publicitaria, la dispersión tipológica que debemos afrontar. La vía de la modelización, en este aspecto, sería tan innecesaria como imposible de ejecutar. En consecuencia, el investigador (y en su día también el documentalista) debe proceder a la "retitulación" para su análisis de contenido (obviamente, el título original permanece en la descripción formal del artículo). En esa retitulación, o enriquecimiento del título, intervienen elementos de transformación semántica y morfosintáctica.

Para evitar posibles escauceos, también del documentalista en la retórica, el sistema debe ofrecerle patrones-corsé que dirijan la retitulación. Tales patrones deben ser lo suficientemente genéricos como para albergar cualquier retitulación y lo suficientemente precisos como para reducir el inventario a dimensiones manejables.

El enriquecimiento es una formalización del enunciado a partir de la comprensión del mensaje global del texto. En consecuencia, debemos estudiar cómo acceder a ese mensaje y propo-

ner reglas comunes para todo mensaje del corpus de referencia. Se trata, en definitiva, de obtener un enunciado global y descriptivo del contenido, es decir y en términos de Van Dijk, una macroproposición. Solamente a partir de formalizaciones de las noticias podemos concebir un proceso de datos periodísticos en Inteligencia artificial.

#### 4.3 CONSTITUCION DE LA MUESTRA: GARANTIAS PREVIAS Y METODO DE VALIDACION

La suma de los datos procedentes de los titulares, resúmenes y descriptores, deberá componer la unidad mínima de operación. Pero sobre su estructura y construcción se hablará más adelante. Ahora, en una aproximación realizada de lo general a lo particular, cumple describir el armazón en el que se articulan tales unidades o muestra, propiamente dicha.

Efectivamente, en apartados anteriores nos hemos detenido en el universo global de referencia. Tomar este corpus exhaustivo como espacio de investigación sería poco operativo e innecesario. Esta afirmación la basamos en el principio sostenido por Krippendorff y aplicado al Análisis de Contenido, cuando nos dice que para obtener una muestra que evite el análisis de la población total sólo se requiere un test que nos proporcione un adecuado tamaño de la muestra (3).

De acuerdo con el profesor de la Universidad de Pennsylvania, "la técnica de la mitad" ofrece un muestreo satisfactorio. La muestra inicial es dividida en dos partes de igual tamaño. Si cada parte aporta el mismo nivel confianza, la totalidad puede ser aceptada como adecuada en sus dimensiones. De lo contrario, el analista debe incrementar el tamaño hasta que las condiciones deseadas se cumplan.

Para la constitución de la muestra hemos optado por la técnica aleatoria, dado que el tipo de variables (fuente, género, autor y extensión) existente no es relevante para la obtención del objetivo (puesto que trabajamos con textos formalizados), lo que excluye la estratificación o reconocimiento de subgrupos (4).

Así, el método aleatorio garantiza, para muestras simples, que cualquier registro del universo total tenga igualdad de oportunidades de estar incluido o excluido de la población de datos seleccionada. Krippendorf apela al sorteo para determinar qué unidades compondrán la muestra, a través de cualquier procedimiento que asigne igualdad de probabilidades.

Puesto que la división del universo de referencia agotaría la posibilidad de ampliación según la técnica de la mitad, es necesario proceder a la selección de una muestra pequeña que sea divisible en dos partes. El tamaño inicial para la verificación no es relevante y, sin embargo, ha de ser operativo. Por lo tanto, partimos del 50% del universo total para constituir la muestra, siendo sus mitades, el 25% de ese mismo universo. No obstante, la selección de las unidades de trabajo no puede ser lineal ya que la información no fue introducida totalmente al azar en la base de datos (los analistas digitaron registros del mismo medio, subtema o periodo en una misma sesión).

En suma, opto por adjudicar, a la primera mitad (A), el primero de cada cuatro registros de la base de datos y, a la segunda mitad (B), el segundo registro del mismo tramo. Esto hace, aproximadamente, 186 referencias por submuestra, evitando decimales, es decir, contamos con una muestra global de 372 registros en dos mitades ampliables al doble de su volumen. Con ello, confiamos en la fiabilidad de los resultados al verificar lo obtenido en A sobre B.

#### 4.4 UNIDADES OPERATIVAS

Mientras que la gramática convencional se centra en la detección de las categorías funcionales (sujetos y complementos) habidas en las oraciones para desmontar los distintos roles de las palabras en las frases, el método que usamos busca la identificación de roles conceptuales más allá de la frase, a saber, en la estructura semántica. Puesto que ésta se halla camuflada bajo estrategias y recursos discursivos, necesitamos trabajar sobre unidades desprovistas de figuras retóricas.

De esta forma, los resultados de la metodología de análisis serían no sólo independientes de los agentes que los efectuaron (5) sino también de cualesquiera formas de expresión que adoptara el mensaje. Por tanto, una vía hipotéticamente válida pero poco operativa sería la de enunciar formalmente una proposición en todas las variantes posibles y, así, ratificar la presencia de unos pocos roles en la dispersión de reiteraciones.

##### 4.4.1 Estructuras formales y estructuras semánticas

Dado que los resultados del análisis sintáctico convencional varían según el enunciado y la posición de los conceptos, es inviable que una simulación de la lectura pueda ser realizada con suficientes garantías y satisfacción sin la intervención humana. Por ello, apostamos por una simulación de la lectura de macroproposiciones, o textos enriquecidos por el experto en la gestión automática, lo que quiere decir que debemos olvidar los milagros de la tecnología inteligente y replantear las tareas del documentalista de prensa del próximo futuro y su intervención o exclusión en determinadas fases del proceso de la noticia.

El observador se topa, no obstante, con estructuras de superficie que aportan datos en diferentes posiciones físicas dentro del registro y debe servirse de ellas para sus propósitos. A través de estas estructuras formales puede llegarse a la estructura de significación de la noticia la cual ejerce, además, un rol conector con otras unidades que conforman el relato en el discurso periodístico.

Tengamos en cuenta que a mayor cantidad de enunciados, mayor aportación de datos irrelevantes y secundarios y, por lo tanto, mayor dificultad para documentalistas de prensa e investigadores de la Documentación periodística en lo que se refiere a resultados y eficacia. Así, es necesario desprender de lo relevante el ropaje añadido por el productor del texto, es decir, aislar las proposiciones más importantes del autor para permitir la realización de la práctica y de la investigación. En ese sentido, es preciso

manipular exclusivamente enunciados breves y altamente significativos, presentes en el texto o, en su defecto, construirlos en laboratorio.

No cabe duda de que, en la redacción de titulares, el redactor concentra, en pocas palabras, la mayor cantidad de información posible (principio de economía). Este recurso sintético ofrece, en algunas ocasiones, enunciados explicativos y descriptivos que facilitan la tarea a los documentalistas pero, en otras, hace incomprensible o provoca confusión en cuanto al sentido del texto: el título se confecciona en función de los objetivos estratégicos de la producción de la noticia: sensacionalistas o ideológicos.

En consecuencia, no puede procederse a investigar este aspecto de la Documentación periodística si no es a través de artificios obtenidos desde las estructuras formales en función de las estructuras semánticas. A medida que se obtengan resultados, dentro de esta tendencia investigadora, será posible relajar el control del documentalista en el reconocimiento de textos y alcanzar mayores cotas de fiabilidad en el reconocimiento artificial. Pero, por el momento, el ser humano debe participar en todas las fases de captación de síntesis.

#### **4.4.2 Construcción de unidades de trabajo**

Los registros manejados no constituyen elementos válidos para operar. Basta observar un listado de titulares para entender, incluso bajo el mismo eje temático, esta afirmación. Figuras estilísticas (Hispatat pone en órbita el castellano), omisiones (Grupo 16 negocia) y sensacionalismo (!Que viene el lobo!), entre otras muchas causas, impiden tomar directa y exclusivamente el título como fuente de información para el análisis documental de contenido. Por ello, es necesario construir nuevas unidades que garanticen una mejor representación del texto entregando suficiente información al investigador y, en su caso, a la máquina. Denominaremos, a esas construcciones, unidades de trabajo (UT).

El número de UT en la muestra coincide con el número de registros en el corpus real a pesar

de existir en éste múltiples enunciados. Esto se debe a que, en ciertos registros, unas noticias inflexionan o entran en dependencia (Cela dice algo en un Congreso inaugurado por el rey) teniendo el mismo nivel de relevancia (palabras de Cela, inauguración) y, en consecuencia, cada enunciado relevante debiera ser analizado como si fuesen textos distintos en el trabajo real. En esta investigación hemos optado por subsumir la noticia bajo un sólo enunciado o elegir el enunciado más genérico, en caso de diversidad.

##### **4.4.2.1 Metodología de construcción**

La construcción de UT es una técnica de investigación documental y, por tanto, útil para la experimentación pero carente de uso fuera de esas condiciones ya que los receptores y peticionarios de la información documental de actualidad se dirigirán en texto libre hacia textos no condicionados. La UT se obtiene mediante procedimientos derivados del denominado enriquecimiento (expanded title) (6). Se procede a una reducción cuantitativa muy notoria respecto al tamaño del registro. Se pretende trabajar con unidades mínimas o proposiciones (7), esto es, conjuntos de conceptos en torno a una sólo acción. A veces el título selecciona una frase brillante del texto, que no tiene por qué coincidir con la proposición más importante del mismo. Por tanto, el título es una guía indicativa, en principio, no vinculante.

La construcción, por reducción, en laboratorio es necesaria para modelizar enunciados que, de otra forma, presentan gran dispersión terminológica, semántica y sintáctica. Es evidente que para poder proponer soluciones generales a problemas dispersos, es precisa una fácil manipulación de los datos. En este sentido, prefiero observar el texto como un conjunto de proposiciones separables con informaciones parciales. La suma de todas ellas es la macroproposición. La macroproposición es representable mediante un enunciado simple que gire en torno a una acción. No obstante, pueden aparecer casos en los que son necesarios más enunciados. Tal situación excepcional sería evitada, sin perjuicio para los resultados, puesto que los enunciados siguientes serían nuevas unidades de trabajo en



un supuesto práctico, si bien procedentes de un mismo registro.

Manipular enunciados simples facilita enormemente la formalización y, en consecuencia, la simulación mecánica, objetivo a medio plazo que persigue esta investigación. Podríamos haber obtenido estos enunciados por vía intuitiva, puesto que en las transformaciones sucesivas del corpus se hubiera ajustado convenientemente a las formas deseadas. Sin embargo, he preferido explicitar los pasos dados desde los registros hasta la elaboración de la unidad de trabajo para garantizar la fiabilidad ulterior.

Ya que, en nuestro caso, contamos con un universo previo bastante uniforme, el método de Van Dijk fue aplicado con relativa facilidad a fin de justificar un mecanismo de explicitación de la reducción. Estas son sus reglas generales:

- Eliminación de proposiciones a través de las siguientes operaciones selectivas:  
eliminación de predicados atributivos.  
eliminación de datos secundarios y profundos.  
localización de hechos (lugar y tiempo) secundarios.  
razones y consecuencias de los hechos secundarios.  
hechos preliminares y auxiliares.
- Sustitución de proposiciones específicas por proposiciones supraordenadas a través de operaciones constructivas:  
sustitución por generalización simple.  
sustitución por construcción propiamente dicha.
- Elaboración de la macroproposición.

La adaptación realizada tiene en cuenta la extensión de nuestros registros. El hecho de ser textos de documentalistas ya supone una reducción de datos secundarios importante, por lo que se diseñó un juego de reglas de simple aplicación, seguidas ordenadamente:

- a) lectura lineal y neurálgica del registro.
- b) eliminación de subordinaciones, coordinaciones y yuxtaposiciones.

c) eliminación de material irrelevante y secundario: redundancias, antecedentes, expectativas y datos concretos y numéricos.

d) sustitución por sinónimo o expresión más común, en caso de término poco usual y metáforas.

e) sustitución por generalización.

f) construcción en voz activa, en impersonal o tercera persona.

g) explicitación de nexos y unificación en caso de sinonimia.

h) unificación de acciones mediante la selección de formas verbales genéricas y aplicación retrospectiva de las mismas.

Como resultado de esta lectura dirigida (por unos objetivos) y de acuerdo con lo demostrado por Amaro, se obtiene una macroproposición o resumen general a partir de todas las propuestas menores de los textos. El enunciado obtenido nos permite comenzar a operar formalizadamente. En suma, esta metodología optimiza la carga significativa de las UT a efectos de esta investigación y, además, la proponemos como modo simplificado y utilizable por el documentalista de prensa a fin de dotarle de un instrumento de primera objetivación de las reducciones antes de la formalización que han de sufrir los textos al pasarlos a formatos legibles por ordenador.

### *Notas bibliográficas*

(1) Por aceptación de la candidatura en la reunión de la Red latinoamericana de Centros de documentación en Comunicación celebrada en Bogotá los días 19-20 de Enero de 1994, según consta en el acta final de la Asamblea de COMNET-AL.

(2) Organizado por el Departamento de Sociología IV de la UCM.

(3) Krippendorf, Klaus: op. cit. p.69.

(4) Ibid., p.66-67.

(5) Condición impuesta como objetivo de investigación en García Gutiérrez, A.: **Análisis documental...** op. cit., p.128.

(6) Idem: **Linguística documental...** op. cit., p.111.

(7) Idem: **Análisis...** op. cit., p.87.