

# EL ANÁLISIS DEL DISCURSO EN LA INVESTIGACIÓN SOCIAL: TEORÍAS Y PRÁCTICAS

Araceli Serrano Pascual  
Ángel Zurdo Alaguero



Contenidos digitales  
[www.sintesis.com](http://www.sintesis.com)

# El análisis del discurso en la investigación social: teorías y prácticas

COLECCIÓN:  
BIBLIOTECA DE SOCIOLOGÍA (MANUAL)

Coordinador:  
*Cristóbal Torres Albero*



Queda prohibida, salvo excepción prevista en la ley, cualquier forma de reproducción, distribución, comunicación pública y transformación de esta obra sin contar con autorización de los titulares de la propiedad intelectual. La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (arts. 270 y sigs. Código Penal). El Centro Español de Derechos Reprográficos ([www.cedro.org](http://www.cedro.org)) vela por el respeto de los citados derechos.

# El análisis del discurso en la investigación social: teorías y prácticas

Araceli Serrano Pascual

Ángel Zurdo Alaguero



Consulte nuestra página web: **www.sintesis.com**  
En ella encontrará el catálogo completo y comentado.

Reservados todos los derechos. Está prohibido, bajo las sanciones penales y el resarcimiento civil previstos en las leyes, reproducir, registrar o transmitir esta publicación, íntegra o parcialmente, por cualquier sistema de recuperación y por cualquier medio, sea mecánico, electrónico, magnético, electroóptico, por fotocopia o por cualquier otro, sin la autorización previa por escrito de Editorial Síntesis, S. A.

Imagen de cubierta original de Noelia Majo del Río

© Araceli Serrano Pascual

© Ángel Zurdo Alaguero

© EDITORIAL SÍNTESIS, S. A.

Vallehermoso, 34. 28015 Madrid

Teléfono 91 593 20 98

[www.sintesis.com](http://www.sintesis.com)

ISBN: 978-84-1357-303-8

Depósito Legal: M. 26.285-2023

Impreso en España - Printed in Spain

# Índice

<b><i>Prefacio</i></b> .....	9
Cómo usar este libro .....	12
<b><i>Introducción</i></b> .....	13
<b>1. Consideraciones sobre el discurso</b> .....	19
1.1. Las nociones de discurso .....	19
1.1.1. <i>El origen etimológico del concepto de discurso</i> .....	20
1.1.2. <i>El discurso como producto o como práctica</i> .....	22
1.1.3. <i>El discurso como práctica lingüística o como práctica social</i> .....	23
1.1.4. <i>La paradójica dualidad materialidad-agencia</i> .....	26
1.1.5. <i>El carácter construido de los discursos: ¿construcción analítica?</i> .....	27
1.1.6. <i>Nociones hipertrofiadas del discurso</i> .....	29
1.1.7. <i>Ensayando una definición de discurso</i> .....	31
1.2. Rasgos constitutivos de los discursos sociales .....	33
1.2.1. <i>La naturaleza ideológica del discurso</i> .....	33
1.2.2. <i>El carácter dialógico de los discursos</i> .....	33
1.2.3. <i>El carácter conflictivo y relacional de los discursos</i> .....	34
1.2.4. <i>La existencia de hegemonía discursiva</i> .....	35
1.2.5. <i>La necesaria interpretación de los discursos</i> .....	36
1.2.6. <i>La multimodalidad de los materiales empíricos</i> .....	37
1.3. Aproximación a una tipología de materiales empíricos .....	38
1.4. Los niveles del análisis .....	40
1.4.1. <i>Nivel informacional</i> .....	41

1.4.2. Nivel estructural .....	42
1.4.3. Nivel sociohermenéutico .....	43
1.4.4. Niveles, tradiciones y perspectivas de análisis .....	44
<b>2. El análisis de contenido de textos: enfocando presencias y frecuencias .....</b>	<b>49</b>
2.1. Hacia la cuantificación como objetivo privilegiado .....	49
2.1.1. El análisis clásico de contenido .....	52
2.1.2. Transformaciones recientes del análisis de contenido cuantitativo .....	56
2.1.3. La lingüística del corpus .....	59
2.2. El análisis temático .....	61
2.3. Limitaciones y debates en torno al análisis de contenido .....	63
Caso práctico .....	68
Recurso electrónico 2.1 .....	69
<b>3. El método de la teoría fundamentada en los datos: inducción, codificación y emergencia de la teoría .....</b>	<b>71</b>
3.1. Antecedentes y supuestos básicos del MTFD .....	73
3.1.1. El origen del MTFD .....	73
3.1.2. Caracterización sintética del método de la teoría fundamentada en los datos .....	75
3.2. Elementos y procesos analíticos básicos del MTFD .....	78
3.2.1. Los datos .....	78
3.2.2. Inductivismo .....	81
3.2.3. Codificación .....	83
3.2.4. Método comparativo constante .....	95
3.2.5. La intercambiabilidad de los índices .....	96
3.2.6. Muestreo teórico y saturación teórica .....	97
3.2.7. La elaboración de memorandos (memos) o notas de análisis .....	101
3.2.8. La emergencia o descubrimiento de la teoría .....	103
3.3. Las variantes del MTFD .....	105
3.3.1. Las variantes objetivistas o positivistas .....	105
3.3.2. Las variantes no objetivistas .....	110
3.4. Limitaciones y debates en torno al MTFD .....	124
Caso práctico .....	131
Recurso electrónico 3.1 .....	132

<b>4. El análisis semiótico: en torno al nivel estructural .....</b>	<b>133</b>
4.1. Análisis como búsqueda de las estructuras subyacentes .....	134
4.1.1. El estructuralismo como conglomerado teórico básico .....	138
4.1.2. Hacia una clarificación de los principales conceptos .....	141
4.2. Decodificación y recomposición .....	146
4.2.1. Las funciones del lenguaje .....	147
4.2.2. Relaciones sintagmáticas y relaciones paradigmáticas .....	148
4.2.3. Mitos y mitemas .....	149
4.2.4. Las lógicas estructurales .....	152
4.2.5. Las metáforas y otras figuras retóricas .....	153
4.3. Limitaciones y debates en el análisis semiótico-estructural .....	156
Caso práctico .....	160
Recurso electrónico 4.1 .....	161
 <b>5. La teoría posestructuralista del discurso en la escuela de Essex .....</b>	 <b>163</b>
5.1. Posestructuralismo, discurso y sujeto del discurso .....	163
5.1.1. La primacía de lo político sobre lo social .....	167
5.1.2. La noción de “discurso” en la escuela de Essex .....	168
5.1.3. El sujeto en la teoría del discurso de la escuela de Essex .....	172
5.1.4. El análisis del discurso como deconstrucción .....	174
5.2. Los pilares conceptuales del análisis sociopolítico .....	177
5.2.1. Hegemonía, antagonismo y lógicas articuladoras .....	177
5.2.2. Puntos nodales, significantes flotantes y significantes vacíos .....	181
5.3. Limitaciones y debates en el análisis en la escuela de Essex .....	185
Caso práctico .....	189
Recurso electrónico 5.1 .....	189
 <b>6. El análisis crítico del discurso: estudio del poder y la dominación ....</b>	 <b>191</b>
6.1. Elementos transversales: poder, contexto y crítica social .....	191
6.1.1. La noción de discurso en el ACD .....	195
6.1.2. El influjo de la obra de Foucault .....	198
6.1.3. Ideología dominante, poder y desigualdad .....	200
6.1.4. Análisis crítico como crítica de la sociedad .....	202
6.1.5. La centralidad del análisis contextual .....	204
6.2. Enfoques y diversidades en el ACD .....	206
6.2.1. El modelo dialéctico-relacional de Fairclough .....	207
6.2.2. El “enfoque histórico del discurso” de Wodak .....	211



6.2.3. <i>El modelo sociocognitivo de Van Dijk</i> .....	215
6.3. Limitaciones y debates en el ACD .....	218
Caso práctico .....	223
Recurso electrónico 6.1 .....	224
<b>7. <i>El análisis sociológico de los discursos: hacia el sistema de discursos</i></b> .....	<b>225</b>
7.1. Tradiciones y referentes en el análisis sociológico de los discursos .....	226
7.1.1. <i>La tradición de la escuela cualitativista madrileña</i> .....	227
7.1.2. <i>El nivel ideológico en la “maraña” de discursos</i> .....	228
7.1.3. <i>Hermenéutica, pragmática y dialogismo</i> .....	231
7.1.4. <i>La noción contextual del discurso</i> .....	240
7.1.5. <i>El complejo y cambiante sistema de discursos</i> .....	241
7.2. La “caja de herramientas” incorporadas a las prácticas .....	244
7.2.1. <i>Herramientas confluentes, transformadas y aportadas</i> .....	245
7.2.2. <i>Una tentativa (flexible) de modelización de las etapas</i> .....	268
7.3. Limitaciones y debates en el análisis sociológico del discurso ....	271
Caso práctico .....	277
Recurso electrónico 7.1 .....	277
<b><i>Bibliografía seleccionada</i></b> .....	<b>279</b>

## RECURSOS DIGITALES



### Bibliografía completa

- Recurso electrónico 2.1. Estudio de un ejemplo de análisis de contenido
- Recurso electrónico 3.1. Estudio de un ejemplo de uso del MTDF
- Recurso electrónico 4.1. Estudio de dos ejemplos de análisis semiótico
- Recurso electrónico 5.1. Estudio de un ejemplo de análisis siguiendo a la escuela de Essex
- Recurso electrónico 6.1. Estudio de un ejemplo de uso del análisis crítico del discurso
- Recurso electrónico 7.1. Estudio de un ejemplo de análisis sociológico del discurso

## 2

# *El análisis de contenido de textos: enfocando presencias y frecuencias*

En el seno de la investigación social, un aspecto relevante en el análisis de los materiales cualitativos es el acceso a su dimensión informacional. En este nivel, el análisis de contenido viene a constituir una de las tradiciones con una trayectoria más temprana y dilatada en el espacio y el tiempo, y está muy presente en la actualidad, en el contexto de las potencialidades desarrolladas con el acceso masivo a los datos (míticamente denominados *big data*). Muchas son las variantes y usos de esta modalidad de análisis, si bien en este capítulo se consideran solo algunas de ellas, las que creemos más relevantes para la práctica empírica de la sociología y otras ciencias sociales. En la primera parte de este capítulo, se abordan tres modos de análisis de contenido, interconectados y con claros entrecruzamientos, orientados por un objetivo de sistematización cuantitativa del análisis. El análisis clásico de contenido, las elaboraciones más recientes de este y la lingüística del corpus constituyen las tres variantes que se consideran, sucintamente, en este primer apartado. En un segundo momento, se contempla una de las orientaciones del análisis de contenido que no persigue el objetivo de la cuantificación: el denominado análisis temático, muy frecuente en la investigación empírica, aunque menos formalizado y reflexionado en los textos metodológicos que abordan esta tradición. Lo sigue un tercer apartado, en el que se da cuenta de las que se consideran las principales limitaciones y debates que se observan en esta tradición, desde la mirada de las personas que firmamos el presente texto. Finalmente, en la cuarta parte del capítulo, se expone un ejemplo prototípico de uso de la variante de análisis de contenido, en sus manifestaciones más recientes, aplicado a un campo en expansión: el análisis de las redes sociales (en este caso, se utiliza la red Twitter). Se trata de un estudio que considera la presencia en dicha red de elementos vinculados al rechazo hacia la población inmigrante y refugiada.

### **2.1. Hacia la cuantificación como objetivo privilegiado**

La tradición del análisis de contenido es una de las más antiguas en el campo del análisis de materiales cualitativos, si bien, frecuentemente, no es considerada dentro de las

corrientes de análisis del discurso. En muchas ocasiones, se aborda como un proceso de análisis previo; en otras ocasiones, se considera una primera etapa del análisis del discurso; en otras, un proceso de análisis radicalmente distinto y con objetivos diferentes; e incluso, en otras, el momento final de la dimensionalización de los discursos. Desde nuestra perspectiva, preferimos considerarlo una aproximación al análisis de textos (que no de discursos) orientada hacia su nivel más básico: la dimensión informacional.

Aunque bajo la etiqueta “análisis de contenido” se encuentra un enorme conjunto de prácticas de muy diverso carácter, en términos generales implica la articulación de una estrategia cuantitativa a la hora de analizar los materiales cualitativos, siendo un análisis que permanece en el nivel más inmediato, el denotativo, y que no implica, propiamente, una consideración del discurso en ninguno de los sentidos ni las conceptualizaciones revisadas en el capítulo 1. Se vincula con un primer nivel de análisis con respecto a los materiales: la orientación externa, según la cual se enfocan exclusivamente los aspectos manifiestos y puramente informacionales en su aproximación a los materiales cualitativos.

Compartiendo el objetivo central de la cuantificación, se puede señalar la existencia de tres variantes en sus usos. La primera de ellas sería el análisis clásico de contenido. Por su parte, la segunda variante abarcaría las aplicaciones y adaptaciones metodológicas más recientes que tratan de dar respuesta a parte de las críticas vertidas sobre los modelos más clásicos y que han tomado forma paralelamente al desarrollo de nuevas posibilidades técnicas en la acumulación, el acceso y el procesamiento de ingentes cantidades de información digitalizada. Por último, la tercera variante se correspondería con la propuesta más reciente de la denominada “lingüística del corpus”. Los tres enfoques —análisis clásico de contenido, transformaciones recientes de este y lingüística del corpus—, aun partiendo de premisas teóricas y metodológicas muy parejas, desarrollan tradiciones académicas diferentes, aunque la lógica bajo la que trabajan es muy similar. El desarrollo de la primera variante se ha concentrado, especialmente, en el análisis de los medios de comunicación, partiendo de un primer momento de precodificación y una posterior fase de cuantificación de apariciones y presencias. Una reflexión más actual sobre las potencialidades y los límites metodológicos y epistemológicos de esta variante se encuentra en aproximaciones de corte más sociológico, que incorporan los nuevos métodos de acceso a los materiales, así como las enormes posibilidades desplegadas recientemente en el tratamiento y el procesamiento de la información. En el caso de la lingüística del corpus, proveniente de una tradición lingüística y, por ello, menos vinculada al estudio de fenómenos sociales, se parte, generalmente, del trabajo de cuantificación (atendiendo a frecuencias de aparición y de colocación, cuyo análisis tratan de automatizar) a partir de grandes corpus de materiales (textos) obtenidos en contextos naturales que se convierten en repositorios de la lengua en uso, que no necesariamente tienen por qué reflejar el contenido de los medios de comunicación. Aunque el espacio de los medios acostumbra a ser uno de los campos de análisis fundamentales, también se encuentran análisis de redes sociales, literatura, libros de texto, grandes repositorios institucionales, etc.

En la variante del análisis clásico de contenido, generalmente, aunque no siempre, se parte de un trabajo previo de establecimiento de códigos, a veces contruidos por el equipo investigador, en otras ocasiones producto de un trabajo de jueces y expertos en el ámbito que se aborda. Estos códigos son, frecuentemente, derivados de la teoría existente (o de investigaciones previas) en relación con el tema que se considere. En la variante de la lingüística del corpus, con un carácter mucho más inductivista en la mayoría de sus aplicaciones, el análisis se basa en considerar la presencia de determinadas señales lingüísticas, el análisis de sus frecuencias, sus colocaciones y sus relaciones (correlaciones o copresencias, fundamentalmente). Las principales modalidades en la tradición del análisis de contenido pueden tomar como base del análisis elementos informacionales diferentes:

- *Las palabras y otro tipo de señales.* Es lo más frecuente en las aproximaciones etiquetadas como lexicométricas (Tournier, 1993). Se hace hincapié en la presencia y frecuencia de determinadas señales lingüísticas, y se realiza un análisis cuantitativo de las mismas. Progresivamente, se ha incluido la consideración de las señales no lingüísticas (icónicas, objetuales, sonoras, musicales, corporales...). Es lo que Palmquist (2010) ha denominado “análisis conceptual”.
- *Las categorías o códigos asignados a un contenido informacional.* En ocasiones, son valoraciones (positivas, negativas o neutras, de aceptación, rechazo o indiferencia, por poner algunos ejemplos), o códigos breves temáticos (etiquetas) asignados a textos o fragmentos de textos; también pueden ser cualquier otro criterio de clasificación que se considere relevante y que podría derivarse del tipo de vocablos usados u otras señales presentes en relación con una determinada temática (problema, grupo, entidad, partido). Para poder categorizar y codificar el contenido, se debe producir un proceso previo de reflexión y un trabajo de establecimiento de los códigos, que se acompañará de un entrenamiento de las personas encargadas de la codificación. En este sentido, se podría hablar de análisis categorial (Bardin, 1996).
- *Las relaciones.* En este caso, lo que se considera en el análisis son las redes conceptuales (algunos autores las denominan redes semánticas) que surgen del registro de la aparición simultánea (coincidencia o coocurrencia) o la proximidad de palabras, imágenes u otras señales. Estas redes conforman conjuntos de densidad variable e implican diferentes grados de cercanía entre los elementos que las conforman. Asimismo, se consideran las relaciones existentes entre dichas redes y los sujetos, grupos, entidades o instituciones que elaboran o producen los textos que se analizan. Comprenderían lo que se ha venido a denominar análisis relacional (Palmquist, 2010). Existe una multiplicidad de herramientas técnicas desarrolladas para realizar este tipo de análisis, muchas de ellas de acceso gratuito (por ejemplo, el TXM, un *software* de lexicometría), herramientas que permiten realizar el análisis de un concepto central considerando las palabras que aparecen antes y después del mismo.

Quien desee profundizar en los procedimientos del análisis de contenido puede tomar como guía los conocidos textos de Krippendorff (2012) (existe versión en castellano de la ya lejana primera edición en inglés, de 1980) y Neuendorf (2017), ambos con varias ediciones ampliadas a sus espaldas. Sobre las técnicas de generación de redes semánticas, recomendamos el reciente trabajo de Escobar y Tejero (2018). Describiremos aquí solo algunas de las características, los procedimientos y las herramientas más presentes en este conjunto diverso de aproximaciones.

### *2.1.1. El análisis clásico de contenido*

Una de las primeras versiones que emergieron, en un intento de abordar el análisis de materiales cualitativos, fue lo que ha venido a llamarse “análisis clásico de contenido”. Este método se desarrolló en Estados Unidos a lo largo de las primeras décadas del siglo xx y fue sistematizado por Berelson en un texto clásico (y que se considera canónico) titulado *Content Analysis in Communication Research*. Si bien surgió en el espacio del análisis de los medios de comunicación, rápidamente aparecieron aplicaciones en la investigación política y social.

Este análisis de contenido, en su origen, articula una metodología puramente cuantitativa. La referencia, en su primer desarrollo, es la celeberrima definición de Berelson (1952: 18) “el análisis de contenido es una técnica de investigación para la *descripción objetiva, sistemática y cuantitativa* del contenido manifiesto de la comunicación”. Aplicado fundamentalmente a los medios de comunicación, desarrolla una amplia línea de trabajo que se extiende al estudio de la opinión pública, con autores como Laswell y Osgood, que se orientaron hacia el uso de estos procedimientos de sistematización objetivista en el análisis de la comunicación. Su desarrollo ha implicado derivas diversas hacia procedimientos cuantitativos que integran, también en ocasiones, elementos cualitativos; si bien, salvo excepciones, el momento cualitativo se reduce a la categorización o codificación, que, aunque constituye una fase crítica de cara a la producción de los datos, siempre aparece subordinada a la finalidad cuantificadora.

Años más tarde, Bardin, otro de los grandes referentes en esta modalidad de análisis, amplía la delimitación clásica del análisis de contenido y pasa a definirlo como “el conjunto de técnicas de análisis de las comunicaciones, tendentes a obtener indicadores por procedimientos sistemáticos y objetivos de descripción del contenido de los mensajes” (Bardin, 1996: 32). Si bien este autor menciona la relevancia de incorporar la consideración tanto del contexto de la producción como del de la recepción de los mensajes, en sus aplicaciones concretas estos contextos pocas veces son tenidos en cuenta, o se vinculan exclusivamente a variables sociodemográficas aplicables a los sujetos o espacios productores (o soporte) de los textos (varón-mujer, rural-urbano, políticos-técnicos-activistas, prensa-redes sociales-blogs...).

Como indica Luis Enrique Alonso, se trata de la forma de aproximación a los textos más cercana al nivel informacional; esto es, remite al aspecto “más inmediato y *descrip-*

tivo y tiende a explorar la dimensión más *denotativa y manifiesta* de los textos. Operativamente, tiende a realizarse de manera *cuantitativa*; de lo que se trata, fundamentalmente, es de analizar y desmigajar un *corpus* [...] en un conjunto de palabras, entre las que se producen *reiteraciones*, en las que hay *asociaciones* posibles y a las que se pueden atribuir lejanías o proximidades semánticas” (Alonso, 1998: 189). Así, siguiendo a este autor, los textos se consideran, tratan y analizan convirtiéndolos en un espacio de frecuencias y asociaciones, de ocurrencias y recurrencias. Se articula, así, un análisis de presencias y, en su caso, asociaciones, de proximidades o distancias, de coocurrencias o ausencias. El proceder concreto de las formas de análisis más clásicas remite al cómputo de la frecuencia de aparición de las palabras u otras señales, de los términos o elementos usados (su redundancia) como elemento clave para concretar su relevancia en el discurso. De esta manera, el léxico empleado, el vocabulario, las palabras y los iconos se transforman en índices de la relevancia en relación con el fenómeno de estudio y las formas de concebir el mismo. Cuanto más se repite un término, un elemento icónico, un sonido, este se muestra, necesariamente, más importante, de manera que aglutinará una parcela mayor de significado. También se habla de “mundos lexicales” (Reinert, 1987, 1993) o “trenes léxicos” para referirse a concatenaciones de palabras evocadas por los enunciadorees (es decir, identificando la relevancia por proximidad). Con ello, y a través de una asignación de las características de los emisores (sujetos, entidades, medios u organizaciones) y de los tiempos en que se ubican, pueden establecerse comparaciones de interés para la investigación. Las lógicas y características básicas de esta manera de operar serían las siguientes:

- *Lógica de la búsqueda de objetividad.* Los resultados deben distanciarse de cualquier forma de subjetividad. Las reglas que se establecen y las categorías que se conforman sirven, así, para reducir al mínimo la posible influencia del analista, de manera que se neutralicen sus prenociones y posicionamientos, que deben orientarse a poder ser replicadas (consiguiendo, así, fiabilidad).
- *Descriptivismo.* Se presenta un proceso de análisis “objetivo”, en el que no hay interpretación, sino pura descripción. De ahí que el analista como sujeto desaparezca, aparentemente, en favor de un análisis —como se ha visto— pretendidamente objetivo de la información. En esa línea, Bardin caracteriza el análisis de contenido en términos de “tratamiento de la información contenida en los mensajes” (1996: 25). Es decir, su orientación analítica se fundamenta en la descripción y el tratamiento de la información explícita contenida en los textos (esto es, el plano denotativo). En esa línea de objetivación y de disolución de la dimensión subjetiva, tanto en el análisis como en el proceso comunicativo “el destinatario es presentado como un decodificador pasivo” (Alonso, 1998: 192). No se considera al sujeto ni en el ámbito de la producción del texto, ni en el de su recepción o su distribución.
- *Mecanización.* El análisis se desarrolla bajo una “lógica ‘taylorizada’ de descomposición del texto en unidades mínimas y de recomposición técnica” (Alon-

so, 1998: 190), con el fin de dimensionar y cifrar. En este sentido, el análisis de contenido proyecta una imagen mecanizada del proceso de análisis, que se desarrolla en función de un entrenamiento previo tendente a la estandarización y la protocolización de los procedimientos. Tiende a ofrecerse, generalmente, una suerte de pasos o etapas que se han de ir cubriendo para analizar un conjunto amplio de textos.

- *Proceso: inductivista o deductivista.* Las versiones más clásicas se organizaban, generalmente, con una orientación puramente inductivista, según la cual se procede a la generación de categorías a partir de la observación reiterada de apariciones de palabras u otras señales. Desde sus inicios, se produjo una orientación en la corriente dominante del análisis de contenido, conforme a la cual “tratar el material es codificarlo. La codificación corresponde a una transformación —efectuada según reglas precisas— de los datos brutos del texto” (Bardin, 1996: 78). Este tipo de análisis es denominado por algunos autores “análisis categorial”. Mediante él, se intenta “tomar en consideración la totalidad de un ‘texto’ para pasarlo por el molinillo de la clasificación y de la enumeración por frecuencia de presencia (o de ausencia) de ítems de sentido” (Bardin, 1996: 28). De esta manera, se prioriza un proceder inductivista. Sin embargo, en muchas ocasiones la lógica obedece a la formulación *ex ante* de códigos derivados de la teoría o de análisis previos, que se definen con reglas de asignación estrictas, que son aplicadas a los textos para comprobar el grado de presencia, las frecuencias y, en algunos casos, las relaciones de proximidad, distancia o coocurrencia.
- *Lógica de la representatividad.* Todo el material que se considere relevante debe ser analizado; esto implica que se ha de trabajar con el universo, o con una muestra importante y, a ser posible, representativa, siguiendo criterios estadísticos; no se trata de trabajar con una selección orientada o designada para ilustrar determinadas hipótesis, o delimitada según criterios estructurales o teóricos. La lógica de la selección de los elementos que vayan a ser analizados es, por lo tanto, una lógica probabilística —cuando no sea posible trabajar con los universos— y, con ello, asentada en la aleatorización.
- *Visualidad de los resultados.* El análisis permite la representación gráfica de las presencias o relaciones en forma de tablas, mapas de palabras, nubes de palabras, sistemas de representación espacial, etc. Así, el análisis concluye con la generación de modelos que permitan confeccionar reorganizaciones formales a partir de frecuencias de términos, vocablos, categorías o relaciones tratadas estadísticamente.

Siguiendo la lógica básica objetivista, automatizada y formalizada de este proceder analítico con materiales cualitativos, la sucesión de los pasos implicados en el método tiende a proponerse de una manera pautada, que frecuentemente adopta la siguiente forma:



- a) Seleccionar el tema y perfilar el problema, el grupo y la etiqueta que analizar, confeccionando y delimitando con precisión los objetivos de la investigación y los conceptos que los sintetizan.
- b) Seleccionar el corpus de los documentos que serán analizados, de manera que se consiga una muestra amplia y representativa. En ocasiones, se trabaja con el universo de los textos localizables en un espacio temporal bien delimitado.
- c) Tomar decisiones sobre si el trabajo se va a vincular a la presencia de señales (palabras, iconos, sonidos, personajes, *hashtags*...) o a las agrupaciones de estas en torno a categorías. Estas categorías son conformadas a partir del estudio previo de la teoría relacionada con el tema o de la participación de los expertos que elaboran la categorización (jueces, como son denominados en ocasiones), o del análisis previo de una muestra pequeña de documentos, que permite elaborar un sistema de códigos que posteriormente se aplica al conjunto del material (códigos que agrupan conceptos o palabras o temas y que se configuran por proximidad semántica). En otras ocasiones, estos códigos se conforman en relación con atributos que se vinculan a aquello que se está estudiando: antiguo, nuevo; rural, urbano; masculino, femenino..., o cualquier calificación o atributo que se considere relevante. Muchas veces los códigos se elaboran, de forma progresiva, durante los procesos de análisis, en procedimientos inductivos y deductivos que se suceden de manera iterativa. Se señala que las categorías deben reflejar el objeto de estudio y ser sensibles al mismo.
- d) Definir de forma clara, exhaustiva y excluyente el conjunto de categorías considerado, de manera que el contenido de los textos, imágenes, sonidos o personajes que se analizan se pueda codificar de manera no ambigua en función de ellas. Resulta estrictamente necesario configurar un sistema de categorías que sean definidas de forma clara y sin solapamientos (tarea que se convierte en uno de los aspectos más difíciles de lograr, dada la naturaleza polisémica, connotativa y contextual que caracteriza el lenguaje y los actos comunicativos).
- e) Explicitar unas reglas que provean de claves para poder asignar señales o términos a las categorías, de manera que cualquier persona codificadora (o máquina) entrenada pueda realizar la asignación sin dudas, ni ambigüedades. Estas reglas deben configurarse atendiendo, especialmente, a los espacios en los que haya una mayor ambigüedad a la hora de realizar la asignación. En las ocasiones en las que se generen problemas de asignación, se construyen reglas o se nombran jueces para las excepciones —que se aspira a que sean limitadas y resolubles—. Como se señalaba previamente, esta es una de las operaciones más cargadas ideológicamente, y es realizable solo a costa de reducir la polisemia, la multidimensionalidad y la contextualidad intrínsecamente vinculadas a la comunicación y, con ella, a los lenguajes. Sin embargo, la complejidad de esta operación y la reducción intrínseca asociada pasan muchas veces desapercibidas sin el despliegue en el análisis de la indispensable vigilancia epistemológica.



- f) Designar y entrenar a las personas (y, a veces, a las máquinas) codificadoras. Frecuentemente, se elaboran materiales y textos de apoyo (a menudo, son manuales completos) que garanticen una cierta estandarización de los procedimientos. En ocasiones, y cuando es posible, se desarrollan mediciones de congruencia de las asignaciones.
- g) Contar y medir el volumen de las frecuencias de las palabras-señales o categorías que se hayan considerado relevantes. En algunos casos, se trabaja con análisis multivariantes.
- h) Considerar su distribución en función de criterios relevantes (género, territorio, clases sociales, medios, canales, entidades, fechas, etc.).
- i) Observar las relaciones existentes entre los elementos que se designen como importantes: identificar proximidades, coocurrencias, etc.
- j) Proceder a componer la visualización del conjunto de los datos obtenidos: elaborar tablas con resultados del cómputo de las frecuencias, nubes de palabras o gráficos que esquematicen las presencias (ocurrencias) de señales o las relaciones (copresencias, proximidades, etc.).

### *2.1.2. Transformaciones recientes del análisis de contenido cuantitativo*

A pesar de que se trata de una modalidad de aproximación a los textos que ha sido objeto de intensas críticas en sus modalidades más clásicas, en los contextos contemporáneos, especialmente tras lo que se ha llamado “el giro digital”, se aplica, con asiduidad y visibilidad e, incluso, con una acogida entusiasta, en las aproximaciones más formalizadas y reconocidas desde parámetros de la ciencia hegemónica, especialmente en sus versiones de lexicometría y de análisis de redes semánticas.

Gracias al descomunal potencial técnico de mecanización de la recogida, el archivo, la organización y la sistematización de la información, derivado del desarrollo computacional producido a lo largo de las dos últimas décadas, se ha generado una enorme capacidad de aplicación de procedimientos automáticos al análisis de textos. Gracias a estas posibilidades, se han multiplicado los análisis que trabajan con *software* de cuantificación textual y análisis de “colocación” a partir de miles, cientos de miles o millones de palabras (u otras señales), textos y mensajes, que son recopilados y seleccionados por aplicaciones y programas cada vez más potentes, accesibles y con mayor grado de difusión en su uso. Se trata de programas que poseen la capacidad de contar y cifrar apariciones en relación con estos enormes corpus de materiales, de considerar las apariciones simultáneas o *coincidencias*, de analizar las colocaciones y proximidades, así como de asignar categorías para aquello que constituya apariciones o presencias de señales (palabras, imágenes, nombres, expresiones, *hashtags*, etc.) en función de un entrenamiento previo de las máquinas (identificado, principalmente, bajo la etiqueta de *machine learning*).

Se señala que este tipo de aproximación puede ayudar a explorar “las macroestructuras léxico-semánticas en materiales textuales” (Scholz, 2019b: 146), considerando

el lenguaje específico de determinados grupos, entidades o comunidades, a veces personajes (por ejemplo, líderes políticos), o los temas (códigos-descriptores-categorías) más presentes en unos determinados materiales, de manera que puedan ser agrupados y representados los usos de léxicos coincidentes, similares, próximos o coocurrentes en función de las características que se consideren relevantes. También permite abordar los cambios y transformaciones en el léxico de los grupos, personas, entidades, repositorios o instituciones en el tiempo (en momentos concretos o cortes temporales).

Como se señalaba previamente, este tipo de aproximaciones se tienden a englobar bajo la etiqueta de “análisis de redes semánticas”. Se analizan, así, las relaciones que se producen entre los términos, las señales o las palabras que se especifican *a priori*, o entre las agrupaciones de estas, de manera que se puedan establecer redes conceptuales o semánticas siguiendo criterios de coocurrencia (simultaneidad de la aparición) o proximidad de las señales o los códigos, pudiendo delimitarse agrupaciones vinculadas a sujetos, entidades, organizaciones, espacios o tiempos, así como analizar las relaciones que se establecen entre ellos.

Para ilustrar algunas de las aplicaciones de estas redes semánticas tomando como base muy diversos tipos de materiales, se recomienda la lectura de alguno de los siguientes textos: Escobar (2009), Escobar y Román (2011), Escobar y Gómez (2015), Soriano *et al.* (2021) y Escobar, Gil y Calvo (2022). Algunos de los materiales más frecuentemente utilizados en estas investigaciones son, por ejemplo, textos periodísticos, tuits, artículos científicos, fotografías, blogs o páginas personales, que sirven como base para el estudio de diversas problemáticas, como puede ser el posicionamiento ante la violencia sexual (condensada en un *hashtag* en Twitter; por ejemplo, *#cuéntalo*), o temas priorizados en áreas científicas, o formas de presentación de los sujetos en redes sociales como vía de acceso a la identidad, o vínculos entre personas relevantes en un campo. Así pues, en múltiples ocasiones el objetivo no es tanto analizar presencias como localizar sus “colocaciones”, proximidades y términos con los que coocurren. Muy a menudo, esta modalidad de análisis tiende a denominarse *natural language processing (NLP)*, esto es, “procesamiento del lenguaje natural”.

La confianza en la potencialidad y el alcance (predictivo, llegan a decir incluso algunos autores) de este tipo de búsquedas y análisis es tal que, en ocasiones, adquiere lo que puede llamarse un carácter mítico. En algún artículo, con evidente ingenuidad, se ha llegado a afirmar que, si se hubieran utilizado estas técnicas de análisis aplicadas a las conversaciones en la red entre los médicos y los pacientes, se podría haber evitado o frenado la crisis sanitaria del COVID-19: “Si se hubiesen usado en tiempo real los análisis semánticos NLP, se habría entendido en modo consistente que lo que estaba sucediendo no era un simple brote de pulmonía, sino una nueva pandemia” (Jiménez y Moreno, 2020).

También en sus desarrollos contemporáneos, en un intento de contrarrestar las intensas críticas recibidas por este tipo de análisis, por su desatención y su descuido con respecto a la dimensión afectiva –irremediabilmente vinculada a los lenguajes–, se ha desarrollado una multiplicidad de investigaciones (y programas informáticos)

centradas en tratar de incorporar dicha dimensión. En este sentido, se intentan vincular los contenidos de los textos que se vayan a analizar con las cargas valorativas asociadas (como indicadores de afectos), que acostumbran a presentarse en forma de categorías (por ejemplo: atributos positivos, negativos y neutros). En esta línea, cada vez es más frecuente encontrar un intento de articular esta tradición de análisis de contenido con orientaciones críticas y de denuncia, que buscan formas de operacionalizar, cuantificar y considerar cómo se manifiestan y articulan los estereotipos y procesos de discriminación con respecto a determinados grupos, colectivos o problemáticas, siempre trabajando con enormes volúmenes de textos. Así, se implementan búsquedas de textos que contengan una o más palabras clave en relación con colectivos o problemas (por ejemplo: feminismo, o colectivo LGTBIQ+, o cambio climático, o refugiados) o acontecimientos (pueden ser “guerra de Ucrania” o “frontera sur de Melilla”) o determinadas claves (entre otras, el análisis de *hashtags*, como #cuéntalo o #metoo) o personajes (líderes políticos, periodistas o personajes públicos) o entidades, organizaciones, etc., y se analizan las preferencias léxicas, las valoraciones asociadas, las colocaciones, las proximidades y las coocurrencias de los términos más frecuentes, tratando de desvelar las implicaciones vinculadas a dichas copresencias. Se hacen, así, búsquedas sistemáticas de conceptos clave que se tratan de analizar (por ejemplo: feminismo o refugiados) utilizando operadores booleanos y tesauros disponibles o en proceso de creación, y se consideran los términos con los que coocurren, incluyendo, también, su orientación de valor, es decir, si son términos que evocan situaciones positivas y de aceptación, o negativas y que generan rechazo (tales como la tendencia a vincular de forma sistemática los conceptos de inmigración y delincuencia, o de feminismo y autoritarismo).

Desde esta perspectiva, se trata de superar (aunque de forma muy limitada, parcial y reductora de la multidimensionalidad, añadiríamos por nuestra parte) ciertas críticas vertidas sobre el análisis de contenido, intentando aproximarse a una de las dimensiones de la polisemia de los conceptos, considerando cómo estos pueden adquirir valores positivos o negativos en función de dónde ocurren, junto con qué otros términos aparecen, o próximos a qué otras unidades léxicas se hallan. Se incide, así, en la consideración del contexto —una definición muy estrecha de contexto, diríamos desde nuestra perspectiva— a la hora de dar sentido a las implicaciones del uso de un determinado vocablo, o una señal icónica o de cualquier otro tipo.

No obstante, al margen de las idealizaciones míticas de la potencialidad abierta por las ciencias de la computación, así como de las tecnologías de la información, la mayor parte de estas modalidades de análisis tienden frecuentemente a presentarse como aproximaciones exploratorias que requieren de profundización cualitativa y otras formas de abordaje complementarias que permitan un análisis de mayor profundidad.

Por otra parte, y en relación con lo anterior, una variante de las formas de análisis de contenido que está cada vez más presente tanto en la literatura de carácter metodológico como en las investigaciones empíricas concretas es aquella que se orienta

a un tipo de cuantificación que se desarrolla como última fase de un proceso previo de análisis del discurso más complejo, que desemboca en un sistema de indicadores categoriales manifiestos, en un intento de cuantificación y dimensionalización de la presencia de determinados elementos discursivos previamente analizados en profundidad (o analizados en contextos investigativos precedentes) y que se traducen en un sistema de códigos, etiquetas y reglas de asignación. A partir de un análisis del discurso, se procede a desarrollar una suerte de categorías o indicadores manifiestos, que se definen de forma operativa (generalmente, en forma de categorías excluyentes) para poder computar su presencia en un corpus de textos mucho más extenso. Por poner un ejemplo de este tipo de uso particular, en un estudio relativamente reciente sobre los llamados discursos del odio dirigidos contra población inmigrante y de etnia gitana (Cabo *et al.*, 2015), se diseñó un sistema de categorías orientadas a tipologizar tres tipos de discursos: discurso intolerante (con cinco variantes, que se etiquetan como “estereotipos y prejuicios”, “odio”, “rumores”, “discurso antiinmigrante o antigitano de baja intensidad” y “de argumento trampa”), discurso alternativo y discurso neutro. A partir de esta categorización, y después de un intenso proceso de entrenamiento de codificadores, se clasificaron más de cuatro mil comentarios de foros de noticias de tres periódicos de ámbito nacional de gran tirada, considerando la distribución de estos discursos, sus cambios en el tiempo, su presencia en los diferentes periódicos y las consecuencias derivadas de pequeñas intervenciones (provocadas) en los foros por parte del propio equipo de investigación. En este tipo de aproximaciones, si bien el objetivo es también la cuantificación a partir de la presencia de elementos en el texto, el proceso previo de construcción de las categorías, a partir de análisis más profundos, dota de calidad interpretativa y comprensiva a las aproximaciones, las cuales no quedan exentas, no obstante, de buena parte de las limitaciones que caracterizan esta tradición y que se considerarán más adelante.

### *2.1.3. La lingüística del corpus*

Una de forma de análisis de contenido que presenta muchas coincidencias con algunos de los desarrollos comentados en el apartado anterior es la que ha venido a denominarse lingüística del corpus. Muchos son los elementos que tiene en común con ellos, pero también cuenta con algunas especificidades y peculiaridades que conviene destacar.

En primer lugar, cabe señalar que es un área disciplinar (que muchos se esfuerzan en etiquetar, reiteradamente y con énfasis, como científica) que se ha desarrollado prioritariamente en el seno de la lingüística, orientada hacia el estudio del lenguaje natural o, como lo describen algunas de las personas más representativas, el estudio del lenguaje en sus “usos reales” (Sinclair, 1991), tomando como base para ello enormes corpus de materiales procesados computacionalmente. Destacan las figuras de Henry Kucera y Nelson Francis, quienes, ya en el final de los años sesenta, hablaban de esta subdisciplina (difícilmente podían imaginar estos autores el grado de desarrollo que adquirirían,

tan solo unas décadas más tarde, los programas informáticos de tratamiento textual). Otro de los principales representantes es John Sinclair, lingüista británico y profesor de la Universidad de Birmingham, que constituye una referencia central en esta tradición. Este autor formó parte de varias iniciativas que consolidaron la lingüística del corpus como subdisciplina (entre ellas, la elaboración de un diccionario del uso del inglés con millones de entradas). Actualmente, existen varios centros de lingüística del corpus en la Universidad de Birmingham, así como el Tuscan Word Centre, fundado por el propio Sinclair para impartir cursos en relación con esta variante analítica.

El origen de esta tradición condiciona, en buena manera, su interés y su centramiento en los lenguajes naturales tal y como se desarrollan en sus usos cotidianos y en las situaciones reales de interacción (simultánea o diferida). Una característica particular es que la lingüística del corpus no acude a la categorización previa (al menos, en el grueso de sus aportaciones), sino que, siguiendo el espíritu del primer análisis de contenido, estudia fundamentalmente la presencia de términos y vocablos, así como su colocación (Stubbs, 2001) y su coaparición, esto es, la proximidad y la concurrencia de palabras u otras señales, analizando las frecuencias asociadas a dicha colocación, como vía exclusiva de acceso al significado del texto. Es decir, el proceder es siempre inductivista, yendo directamente a los textos y los materiales, sin partir de un proceso de codificación de estos. Generalmente, se toma este proceso inductivo como un “acceso diáfano” a los usos generales y cotidianos del lenguaje.

La lingüística del corpus es dependiente de la posibilidad de creación de enormes corpus de materiales lingüísticos, que se someten a análisis, potenciados e impulsados por el desarrollo computacional del tratamiento cuantitativo de los textos, el análisis de redes sociales y otros espacios de almacenamiento de grandes cantidades de información (coincide con la tendencia ya señalada en el apartado anterior). Además, este acceso masivo permite el desarrollo de análisis multivariantes de los datos obtenidos. Generalmente, es utilizada en aproximaciones a volúmenes muy grandes de textos presentes en medios de comunicación, internet o redes sociales (muchas veces, se trabaja con millones de palabras o imágenes, miles de artículos, noticias, programas, tuits...) y está potenciada por la creación paralela de grandes corpus representativos *–a priori–* del lenguaje natural (Jaworska y Krishnamurthy, 2012), tal y como se manifiesta en los contextos naturales digitales. Así, estos contextos permiten trabajar con enormes cantidades de textos, disponibles en internet y, de manera más concreta, en las redes sociales, en lo que pueden denominarse “contextos de comunicación digital” (Scholz, 2019b). Con la proliferación y el desarrollo del *big data* y la inteligencia artificial, así como de las técnicas de análisis automático de textos y los motores de búsqueda, y con la facilitación de la clasificación de los textos (a partir de palabras clave, *hashtags*, búsquedas en buscadores, etc.), este tipo de análisis han proliferado exponencialmente, aplicándose a conversaciones en las redes sociales –inscritas en espacios que permiten registros sistemáticos–, en el seno de aplicaciones orientadas a la publicación (más o menos abierta) o el intercambio de mensajes, o el análisis masivo de textos almacenados en grandes repositorios –por ejemplo, de prensa, literatura, etc.–.

Desde esta perspectiva, se multiplican las investigaciones que persiguen el análisis de variantes lingüísticas y sus transformaciones, los usos de la lengua de determinados grupos sociales (adolescentes, minorías, grupos caracterizados por determinadas problemáticas u otros colectivos que centran el interés de la investigación). Asimismo, se consideran fenómenos conversacionales delimitados, en los que se analiza el número de veces que aparece una serie de vocablos (usos más frecuentes) o las señales de diverso carácter, que son las que funcionan como localizadores (o algoritmos) de búsqueda: nombres, marcas, palabras clave (por ejemplo, en discursos prototipados como discursos del odio), síntomas de enfermedades, marcas institucionales, movimientos sociales, grupos vulnerables o cualquier otro fenómeno en torno al que se considere relevante recrear las conversaciones (conversaciones que conformarán el corpus).

Así, el análisis realizado desde la lingüística del corpus presenta especificidades que, de forma sintética, resumimos como las siguientes: estudio lingüístico del lenguaje natural, proceder inductivista, tratamiento de grandes cantidades de textos (corpus voluminoso) y análisis mediado por programas informáticos de recolección de textos y de medición de ocurrencias, copresencias y colocaciones. Se trata de una orientación fundamentalmente cuantitativa; aunque, en ocasiones, se encuentran aportaciones que incorporan algún elemento o momento cualitativo (Gray y Biber, 2011) e, incluso, derivaciones que intentan avanzar hacia un verdadero análisis del discurso (en lo que se ha venido a llamar estudios posdisciplinares del discurso; puede consultarse, por ejemplo, la reciente compilación de Scholz, 2019a). Con este breve resumen, se sintetizan las características de esta variante, que, como se ha señalado, se solapan, parcialmente, con las consideradas en el apartado anterior, si bien las aquí englobadas constituyen un espacio categorizado como campo o, incluso, disciplina diferenciada que ha generado una línea de trabajo, una comunidad científica y centros de estudio y revistas especializadas (como pueden ser *Corpus Linguistics Theory* o *International Journal of Corpus Linguistics*).

## **2.2. El análisis temático**

Además de estas formas de aproximación, mediadas por el conteo de palabras (u otras señales), por sus relaciones o por las valoraciones que se les atribuyan, encontramos otra modalidad de aproximación al contenido de los textos que implica el acceso al nivel informacional: se centra en el análisis de los temas presentes en los textos. Su consideración se basa en la presencia ya no de palabras (u otras señales) en concreto y su frecuencia, ni en sus relaciones o colocaciones, o de las categorías que se puedan generar, ni siquiera de las valoraciones que se les puedan atribuir, sino de un tratamiento de los bloques de significado, frases (o secuencias o conjuntos icónicos) y aspectos considerados de forma explícita, que permiten una organización del contenido con base en tópicos o espacios relacionados con el significado explícito y la función referencial de los lenguajes (esto es, la referencia a una parcela del mundo o de una realidad a la cual



remiten). Los temas constituyen el núcleo conceptual de un texto, haciendo referencia a sus ideas principales o dominantes, y en ellos se condensan las informaciones más importantes que lo articulan (Maldonado, 2003). Esto conformaría la modalidad considerada como análisis temático, en la cual se tratarían de organizar los textos, sintetizando el significado de cada una de las partes relevantes de los mismos, a modo de grandes etiquetas clasificadoras, espacios o elementos relevantes presentes de forma manifiesta, pero sin buscar en ningún caso la cuantificación.

Esta es la variante del análisis de contenido que, desde nuestra perspectiva, presenta mayor interés para el análisis sociológico. Así pues, este análisis temático (Ruiz, 2009), que correspondería a una variante no cuantitativa del análisis de contenido, viene a centrarse más en los significados que en los significantes, y, por lo tanto, tiende a desbordar en sentido estricto el contenido textual. El análisis temático implica lo siguiente:

... clasificar el *corpus* de textos de la investigación desde la perspectiva de delimitar y organizar un conjunto de temas representativos del contenido de los textos analizados que sean, al mismo tiempo, pertinentes de cara a los objetivos de la investigación (Conde, 2009: 24).

Se trata de prestar atención al contenido para señalar y clasificar los distintos tópicos abordados, la relación que se establece entre los temas –esto es, su organización interna–, el orden de aparición, su extensión, etc. (Ruiz, 2009), como punto de partida para posteriormente abordar otras formas de análisis. En ocasiones, se conforma como el objetivo central cuando no se considera necesario el acceso a otros niveles más profundos. Por ejemplo, el análisis temático es una modalidad de aproximación muy frecuente en el análisis de textos producidos por expertos, cuando los objetivos se centran y limitan a analizar lo que saben o conocen en relación con una determinada problemática.

Desde esta perspectiva de análisis, que constituye el primer paso en muchas aproximaciones, se procede, generalmente, a realizar una clasificación de los temas, previamente categorizados o codificados, sobre el objeto de estudio, de tal forma que, a partir de códigos o categorías temáticas, se realiza un análisis semántico de los tópicos que emergen y que más frecuentemente se abordan, considerando una serie de criterios relevantes para la investigación (ya sean el país de procedencia, el género, la clase social, el medio de comunicación o cualesquiera otros que se consideren importantes para el análisis en desarrollo). Como se señalaba, en muchas investigaciones esta constituye la primera fase, antes de desarrollar otros tipos de análisis que impliquen ahondar en niveles que trascienden la consideración de los elementos meramente presentes en los textos.

En ocasiones, se habla de la realización de un “vaciado temático”, como primera aproximación; metáfora esta –la del vaciado– que da cuenta del nivel superficial bajo el que se aborda el tratamiento de la información; como si de un espacio lleno se extrajeran los componentes explícitos (vaciando el texto). En muchas ocasiones, en la investigación que se produce en el mercado de la investigación social, habitualmente