Se hace camino al andar

En las páginas precedentes se ha llevado a cabo un recorrido sobre los principales temas y enfoques teóricos adoptados en los últimos años para abordar las transformaciones de la lengua española en los contextos de la comunicación digital, así como el impacto que han tenido la innovación tecnológica, la difusión de móviles y la llegada de los algoritmos en nuestras prácticas discursivas. El camino andado es largo y complejo y, como apuntaba en la conclusión del segundo capítulo, se nos presentan diversas encrucijadas. Sin duda, la investigación sobre la comunicación digital está bien asentada, si se atiende a las perspectivas teórico-metodológicas adoptadas, dentro de los estudios lingüísticos y sociolingüísticos; estudios relacionados con determinados géneros discursivos y con las lenguas de especialidad; estudios pragmáticos y sociopragmáticos; análisis críticos del discurso público; y análisis multimodales y computacionales.

Frente a la indicación de las perspectivas teóricas y metodológicas adoptadas en gran parte de la investigación, sorprende que las cuestiones de método relativas a la creación y anotación de corpus de interacciones digitales se reseñen brevemente. Cada vez más, se explicita en los estudios el método de extracción o de recopilación de los datos que se analizan, el método de análisis cuantitativo, cualitativo o mixto empleado, la importancia otorgada a los datos verbales respecto a las imágenes que conviven en un mismo mensaje; sin embargo, no se menciona si los corpus han sido anotados sintáctica o semánticamente o qué precauciones o qué cuestiones éticas se adoptan al tratar datos públicos y privados de distintas plataformas.

En la última década se han publicado distintas monografías en inglés dedicadas a la metodología en la investigación, como *Researching Language and Social Media* (Page *et al.* 2014) y *Research Methods for Digital Discourse Analysis* (Vásquez Roca 2022); en cambio, *Corpus Approaches to Social Media* (Rüdiger y Dayter 2020) y *Corpus Approaches to Language in Social Media* (Di Cristoforo 2023) aplican la lingüística de corpus al análisis del discurso en las redes. En el ámbito del español, Parini (2023) señala ya los desafíos, principalmente de orden metodológico, teórico y ético, que se presentan a los estudios sobre la comunicación digital.

Cómo citar: Pano Alamán, Ana. 2024. «Se hace camino al andar». En *La comunicación digital en español. Enfoques, métodos y perspectivas,* 155-208. Madrid: Ediciones Complutense. https://dx.doi.org/10.5209/ling.002.04

A partir de las cuestiones tratadas en estas monografías y en función de la experiencia en este campo y del fructuoso intercambio de ideas con investigadores citados previamente, exploro en este capítulo algunos puntos relacionados con la metodología de análisis de la comunicación digital, con la extracción, almacenamiento, anotación y análisis de corpus multimodales y con aspectos relativos a la recopilación y al manejo de los datos que circulan por la red desde la ética investigadora, teniendo en cuenta que no hay soluciones únicas o definitivas y que, como escribía el poeta Antonio Machado, se hace camino al andar.

1. Lo bueno conocido: métodos tradicionales en el contexto digital

Una de las preguntas clave planteadas en el plano metodológico es hasta qué punto las propuestas desarrolladas para entender y describir la comunicación en la interacción cara a cara o en medios tradicionales pueden aplicarse al discurso digital. Si en algunos casos esto es posible, en otros, se deben adaptar o revisar los modelos existentes. En este sentido, Herring, pionera en la investigación de la llamada *Computer-Mediated Communication* (CMC), propone en 2007 un marco teórico-metodológico que se conoce como *Computer-Mediated Discourse Analysis* (CMDA) y que se apoya en gran medida en los supuestos del análisis del discurso, teniendo en cuenta los parámetros tecnológicos y sociosituacionales de la comunicación digital. También los planteamientos etnográficos de Androutsopoulos (2008, 2013) resultan útiles en este campo.

Sin embargo, a medida que el panorama del discurso digital se transforma y se vuelve cada vez más complejo parece necesario contar con una pluralidad de métodos y enfoques. Como señalan Jones *et al.*:

> In order to cope with the fast-changing landscape of digital media, discourse analysts need to both draw upon the rich store of theories and methods developed over the years for the analysis of 'analogue' discourse, and to formulate new concepts and new methodologies to address the unique combinations of affordances and constraints introduced by digital media (2015, 1).

Como se ha visto, el análisis de la comunicación digital recurre a distintas teorías en el seno de disciplinas como la lingüística, la etnografía, la sociología,

la sicología o la comunicación, se apoya en propuestas teórico-metodológicas aplicadas, por ejemplo, al estudio de los géneros textuales, de los actos de habla, de la (des)cortesía, entre otros, y en los métodos cualitativo y cuantitativo del análisis lingüístico y pragmático. Podría decirse que aprovecha lo «bueno» conocido, las teorías y modelos surgidos en el seno de la lingüística, la sociolingüística, el análisis del discurso, el análisis crítico del discurso, la pragmática, la pragmática sociocultural, la lingüística de corpus y, en los últimos años, la lingüística computacional, que permiten describir los mecanismos que rigen las prácticas comunicativas de los hablantes en todo tipo de contextos, incluido el digital. En este ámbito se asiste, pues, a un «alto grado de dispersión, fragmentación e inconsistencia teórico-metodológica» (Gallardo Paúls 2023, 146), que es habitual cuando se explora un nuevo objeto de estudio.

La gran variedad de situaciones comunicativas que se producen hoy en la red obliga a considerar en la investigación los parámetros tecnológicos y sociosituacionales asociados a cada contexto, asumiendo, por otra parte, que la noción de «contexto» aplicada a los entornos en línea es problemática y que debe abordarse desde una perspectiva polifocal (Parini 2023, 396). En los últimos años, a raíz del creciente interés por la comunicación en redes y a través del móvil, la mayor parte de los estudios asume que el discurso digital es una forma de (inter)acción social. Las preguntas que suelen guíar este tipo de estudios son: ¿cómo se estructuran estos discursos?, ¿cómo se logran las acciones sociales?, ¿cómo se negocian las identidades y cómo se construyen las ideologías?

Al mismo tiempo, suelen adoptarse metodologías propias del estudio de géneros (López Alonso y Séré 2006), del análisis de la conversación (Sanmartín 2007; Vela Delfa y Cantamutto 2016), la lingüística de corte variacionista (Estrada Arráez y De Benito Moreno 2016), la pragmática (Mancera Rueda y Pano Alamán 2020) o la pragmática sociocultural (Fuentes Rodríguez y Placencia 2014).

Como se ha planteado en las páginas previas, las categorías *macrogénero, constelación de géneros, tradición discursiva, coherencia, cohesión, marcador, conector, alternancia de turnos, par adyacente, registro, variación, relevancia, implicatura, atenuación, intensificación, acto cortés* o *actividad de imagen* son constructos teóricos útiles en el análisis de la interacción digital. Aplicadas al análisis de la comunicación escrita y de la interacción hablada han sido replicadas en este ámbito mediante procesos más o menos explícitos de operacionalización. Sin embargo, otras categorías han mostrado que el modelo puede ser limitado si se aplica a los entornos digitales. La replicabilidad, esto es, el

empleo de las mismas categorías para analizar diferentes datos, ha demostrado que puede ser funcional si estas se adaptan a los contextos de uso (Pano Alamán y Mancera Rueda 2014). Además, la comparación con situaciones comunicativas anteriores a la comunicación digital resulta productiva en la medida en que permite identificar modelos teórico-metodológicos fuertes o débiles de cara al análisis del nuevo paradigma comunicativo. Este es el caso, por ejemplo, de la posibilidad de emplear las unidades de la conversación para entender y describir las interacciones que se producen en los entornos digitales.

En resumen, para explorar adecuadamente los procesos y dinámicas de formación y transformación de las modalidades de comunicación digital, parece necesario adoptar modelos descriptivos y explicativos que no deben ser necesariamente nuevos pero que pueden permitir elaborar un metalenguaje apropiado de las estrategias comunicativas presentes en este tipo de intercambios, así como formalizar sus constantes lingüísticas en tipos de discursos concretos.

1.1. A vueltas con la conversación

Como apuntaba en el primer capítulo, una de las cuestiones más debatidas en este campo es si es posible hablar de conversación cuando interactuamos con otros en los comentarios de Facebook o de una noticia en *El País*, en Instagram, WhatsApp o TikTok. Los primeros estudios sobre los chats, en los que la interacción era síncrona, aplicaron las categorías asentadas en el análisis de la conversación oral coloquial para describir tanto la manera de gestionar los turnos de palabra como las manifestaciones de la oralidad en la escritura (Sanmartín Sáez 2007; Pano Alamán 2008). Como hemos visto, los enfoques pragmáticos adoptados en numerosas investigaciones asumen que la conversación en este contexto es un comportamiento comunicativo y social que resulta de la negociación entre dos o más interlocutores a través de principios como el de cooperación (Grice 1975), relevancia (Sperber y Wilson 1995) y cortesía (Brown y Levinson 1987), que bien pueden aplicarse al análisis del discurso digital.

En todo caso, ¿se pueden adoptar las nociones empleadas habitualmente en el análisis conversacional para describir este tipo de situaciones comunicativas? En general, como advierte Jucker (2021), el análisis de la comunicación digital se apoya a menudo en los conceptos del análisis de la conversación. Por ejemplo, Meredith (2019) considera los turnos, la organización de secuencias, la

reparación y las aperturas y cierres, interrogándose sobre su posible replicabilidad en la conversación digital. En el ámbito del español, Alcántara-Plá (2014) propone una reflexión similar para identificar las unidades del discurso en WhatsApp. Entre otras cosas, señala que los participantes en las interacciones en línea no tienen la misma posibilidad de seguir el progreso de una unidad constitutiva de turno ni pueden predecir cúando acaba un turno o empieza otro. Las reparaciones y las secuencias de apertura difieren también de las formas que adoptan en las conversaciones cara a cara, debido no solo a la interfaz de la plataforma o de la aplicación que se emplea sino también a la percepción que los hablantes tienen de la situación comunicativa (Cantamutto y Vela Delfa 2020).

Un concepto empleado en el análisis de las redes sociales virtuales, es el de «hilos de escritura»[1], que se han definido como «conversaciones multifuente –de varios usuarios– que no necesariamente comparten tiempo ni momento de incorporación y que pueden utilizar varios entornos en una misma charla» (López Sobejano 2012, 170). Efectivamente, un usuario puede incorporarse a una «conversación» iniciada en una determinada plataforma como reacción a un mensaje ya publicado o como intervención espontánea sobre el tema que se está debatiendo en esa u otra plataforma, abriendo así una nueva ramificación temática, un nuevo hilo. Asimismo, los hilos son «multientorno», en la medida en que pueden iniciarse en una plataforma y acabar en otra, y «multiherramienta», puesto que facilitan el empleo de distintos modos de comunicación, como la escritura, las imágenes, los vídeos y otros dispositivos hipertextuales.

Otros desafíos que se plantean en el «análisis de la conversación» en entornos digitales tienen que ver con la diversidad de aplicaciones y plataformas existentes, un aspecto que dificulta las generalizaciones, y con los múltiples niveles analíticos del comportamiento comunicativo y social de los hablantes en línea y fuera de línea, dos contextos que están estrechamente interrelacionados (Parini y Yus Ramos 2023).

Por ejemplo, en los últimos años se ha asistido a un aumento vertiginoso del uso de sistemas de videoconferencia, como Skype, FaceTime, MSTeams, Zoom, Google Meet, Webex, Collaborate, por mencionar los más utilizados. Al mismo tiempo, los sistemas de comunicación basados en texto escrito se

[1] Este concepto, aplicado en general a cualquier intercambio de mensajes en un foro o en una red social, no debe confundirse con el de *hilo* utilizado en Twitter/X. En el microblog, designa un conjunto de tuits publicados en secuencia por un mismo usuario para desarrollar un tema. Generalmente, el autor o la autora del hilo advierten a los usuarios de que su mensaje está constituido por más de un tuit, mediante la fórmula «va hilo» o «hilo dentro».

han expandido, tanto en plataformas de redes como en sistemas de mensajería instantánea, donde impera de hecho la comunicación multimodal, de uno a uno, de uno a muchos o incluso de muchos a muchos, como se decía. Esto plantea una serie de desafíos metodológicos al análisis de la interacción, por ello cabe plantearse en qué medida los conceptos y categorías del análisis conversacional pueden aplicarse en este campo o si es necesario adaptarlos.

El modelo de microanálisis de datos en línea, o MOOD (*Microanalysis of Online Data*), que proponen Giles *et al.* (2015, 46) para estudiar las interacciones en línea desde la perspectiva conversacional, aborda, en primer lugar, la naturaleza de los datos y la manera de acceder a ellos. Mientras que los analistas de la conversación exploran qué aspectos no verbales pueden integrar o enriquecer su análisis –prosodia, gestos, expresiones faciales, mirada fija–, en este ámbito es necesario identificar qué aspectos multimodales son importantes: el contexto visual, el diseño gráfico de una interfaz, los emojis, los *hashtags* o los vídeos breves. El modelo MOOD sugiere atender a la sincronía o asincronía de las plataformas y aplicaciones que se emplean, sin embargo, como se apuntaba en el primer capítulo, esta distinción ha dejado de ser operativa en el análisis de la transmisión y recepción de los mensajes, puesto la comunicación digital se basa en conexiones intermitentes. En todo caso, como apuntan Giles *et al.* (2015), es importante considerar las posibilidades técnicas de cada entorno, evitando la mera descripción de sus posibilidades técnicas. Por ejemplo, se puede observar si restringen la duración de las conversaciones o cómo muestran cada contribución en la pantalla de ordenador o de móvil, o determinar si la cronología de los mensajes incide en la gestión de la interacción. Cuando visualiza un «hilo» de comentarios a una publicación en Facebook en su propio perfil, el usuario puede reaccionar a cualquiera de los comentarios, independientemente de si es reciente o antiguo en ese hilo. En WhatsApp, en cambio, hay que tener en cuenta que cada usuario tiene en su dispositivo una versión personalizada. En este sentido, en la investigación que se lleva a cabo es necesario describir el contexto en el que se produce cada interacción[2].

[2] Como señala Yus Ramos (2001, 92-96), las interrupciones en el chat podían darse «debido a una falta de sincronización en los turnos de habla por razones de índole tecnológica. [...] Es posible recibir un mensaje antes de que se haya terminado de escribir el mensaje en la otra área». Para mitigar el impacto negativo de la multiplicidad de interacciones que se producían en ese medio era necesario incluir el apodo o el nombre del interlocutor al que iba dirigido el mensaje o utilizar mecanismos de retroalimentación recurriendo a la repetición, como se ha ilustrado en el ejemplo (21). Esto se observa aún en algunas redes, en foros y WhatsApp.

Volviendo a la estructura dialógica de las interacciones digitales, es también fundamental revisar las nociones de turno o alternancia de turno*s* antes de aplicarlas al estudio de la comunicación en una determinada plataforma. En la interacción cara a cara, la sincronía determina generalmente la alternacia de turnos y la organización de los turnos verbales; la alternancia se relaciona, además, con las pausas, la elección del hablante sucesivo y los posibles solapamientos entre intervenciones. Para delimitar la alternancia de los turnos en una conversación cara a cara se ha recurrido a los puntos de relevancia transicional o de transición, que se basan en una serie de indicadores lingüísticos como la estructura sintáctica o la entonación.

Como apunta Briz Gómez (1998, 52), existen, en función de su organización, dos tipos de conversación: estructurada y preparada previamente, y 2) improvisada y no planificada de antemano. En ambas, la estructura se articula en torno a la apertura, orientación, objetivo, conclusión y cierre. Los interlocutores se orientan en función de la posibilidad y relevancia de estos cinco momentos interactivos, que pueden aplicarse solo a algunos tipos de interacción digital. Por ejemplo, como explica Vela Delfa (2021), en el correo electrónico el discurso se articula en torno a una secuencia de apertura que puede contener fecha, preinicio, ritual de saludo y eventualmente excusas (*e.g.* por el retraso en responder); una secuencia central de carácter informativo; una secuencia de cierre o precierre (anuncio de cierre), cierre (generalmente, despedida y firma) y, a veces, poscierre (posdata).

Al contrario, en WhatsApp, donde discurren conversaciones ininterrumpidas que pueden durar incluso varios días, se asiste a una interacción no planificada o menos planificada que no presenta este tipo de secuencias o no todas. Un mensaje, considerado como una intervención, puede ser continuo o discontinuo respecto a un mensaje previo, y si bien es posible identificar la orientación y el objetivo de ese mensaje, no siempre se puede saber si constituye el inicio o el cierre en aquellas interacciones que se producen entre dos o más interlocutores de forma frecuente (Alcántara-Plá 2014).

Por otro lado, si atendemos a las opciones de visualización por defecto de los mensajes en WhatsApp, vemos que cada vez que uno de nuestros contactos envía uno o varios mensajes seguidos de texto, aparece una suerte de bocadillo con un fondo generalmente blanco a la izquierda de la pantalla; nuestro mensaje de respuesta o reacción a ese primer mensaje presenta la misma forma, aunque tiene un fondo distinto y se sitúa a la derecha. De este modo podemos distinguir nuestras intervenciones de las de nuestro interlocutor o nuestros interlocutores, lo cual permite operar distinciones entre los participantes en el acto comunicativo y clasificar más fácilmente los mensajes en función de la

orientación, del objetivo y de las intervenciones iniciativa o reactiva que se publican. En el caso de los grupos, la aplicación permite hoy en día identificar al autor de cada mensaje mediante no solo el nombre o el número de teléfono del contacto sino también la inserción del avatar o fotografía que le identifica en ella en el borde superior izquierdo de cada bocadillo. Esto ocurre también en las redes sociales, donde las interfaces de nuestros perfiles pueden personalizarse. No obstante, las intervenciones iniciativas –*tiktoks*, publicaciones, post o vídeos– no siempre se distinguen formalmente de las intervenciones reactivas, que suelen aparecer en el espacio de nuestros perfiles de forma secuencial y con el mismo formato.

Asimismo, aunque es posible identificar mensajes de apertura, de orientación hacia uno o más usuarios, o bien de cierre, es habitual que en las aplicaciones y plataformas 2.0 se acumulen mensajes de un mismo autor dirigidos a distintos interlocutores y con objetivos diversos. De ahí que el concepto de turno de habla tampoco pueda aplicarse directamente a todas las interacciones digitales sin revisar primero el tipo de interacción que facilita cada plataforma y la manera como los usuarios perciben y se apropian de estos espacios. Por ejemplo, aunque es posible que puedan darse pares de adyacencia[3] en aplicaciones de mensajería instantánea o de correo electrónico, es menos probable que aparezcan en espacios como Reddit, en comentarios en la prensa, en Twitter/X, YouTube, Instagram o Facebook, donde un primer mensaje considerado como turno iniciativo puede vehicular distintos actos de habla. Los comentarios en la prensa digital dirigidos a otro comentario pueden contener actos asertivos, directivos y expresivos que admiten distintos tipos de reacciones; además, un comentario que contiene un acto perlocutivo, una pregunta o una provocación, no siempre recibe una respuesta por parte de otros usuarios. Los comentarios se publican, además, en secuencia bajo la noticia en línea y pueden bien dirigirse al autor del primer mensaje, bien a los autores de mensajes precedentes en la cadena de comentarios, lo cual complica la identificación de pares y de turnos (Pano Alamán 2012, 2015).

En todo caso, como se ha mostrado anteriormente, cualquier mensaje en redes o en otro tipo de entornos digitales puede obtener respuestas y reacciones directas de diverso tipo. Los dispositivos multimodales empleados y las

[3] De acuerdo con el análisis de la conversación, un par adyacente o de adyacencia está formado por dos turnos conversacionales consecutivos; el primer turno prevé que el segundo contenga una respuesta o reacción determinada. Algunos de los pares más frecuentes son del tipo: saludo-saludo; pregunta-respuesta; aserción-acuerdo/desacuerdo.

estrategias lingüísticas que adoptan los usuarios permiten observar cómo los participantes se orientan pragmáticamente hacia uno u otro contenido o hacia uno o más interlocutores. Cabe asumir, por tanto, que la gestión de lo que podemos llamar por comodidad turnos de habla está determinada en parte por el sistema, en parte por los mismos interactuantes. Estos se adaptan a los dispositivos que ofrece la plataforma y a los condicionantes sociocomunicativos propios de este tipo de interacciones. En este sentido, es posible hablar de «recontextualización interdiscursiva», entendida como:

> the dynamic transferand-transformation of something from one discourse/text-incontext (the context being in reality a matrix or field of contexts) to another. Recontextualization involves the extrication of some part or aspect from a text or discourse, or from a genre of texts or discourses, and the fitting of this part or aspect into another context (another text or discourse or discourse genre) and its use and environment (Linell 1998, 154).

Los recursos lingüísticos y semióticos de que disponemos como hablantes y usuarios de las tecnologías se recontextualizan en un nuevo contexto comunicativo, en este caso, la interacción mediada, añadiéndoles nuevas funciones o bien reemplazándolos por otros de forma innovadora. Cuando nos enfrentamos a nuevas situaciones comunicativas adoptamos soluciones ingeniosas para que el proceso funcione, mediante la adaptación y la modificación de nuestras prácticas de acuerdo con las posibilidades de las innovaciones técnicas. Al fin y al cabo, lo que importa es «dialogar» de manera continua o discontinua. Como planteaba en Pano Alamán (2008), es posible entender las interacciones que se llevan a cabo en la red a través del concepto de «dialogismo», en referencia a un tipo de interacción entre personas a través de medios simbólicos en la que se integran tanto el diálogo como el monólogo. Desde esta perspectiva es posible asumir que cualquier acción comunicativa, como interacción social, está orientada mutuamente hacia el otro, y que la orientación hacia el otro existe aunque este no esté copresente (Linell 1998, 35).

El diálogo es la forma predominante de interacción en la red[4]. Un individuo actualiza su estado y comenta el de otro, escribe una entrada en el propio blog y glosa la entrada en un blog ajeno, carga un vídeo o una fotografía para suscitar comentarios y reacciones del público. En todos los casos, se busca fomentar la interacción en torno a un texto, un vídeo o presentación, de modo

[4] Hine (2000) hablaba hace años de «web dialógica».

que se produzca un intercambio de puntos de vista, una suma de opiniones, que nace de la amplia participación de las personas en la red, especialmente desde los móviles. Las interfaces de las plataformas y aplicaciones que empleamos permiten activar la visualización de los mensajes, donde es posible identificar inmediatamente el mensaje inicial y la o las respuestas que recibe (Fig. 3). Percibimos así que participamos en espacios en los que predomina el intercambio breve.

Desde este enfoque, cabe considerar también el alcance de la mutualidad de los entornos cognitivos que son manifiestos en estos contextos (Yus Ramos 2001, 2010). Los supuestos manifiestos pueden inferirse a partir de un determinado tipo de información transmitida de forma textual o visual, del orden en que aparece la información, del uso de determinadas estructuras y de la selección léxica. Los temas, intereses y experiencias más o menos compartidos se actualizan, se modifican, se amplían o se reducen a partir del tipo de interacción en el que se participa, del contexto privado o público de la plataforma o aplicación, del objetivo comunicativo perseguido, del número variable y percibido de los participantes y de la intención comunicativa de cada uno de ellos, de acuerdo con lo que Gibbs (2001) llama an *emergent-interactive view of intention*.

Los análisis de la comunicación digital en español han avanzado en esta dirección para entender cómo el grado de conocimiento previo entre los participantes en un foro de debate o en Twitter/X o el grado en el que se comparten intereses en Facebook, Instagram, Twitch o TikTok explican la presencia de heterografías en el texto escrito o bien el empleo de fórmulas indirectas que permiten expresar un cumplido o criticar irónicamente al interlocutor.

A partir de los conceptos de dialogismo y de recontextualización discursiva, es posible aplicar algunos conceptos y categorías del análisis de la conversación al estudio de la interacción en la red. En la conversación oral coloquial, la intervención es «cada una de las emisiones de un hablante, esto es, un enunciado o conjunto de enunciados (acto o actos de habla) emitidos por un interlocutor de forma continua o discontinua y vinculados por una estrategia única de acción e intención» (Briz Gómez 2000, 228). La realización de dos intervenciones sucesivas, de inicio y reacción, da lugar a un intercambio, unidad dialógica marcada por el cambio de hablante.

El mensaje enviado a una plataforma como unidad comunicativa cerrada o completa (tuit, publicación, comentario, entrada, etc.) constituye una primera intervención. Esta puede ser de inicio o iniciativa, cuando el mensaje busca

provocar una reacción en forma de respuesta verbal o a través del clic en *me gusta*, *retuitear, comentar, compartir* o señalar como favorito. El mensaje de reacción puede ser también la respuesta verbal a una pregunta, la expresión de acuerdo o desacuerdo con lo dicho en la intervención previa o la formulación de un cumplido o una crítica a través de texto o de contenido multimodal. También pueden considerarse intervenciones reactivas aquellas señales manifiestas que dejan una huella digital en la plataforma en la que se interactúa mediante el número de estrellas en un sitio de reseñas, de citas con o sin comentario, de comentarios y de respuestas, etc.

Se ha dicho que la compleja interfaz de las plataformas de redes sociales, de los foros de debate, de los chats o de la mensajería instantánea, y la acumulación en poco tiempo y en cadena de mensajes que reaccionan de distintas maneras a un mensaje inicial, un vídeo o un meme, dificultan la identificación de turnos, la visualización de las intervenciones que se elaboran en torno a un mismo tema o el seguimiento de hilos coherentes, favoreciendo la fragmentación temática y disminuyendo la focalización colectiva. Esto hace que, desde el punto de vista del análisis, sea complejo identificar las intervenciones iniciativas y reactivas y los intercambios. Sin embargo, los sistemas actuales de gestión de mensajes permiten visualizar las categorías de debate en forma de listado y las intervenciones de modo secuencial y organizado, facilitando la aplicación de las unidades de análisis que se acaban de describir. La interfaz de Facebook, de Twitter/X, de los foros o de WhatsApp contiene un dispositivo para «citar» el mensaje al que se quiere dar respuesta o que se pretende comentar, de manera que el propio mensaje contiene ya el texto con el que se relaciona temáticamente, la intervención iniciativa que produce una reacción. Este aspecto ha sido descrito también respecto a los comentarios a las noticias, que pueden visualizarse de forma «llana» cuando se publican con criterio cronológico o «anidada» (Sal Paz 2012), forma que, en los medios en los que se pueden comentar las noticias, se asocia al modo conversación[5] ilustrado en la Figura 3 en el primer capítulo. El sistema vincula unos mensajes a otros en forma de réplicas y contrarréplicas. De forma similar, en el correo electrónico nuestra «respuesta» a un mensaje previo integra automáticamente ese mismo mensaje a modo de cita.

Por otro lado, lo primero que se nota al analizar los mensajes que se intercambian en las redes es que, a pesar de que muchos se dirigen a uno o más

[5] Por ejemplo, en los comentarios de *ElPais.es* es posible visualizar los comentarios en este modo, lo cual permite acceder a los mensajes como intervenciones iniciativas y reactivas.

destinatarios, no es necesario que estos contesten; y si una intervención provoca o busca provocar una reacción por parte de uno de los contactos de la red, esta no tiene por qué ser inmediata. La interacción es además, en muchos casos, pública, por ejemplo en los comentarios a una noticia o a un vídeo, el espacio del chat en Twitch, un muro o una cronología. Caso distinto es el de las aplicaciones en las que predomina la comunicación privada entre dos interlocutores o entre varios dentro de un grupo cerrado, como en el correo electrónico o WhatsApp, o en redes como Instagram o Facebook, en las que es posible chatear privadamente con otro usuario. Estos aspectos obligan a adoptar determinadas estrategias comunicativas al servicio de la coherencia textual. En particular, el uso del vocativo, como se señalaba en el primer capítulo, o el discurso referido, que permite retomar el mensaje de otro usuario para elaborar la propia contribución al intercambio. Véase (28), ejemplo extraído de un corpus de mensajes en Facebook (Mancera Rueda y Pano Alamán 2013a, 67), donde la cooperación entre interlocutores y la retroalimentación acercan el intercambio a la conversación coloquial:

(28) A1: «El 25 de abril, estaré en el Liceo de Málaga, participando en el ciclo de narradores.Será el jueves alas 7 y media de la tarde, nos vemos y nos 'oímos'»
B1: te deseo lo mejor Amiga!!
A2: Será genial, gracias nombredeusuario, un abrazo
B2: después nos compartes, si?..beso*)
A3: Jjajja, claro..
B3: buen domingo!!!!!!!!!!!!!!
A4: Lo mismo para ti cielo :))
C1: Suerte nombredeusuario,un beso desde los Alisios.
D1: COMO ME GUSTARÍA ESTAR ALLÍ, PARA VERTE Y DISFRUTAR DE MI CIUDAD, TE DESEO LO MEJOR DE LO MEJOR, UN BESO.
A5: Me encantaría a mi también nombredeusuario, disfrutaríamos juntas de Málaga,mi madre es malagueña. Estando allí estoy en casa, un abrazo.
A6: gracias, besosssssssss (Facebook, 07-04-2013).

El primer hablante abre la secuencia anunciando, por medio de una publicación en Facebook, su participación en un encuentro de escritores. Algunos de sus contactos en la red reaccionan a esa primera intervención, bien

para felicitarla por la iniciativa y animarle, bien para pedirle otro tipo de informaciones. Se observa en este caso un deseo de interactuar con los participantes y un tenor funcional comunicativo al que se añaden determinados rasgos situacionales que acercan el discurso al registro coloquial. En estas intervenciones destacan, por ejemplo, los agradecimientos, el uso de apelativos cariñosos (*cielo, amiga*) propios del español peninsular en función de vocativo, y las demostraciones de afecto (*un abrazo* o *besos*) que expresan cercanía y que denotan una relación vivencial de proximidad que explica en gran parte la coloquialización de los enunciados a nivel ortográfico, morfosintáctico y léxico: véanse también la reproducción gráfica de la risa (*Jjajja*), el alargamiento de la *s* (*besosssssss*), el recurso a las mayúsculas para enfatizar lo que se dice (D1), el emoticono (B2, A4), la desaparición de los signos de exclamación o de interrogación iniciales (B1, B3), o la contigüedad de oraciones (A5).

La interacción se estructura en microintercambios, la primera intervención exhorta al interlocutor a compartir algo, por ejemplo, un vídeo o una fotografía, sobre ese encuentro: «después nos compartes, si?», lo que lleva a la hablante inicial a contestar afirmativamente por medio del marcador *claro*. En otra intervención, se desea que el interlocutor pase un buen domingo, a lo que el otro responde: «Lo mismo para ti». En este tipo de intercambios mínimos, frecuentes en entornos digitales, los hablantes suelen mostrarse cooperativos, ya que asumen las intervenciones de otros participantes para elaborar sus mensajes, haciendo progresar la conversación en un tiempo indefinido.

Un aspecto común a estos intercambios es que la mayor parte de los participantes tiene como objetivo sumarse al grupo, participar en un espacio en el que distintas voces se superponen en torno a un mismo tema. Lo demuestra la acumulación de cumplidos a una misma fotografía en Instagram y de tuit*s* irónicos o humorísticos publicados en torno a un meme o una etiqueta en Twitter/X. En estas ocasiones, se comenta sobre un mismo tópico o sobre nuevos tópicos, sin necesidad de interactuar con los demás, simplemente añadiendo el propio grano de arena en un espacio y tiempo breves y ante un auditorio variable, el de los propios seguidores y el de los seguidores de estos. A través de este tipo de enunciados, aislados o monológicos, cada uno aporta su punto de vista sin tener en cuenta el mensaje que antecede en la secuencia.

Este es, pues, un discurso colectivo en el que los enunciados se acumulan a partir de una primera intervención, texto, imagen, vídeo, meme, y en el que se mezclan modos y registros diversos (Mancera Rueda y Pano Alamán 2013a,

26). De hecho, la polifonía y la heteroglosia juegan un papel fundamental en la comunicación digital.

1.2. Paisajes lingüísticos virtuales: variación y contacto en las redes

En su blog *Se me va de la lengua*, Carlota De Benito Moreno publicaba en julio de 2020 la entrada *Tuits plantilla: las tradiciones discursivas tuiteras*[6], donde se afirma que:

> Hay dos cosas que hacen especialmente interesante a Twitter. Por un lado, es una comunidad de habla (o de escritura) muy grande y abierta […]. Por otro, es una red social de vocación fundamentalmente conversacional y es en la conversación, en el coloquio, donde los hablantes solemos ser más creativos. Sin embargo, esto es solo cierto de las conversaciones informales: en Twitter entran en contacto a distancia y a través de la escritura personas que jamás se han visto, lo que no parece el paradigma de contexto informal. Sin embargo, las características lingüísticas de gran parte de lo que se escribe en Twitter son propias del habla coloquial […]. Así, Twitter se ha convertido en un sitio en el que se crean innovaciones lingüísticas continuamente.

En la entrada, la investigadora discurría sobre las palabras y expresiones coloquiales utilizadas de manera frecuente en el microblog y que había ido recogiendo en una plantilla, gracias a la colaboración de los usuarios, para mostrar, precisamente, la gran creatividad lingüística de los hablantes (v. también McCulloch 2019). En algunas investigaciones (Estrada Arráez y De Benito Moreno 2017; De Benito Moreno 2022), da cuenta de la utilidad y el interés que presenta el uso de las redes sociales o, en general, las interacciones digitales, para explorar la lengua desde una perspectiva variacionista[7].

De Benito Moreno y Estrada Arráez (2018) llevan a cabo una comparación entre dos corpus, uno general que contiene distintos tipos de textos digitales recogidos en el corpus esTenTen[8] y un corpus de tuits, a partir del análisis de

6 http://www.semevadelalengua.es/?p=655

7 Androutsopoulos (2011) señalaba previamente que el concepto clave en torno al cual se movía buena parte de los estudios dedicados al discurso digital, sobre todo en las plataformas 2.0, era el de variación.

8 El corpus web en español (esTenTen) es un corpus de texto creado a partir de los textos recopilados de Internet. Pertenece a la familia de corpus TenTen, conjunto de corpus web

la elisión de la /d/ intervocálica, las formas de tratamiento en plural, el uso no referencial de *ello*, la pluralización de *haber* existencial y el uso coloquial del sufijo superlativo *-érrimo*. Los resultados muestran que, si bien un macrocorpus de comunicación digital como esTenTen, compilado de forma automática, ofrece una gran cantidad de datos útiles para analizar fenómenos de variación léxica o morfosintáctica, este tiene la desventaja de incluir ruido estadístico[9]. Además, al recopilar los datos generados en «entornos digitales monológicos» (Pano Alamán 2008), como las páginas web o los blogs, esTenTen resulta limitado para documentar fenómenos marcados diatópica, diafásica o diastráticamente. Aquellos corpus controlados en los que se privilegian contextos prevalentemente dialógicos, como las redes sociales, aunque resultan más difíciles de compilar, son preferibles para estudiar las variables lingüísticas propias de la inmediatez comunicativa.

La investigación en torno a la variación lingüística con datos de interacciones digitales convive hoy con los estudios lingüísticos y pragmáticos sobre el discurso digital que se han presentado en los capítulos previos. No obstante, no es fácil distinguir estas dos perspectivas. En los estudios variacionistas, se adopta el sintagma *lengua en la red* en referencia a los datos lingüísticos recogidos en grandes corpus para documentar cómo evoluciona el español en microdiacronía, generalmente, en los planos léxico o gramatical. En este sentido, las redes sociales como Twitter/X o Facebook son una fuente válida para analizar fenómenos sintácticos o léxicos difícilmente documentables a través de otro tipo de recursos. La ventaja de los datos que Estrada Arráez y De Benito Moreno (2016) llaman *twilectales* es que son numerosos y en gran parte accesibles. Con el objeto de mostrar la validez de este tipo de datos, las investigadoras plantean dos casos de estudio: el uso adverbial de *puto*, «fenómeno muy coloquial y principalmente extendido entre los hablantes jóvenes» (p. 77) y la extensión de la forma *se* a otras personas en el paradigma reflexivo, un fenómeno que, de acuerdo con su análisis, se restringe a determinadas variedades diatópicas.

en 40 idiomas, procesados con un tamaño objetivo de más de 10 000 millones de palabras, a los que es posible acceder a través de la plataforma *Sketch Engine*. EsTenTen contiene subcorpus basados en las variedades de español europeo y español americano con datos extraídos de dominios web europeos y americanos. En la base de datos de Sketch Engine, se accede a una lista de los corpus de TenTen en español disponibles, como esTenTen18 (16900 millones de palabras en español europeo, español estadounidense y Wikipedia en español); y esTenTen11 (9500 millones de palabras web en español europeo, en español americano, parte de Wikipedia en español). Información disponible en: https://www.sketchengine.eu/estenten-spanish-corpus/

9 Conjunto de términos no pertinentes o palabras no significativas.

Estudios similares centrados en Twitter/X como fuente de datos lingüísticos, se han ocupado, en cambio, de la evolución de las expresiones malsonantes de contenido procedimental (Fuentes Rodríguez 2022) y de fórmulas más o menos fijadas en la red y no solo como «Yo + cuando + *GIF*/imagen» o «Acompáñenme en esta historia» (Calò 2022). Si bien estos datos permiten documentar fenómenos anti o heteronormativos de la lengua, cabe considerar que las funciones no estándar en la comunicación digital no siempre se correlacionan con su uso en el habla (De Benito Moreno 2021), especialmente en la conversación oral cara a cara, sino que pueden tener en esos contextos funciones como la de jugar con la propia identidad virtual a través, por ejemplo, de la manipulación creativa con propósitos lúdicos o de refuerzo de las relaciones intragrupales.

Un segundo aspecto interesante de la perspectiva sociolingüística aplicada al estudio de la comunicación digital es ver cómo en estos espacios conviven no solo diferentes registros y variedades de una lengua, sino también diferentes lenguas que, en algunos casos, se convierten en objeto de debate en las interacciones digitales. Diversos análisis se han ocupado de las llamadas narrativas migrantes, así como del contacto de lenguas en las redes sociales (Garcés-Conejos Blitvich y Bou-Franch 2014; Mapelli 2019), poniendo en evidencia la importancia de estas plataformas como espacios de mantenimiento lingüístico en los que es habitual alternar o mezclar códigos en función de las relaciones sociales entre hablantes y de los objetivos comunicativos de cada participante. En estos casos, además de observar las prácticas lingüísticas plurilingües que adoptan los usuarios, es interesante explorar las ideologías lingüísticas que vehiculan dichas prácticas y las creencias y actitudes que se manifiestan en los mensajes.

No en vano, los medios digitales fomentan el desarrollo de comunidades de habla en línea que incorporan no solo prácticas comunicativas sino también normas y creencias compartidas en relación con esas prácticas. Se ha dicho, por ejemplo, que los comentarios en YouTube constituyen una instancia virtual del paisaje lingüístico en que es posible investigar distintos aspectos sobre las actitudes y valoraciones de los hablantes hacia su idioma o hacia otros (Ivković y Lotherington 2009), que se expresan de forma espontánea en los comentarios[10].

En resumen, los grupos sociales negocian sus prácticas lingüísticas en línea y crean sus propias políticas lingüísticas en línea, un aspecto particularmente

[10] Por ejemplo, Ivković (2013) ha indagado sobre las creencias y actitudes de quienes comentan los vídeos sobre Eurovisión en relación con las lenguas y variedades que se emplean en las canciones.

interesante en interacciones como las que tienen lugar en YouTube. Como apuntaba en otro lugar (Pano Alamán 2016), esta plataforma, además de proporcionar dispositivos tecnológicos que fomentan la participación y el debate sobre vídeos con distintos contenidos, ofrecen información muy útil, a través de los comentarios, sobre las actitudes que pueden tener los usuarios hacia las lenguas y las variedades lingüísticas que se emplean en los vídeos. En ese estudio, se investigan las actitudes lingüísticas hacia el español en los Estados Unidos a través del análisis de un corpus de comentarios publicados en los canales Univisión y Telemundo en YouTube, en respuesta a vídeos (noticias, entrevistas) centrados en el uso del español en ese país. De acuerdo con Garcés-Conejos Blitvich y Bou-Franch (2014, 439), YouTube es «un espacio idóneo para el estudio de la identidad social, ya sean identidades híbridas que reflejan la tensión entre el país de origen y el de acogida, ya sean identidades dobles que se negocian y (re)crean en el discurso». Dependiendo de la interacción en la que participe, el hablante modificará su actitud frente a la lengua o la variedad de lengua sobre la que se centra la interacción. En este caso, cabe, pues, atender al contenido de los vídeos seleccionados que se comentan, y al tipo de interacciones (intercambios dialógicos) que pueden darse entre quienes comentan. Para ello cabe adoptar un enfoque que considera la relación dinámica entre los contenidos del vídeo, el título o el breve texto de acompañamiento del mismo, si aparece, y la interrelación entre estos elementos.

1.3. Más multimodalidad, por favor: la urgencia de analizar los modos

Hoy en día no es posible ignorar que la comunicación digital se caracteriza en gran parte por la combinación de distintos códigos semióticos, cuyo propósito es construir textos comunicativamente coherentes (Duque 2020, 2021). Por este motivo, en esta sección se subraya la necesidad de prestar una atención mayor a la cohesión multimodal y a las relaciones discursivas que se establecen entre los contenidos visuales, orales, verbales o hipertextuales que predominan en los entornos digitales.

Como es sabido, el análisis multimodal es un método de análisis que tiene en cuenta el empleo de múltiples modos de comunicación, verbal, auditivo, visual, cromático, etc. La noción de «multimodalidad» hace referencia al conjunto de fuentes semióticas, o modos, que permiten generar un significado

socialmente modelado y mejorado culturalmente (Kress 2010). Este tipo de análisis se centra en cómo estos modos crean y transmiten significados interactuando entre sí, asumiendo que el significado siempre se sitúa en un contexto particular y que el uso de diferentes modos puede variar según el contexto y el propósito comunicativo. Asimismo, el análisis multimodal implica prestar atención a los detalles adoptando una perspectiva sistémica. En general, se trata de un método flexible y versátil que se puede aplicar a una amplia gama de contextos y medios y puede proporcionar información valiosa sobre cómo se transmite el significado en entornos comunicativos complejos y dinámicos[11].

Buena parte de las investigaciones que se llevan a cabo desde esta perspectiva se apoyan en la Teoría semiótica multimodal (Kress y van Leeuwen 2001; Kress 2000, 2010; Jewitt y Kress 2003) o en el Análisis del Discurso Multimodal (O'Halloran 2004). Sin embargo, la aplicación de este tipo de análisis presenta varios desafíos. Destacan, entre otros: la complejidad de analizar múltiples modos difíciles de interpretar; la subjetividad en la interpretación de características visuales, acústicas o gestuales sujetas a múltiples variables, entre otras, tecnológicas, sociales y culturales; la limitación o la complejidad de las herramientas y recursos disponibles para la recolección y análisis de los datos; las cuestiones éticas que conlleva todo tipo de investigación, especialmente si esta prevé el registro de voces o de imágenes de los participantes; por último, la ausencia de estandarización en los métodos y técnicas utilizados en el análisis multimodal, lo que puede dificultar la comparación o incluso la replicabilidad.

En todo caso, el modelo de análisis multimodal hace posible identificar unidades mínimas de significado de diferente naturaleza que interactúan de forma simultánea dentro de una misma entidad comunicativa, como puede una plataforma de reseñas, donde se combinan los modos textual, audiovisual, hipertextual y alfanumérico. Este método de investigación permite entender cómo los usuarios crean significados a través de la selección estratégica de los recursos semióticos más adecuados dentro del conjunto de modos disponibles en la plataforma. Para identificar esos significados, el análisis multimodal se centra en cómo se utilizan en un momento comunicativo concreto y en un entorno específico, en vez de intentar fijar un inventario universal para cada modo. Y es que solo el contexto situacional o cultural

[11] Véase Herring (2015) para una amplia revisión del concepto de «comunicación interactiva multimodal» en entornos digitales.

compartido determina la producción e interpretación óptima del mensaje (Yus Ramos 2010).

En el ámbito del discurso digital en español, el análisis multimodal es un enfoque que se está asentando de forma lenta. Como señala acertadamente Duque:

> A pesar del llamado giro multimodal en el análisis del discurso y de que la combinación de códigos semióticos es considerada una de las propiedades que definen al discurso digital, la multimodalidad en la comunicación digital en español ha recibido una más discreta atención que los temas anteriormente mencionados, al menos como objeto central de estudio (2020, 144).

Pueden citarse algunas investigaciones en las que se adopta explícitamente un enfoque multimodal en el estudio de la comunicación digital. La mayor parte se ocupa de la interacción entre contenido verbal y visual a través del análisis de sitios web, vídeos de YouTube, emojis, memes y *affordances*[12], que son dispositivos tecnológicos como el *hashtag* o el *me gusta* con los que es posible realizar ciertas acciones y cuyo significado no es inherente al dispositivo, sino que surge de las formas en que se usan e interpretan dentro de un contexto particular. Por ejemplo, López Pena (2021) señala la utilidad del Análisis del Discurso Multimodal (ADM) para explorar en detalle el significado de texto, imágenes y vídeos presentes en dos sitios web de promoción de la ciudad de Santiago de Compostela y comprender cómo las personas interpretan esos recursos semióticos en dicho contexto. De acuerdo con el investigador, estos espacios son:

> […] artefactos multimodales y multimedia creados en un contexto sociohistórico concreto dentro de un paisaje semiótico determinado que suele recopilar […] vídeos, imágenes, logotipos, texto o audios; al igual que distintos tipos de publicaciones digitales en las que se recoge un amplio abanico de géneros textuales tales como, entre otros, folletos, dosieres, informes o catálogos (López Pena 2021, 841).

[12] El concepto de *affordance* es fundamental en el campo del diseño; proviene de la teoría ecológica de la percepción visual de Gibson (1979) y hace referencia a las posibilidades de acción que ofrece un entorno comunicativo.

Asimismo, los modos presentes en los sitios web dedicados a la promoción turística tienen funciones informativas y persuasivas esenciales (Calvi 2010), lo que obliga a considerar las relaciones de significado que presentan en función de los objetivos comunicativos de la web en cuestión. El análisis de López Pena (2021) muestra que en los dos sitios predominan las imágenes sobre el texto, aunque la relación semiótica que se establece entre texto e imagen es coherente. El autor señala también que ambos utilizan una combinación de lenguaje informativo y persuasivo, mediante adjetivos calificativos y verbos sensoriales, o bien a través de imágenes de la ciudad pensadas para activar las emociones en el usuario y potencial viajero.

Distinto es el análisis de Pardo Abril (2008), quien explora la mediatización y la multimodalidad en el discurso de YouTube a propósito de la pobreza, a través de un caso de estudio, el vídeo-canción «La rutina», y centrándose en los modos visual, verbal y sonoro del vídeo. El estudio destaca por su atención a la metodología empleada. Como declara la autora, la propuesta «aspira a recuperar el conjunto de categorías y relaciones que pueden ser punto de partida para la formulación de un análisis sistemático de los discursos propios de un sitio como YouTube y, en general, de los discursos multimodales típicos de la Internet» (Pardo Abril 2008, 104). YouTube es un «portal» dentro de la Web 2.0, puesto que el usuario puede publicar y compartir vídeos propios –*Broadcast yourself* decía el lema de la compañía– o ajenos, evaluar si le gustan o no, visualizar contenidos relacionados con esos vídeos, enviarlos a cualquier otro usuario mediante un enlace y comentarlos, siempre y cuando disponga de una cuenta propia.

La reflexión de Pardo Abril (2008, 104) pone en relación las maneras como se perciben los significados en los textos de la producción multimedia que se publican en YouTube. Para el análisis se apoya en el programa ELAN 3.6.0 de anotación de texto y de imágenes, programa que facilita el reconocimiento, la categorización y la sistematización del corpus en los niveles verbal, visual y sonoro, y en sus distintas expresiones y recursos semióticos.

Como apuntaba en 2.5, la comunicación por medio de dispositivos móviles ha acelerado y amplificado el uso combinado en un mismo mensaje de textos, imágenes estáticas o dinámicas, vídeos, audios, enlaces y otros modos. En los estudios dedicados al análisis de los emojis en WhatsApp (Sampietro 2016a, 2016b, 2019; Noblía 2018; Cantamutto y Vela Delfa 2019; Vela Delfa y Cantamutto 2021), se adoptan enfoques multimodales que, de acuerdo con Sampietro (2016a, 286), demuestran que las funciones de los emoticones y de los emojis, como el pulgar hacia arriba, no son solo las de completar, clarificar o

desambiguar el mensaje verbal o vehicular ciertas emociones. En buena parte de estos estudios se observa cómo estos dispositivos acompañan, reforzando o atenuando, diferentes actos de habla y contribuyen a estructurar el discurso y a negociar las posiciones y los roles durante la interacción.

Por su parte, Duque (2020) indaga sobre las conexiones de tipo causal, de contraste o de adición que a menudo señalan los marcadores de discurso asumiendo que las relaciones desarrolladas para el modo verbal son aplicables a los «géneros multimodales»[13]. En particular, a partir de un corpus de memes, se centra en las relaciones de ampliación, que se manifiestan cuando la imagen de un meme añade información al contenido verbal estableciendo relaciones cohesivas de correferencia o identidad entre texto e imagen; de semejanza, cuando texto e imagen mantienen relaciones de oposición o de analogía, o cuando, en el plano enunciativo, establecen relaciones de tipo contrastivo, pudiendo dar lugar a interpretaciones irónicas e incluso humorísticas (Pano Alamán y Mancera Rueda 2023); por último, atiende a las relaciones de causalidad, por ejemplo, cuando las imágenes se conectan directamente con el contenido del mensaje, actuando como causas del decir (Duque 2020, 155-156).

Los enfoques multimodales se aplican prevalentemente al análisis de los modos visuales (vídeos, fotografías, emojis y memes), dejando de lado dispositivos hipertextuales como pueden ser las etiquetas o los enlaces. Se trata de mecanismos incrustados en enunciados verbales o colocados fuera del mensaje de forma aislada, que también establecen relaciones semióticas con otros elementos (verbales, visuales, hipertextuales) del mensaje y que desempeñan distintas funciones discursivas y pragmáticas, llevando en algunos tipos de discurso a modificar las prácticas comunicativas, como se ha mostrado en la sección 5.1 del segundo capítulo, dedicada a los *hashtags*.

El análisis de los modos hipertextuales (menciones, etiquetas, enlaces) y audiovisuales (emojis, vídeos, fotografías, pósteres, infografías) en un corpus de tuits de cinco partidos políticos españoles muestra cómo la combinación de esos modos transforma el mensaje político, que se adapta al espacio condensado y fragmentado del tuit (Pano Alamán 2019). Es evidente en ese contexto la tendencia a la atomización del discurso político, mediante el empleo de textos cada vez más breves combinados con distintos modos. El contenido

[13] García Asensio y Palomeque Kovacs (2012) hablan precisamente de *blog multimodal*, destacando la potencialidad comunicativa y de representación de la imagen en interacción con sonidos y texto.

verbal se resume en forma de titular que funciona a modo de gancho para que accedamos a los numerosos vídeos y documentos en los que, por otro lado, predomina la imagen del miembro del partido, como se observaba más arriba.

Se reserva también mayor espacio a los vídeos y a los emojis con función persuasiva. El texto, el contenido verbal, se convierte así, a menudo, en un pretexto para introducir otro tipo de contenido (audiovisual), al que se yuxtapone y hacia el que converge semántica y pragmáticamente, en un formato más inmediato y conforme a la micropantalla.

Por último, cabe recordar que nuestras interacciones en línea se producen generalmente, a veces de forma simultánea (Alcántara Plá 2014), en diferentes plataformas y entornos. Una de las cuestiones planteadas en las reflexiones metodológicas recogidas en Vásquez Roca (2022) se refiere a las semejanzas y diferencias que presentan cada uno de los entornos digitales desde un punto de vista tecnológico y comunicativo. Los estudios que analizan comparativamente la interacción en dos o más entornos digitales (Urra González 2007; Pano Alamán 2008, 2020; Cabedo Nebot 2009; López Quero 2010; Pérez Sabater 2011; Vivas Márquez y Ridao Rodrigo 2015; García Rivero *et al.* 2022) demuestran el interés de adoptar metodologías tanto multimodales como multiplataforma, combinando la cuantificación de ciertas características, por ejemplo, las formas de participación o determinadas prácticas discursivas, con un análisis cualitativo de ejemplos de interacciones en cada uno de esos contextos. En este sentido, los conceptos de «hibridación mediática» y de «remediación» permiten explicar por qué los límites del contexto comunicativo asociado a un medio específico se difuminan. Tampoco hay que olvider que, cuando nos ocupamos de interacciones en las redes o en otro tipo de entornos, estamos ante contextos comunicativos que cambian rápidamente.

Esto es particularmente evidente cuando asistimos a una multiplicación de plataformas cada vez más complejas y sofisticadas desde el punto de vista de su diseño y de las posibilidades comunicativas que ofrecen[14]. Dos ejemplos de esta tendencia son TikTok y Twitch, donde se combinan modos semióticos distintos con fórmulas inéditas de participación (García Rivero *et al.* 2022) como son, respectivamente, los *desafíos* y los *playthrough*, esto es, vídeos en directo en los que un videojugador juega ante su audiencia con el fin de provocar distintas reacciones en un canal de chat.

[14] Aunque no es posible detenerse aquí sobre los videojuegos, parece importante señalar la investigación de Pereira y Alonzo (2017), quienes abordan los videojuegos como discursos complejos.

1.4. Con un ojo crítico: riesgos de la interacción en redes

El apartado 2.5 consideraba el impacto de los algoritmos en nuestros comportamientos de consumo y en nuestras prácticas comunicativas en la esfera digital. Se ha hablado de noticias falsas, de *chatbots* que pretenden interrumpir un debate o simplemente sembrar cizaña, y de iniciativas surgidas en estos años, cuyo objetivo es verificar los contenidos falsos de forma semiautomática, en algunos casos, mediante el análisis lingüístico y de contenido de las narrativas empleadas. Está claro que ante este panorama y frente al desarrollo vertiginoso de las técnicas de manipulación de todo tipo de datos es necesario crear «herramientas [de análisis] novedosas que tengan en cuenta las nuevas formas de creación, amplificación y circulación de la información» (Levi 2019, 122).

Son numerosos ya los estudios dedicados a desarrollar grandes corpus de datos extraídos de redes y anotados para entrenar sistemas de detección automática de discurso de odio. No obstante, las discrepancias que existen en la definición de lo que es discurso de odio (*hate speech*, en inglés) hace que el conjunto de etiquetas empleadas para anotar las expresiones susceptibles de vehicular odio en corpus de entrenamiento sean diferentes. En el ámbito del español se puede mencionar el proyecto MEX-A3T (Álvarez Carmona *et al.* 2018), que contiene tuits de 5000 perfiles de español mexicano en el que se han anotado determinadas palabras (insultos) para clasificar los mensajes como «agresivos» o «no agresivos».

El proyecto IberEval18 (Fersini *et al.* 2018), en cambio, tiene como objetivo identificar la expresión de la misoginia en las redes sociales, para ello se creó un corpus de 3307 tuits que se clasificaron después como pertenecientes a una de las siguientes categorías: estereotipo; dominio; descarrilamiento: justificar el abuso de la mujer, rechazando la responsabilidad masculina; interrupción de la conversación; acoso sexual y amenaza; y descrédito. Por otro lado, Amores *et al.* (2021) han diseñado un detector automático de discurso de odio por motivos ideológicos en Twitter/X en español a partir de técnicas de aprendizaje automático supervisado y utilizando un total de ocho modelos predictivos.

Cuando se analiza el discurso de odio, cabe tener en cuenta la estructura multifacética y las características complejas de este fenómeno y considerar los aspectos semánticos, pragmáticos y socioculturales que permiten identificar una unidad de texto (por ejemplo, un comentario o una publicación) en su contexto apropiado, para evaluar si la expresión o la palabra que se emplea en ella cae o no dentro de la definición de discurso de odio. En proyectos

europeos[15] el fenómeno se analiza atendiendo a las características léxicas y semánticas de palabras o expresiones, cuyo significado intrínseco puede ser insultante o negativo. No obstante, el significado literal de algunas de estas palabras puede no ser siempre ofensivo, mientras que el significado implícito puede serlo en algunos contextos debido a las connotaciones que ha adquirido a lo largo del tiempo; también es posible determinar el grado de ofensividad (alto, medio, bajo) de las palabras incluidas en un modelo de aprendizaje o en taxonomías, como ilustra la base de datos colaborativa multilingüe https://hatebase.org o el proyecto MeOffendEs (v. nota 112).

Igualmente, cabe tener en cuenta la dimensión pragmática de las expresiones susceptibles de vehicular odio y otras manifestaciones de la agresividad, considerando los efectos comunicativos que producen en las personas, un determinado contexto. Algunos de los parámetros que determinan los efectos son el medio oral, escrito, público o privado; el emisor y el destinatario, identificados como individuo o grupo social que emplea la expresión de odio y la comunidad o grupo al que se dirige el mensaje; o los actos de habla que acompañan la palabra o expresión.

Otro de los métodos empleados para investigar la agresividad verbal en la red es el del ya mencionado análisis de las emociones, estrechamente relacionado con el análisis de sentimientos, como se explicaba en Mancera Rueda y Pano Alamán (2020, 25-27). Este es el conjunto de métodos y herramientas de Procesamiento de Lenguaje Natural y de aprendizaje automático, cuyo objetivo es identificar qué sentimientos negativos, positivos o neutros predominan en los millones de mensajes que se publican cada día en entornos digitales. En esta línea, el programa *Lingmotif* (Moreno Ortiz 2017, 2019) se apoya en metodologías sofisticadas que permiten comparar la «estructura afectiva» de diversos textos a la vez.

El análisis de emociones se apoya, en cambio, en categorías descriptivas de emociones como la ira, la sorpresa, la alegría o la tristeza. Entran aquí en juego distintos factores de emocionalidad, determinados por las funciones, los dispositivos y las modalidades expresivas habilitadas por las plataformas[16], los temas sobre los que se debate o el grado de anonimato. El desarrollo de

[15] Véase FAST-LISA, cuyo objetivo es «to create a consolidated and cross-border standard tool and protocol for the understanding, detection and counteraction of online hate speech». Información disponible en: https://fastlisa.eu/

[16] Recordemos que en YouTube se ha desactivado la posibilidad de utilizar el botón del pulgar hacia abajo para evitar la acumulación de mensajes negativos y agresivos.

estos sistemas ha llevado a elaborar reglas de contexto y métodos para controlar los elementos formales que actúan como marcas de afectividad, no solo verbales sino también visuales, como la expresión facial, los gestos o la postura.

Sabemos también que las plataformas, los mótores de búsqueda o las aplicaciones cuentan con tecnologías que registran cada página que visitamos e incluso los lugares y espacios físicos por los que nos movemos, y que conocen nuestras preferencias e intereses, de tal manera que orientan nuestras búsquedas y, en algunos casos, nuestros comportamientos. Estos datos pueden ser empleados por las grandes tecnológicas, que los utilizan, a menudo sin autorización o gracias a la autorización que damos de forma más o menos consciente al abrir una cuenta, para elaborar campañas de *marketing* o informar a las agencias de inteligencia (Cruz Rodríguez 2014).

El empleo de datos generados en las redes y en otras muchas plataformas y aplicaciones digitales para controlar las opiniones y las acciones de periodistas, abogados y activistas que en algunos países se consideran peligrosos para la seguridad del Estado, supone un enorme riesgo y un grave límite para el ejercicio de las libertades de expresión o de manifestación y para el respeto de los derechos humanos. De acuerdo con la *Declaración Deusto sobre Derechos Humanos en Entornos Digitales* (26 de noviembre de 2018), un modo de responder a este desafío sería concienciar a los ciudadanos de que detrás de las plataformas existen corporaciones como Google, Microsoft o Meta, que determinan las posibilidades y modalidades de participación en ellas y que, aun cuando se participa en esos espacios con la intención de contribuir al debate o movilizarse, se está inmerso en discursos que posicionan al sujeto como consumidor y como agente social y político (Mancera Rueda y Pano Alamán 2020, 28-29).

Si, por un lado, es fundamental considerar los riesgos y las amenazas que supone para las personas el acceso y consumo irreflexivo de contenidos digitales dependientes de los algoritmos (Alcántara-Plá 2022), por otro, es importante tener en cuenta que la extracción automática y el análisis de datos de interacciones digitales puede ser útil para combatir la criminalidad. Véase el caso del llamado *digital grooming* o acoso o abuso sexual en línea. En Lorenzo-Dus (2022) se explora este fenómeno como una práctica discursiva, de ahí que la investigación en este campo emplee los conceptos de posicionamiento (*stance*) y registro (*style*) para estudiar en profundidad las técnicas de manipulación que emplean los acosadores en la red. El volumen recoge diversos análisis empíricos de grandes corpus de datos, prevalentemente en inglés, para llevar a cabo análisis textuales y visuales de casos de acoso y abuso. Unido a

este tipo de investigación, cabe mencionar el proyecto pionero en España, *Stoponsexgroom*[17], cuyo objetivo es «combatir el ciberacoso sexual a menores a través del análisis detallado y riguroso de las estrategias de comunicación que utilizan los acosadores para engañar a los/las menores».

Los resultados del análisis han permitido desarrollar un modelo analítico de Online Grooming (OG) en español que incorpora tanto mecanismos de la conducta verbal de los acosadores como de la de los/las menores. El equipo reúne a lingüistas, psicólogos, criminólogos y juristas, un aspecto que fomenta la necesaria interdisciplinaridad y la creación y aplicación concreta de herramientas y recursos para afrontar este problema en la sociedad.

Otros aspectos críticos que cabe mencionar en este apartado en relación con el impacto sobre nuestras prácticas sociales y comunicativas, de algoritmos, modelos de aprendizaje automático y aprendizaje profundo y modelos masivos de lenguaje, es el de los sesgos basados en estereotipos de género, raza, orientación sexual, proveniencia geográfica, edad o lengua, que subyacen a los millones de datos que circulan por nuestras aplicaciones y que, de hecho, alimentan esos modelos. Ante el complejo panorama delineado, urge abrir nuevas pistas de investigación y métodos de análisis del discurso digital que permitan observar con ojo crítico lo que sucede comunicativamente en la red.

2. De cómo tratar el dato digital: desafíos y propuestas

Hoy en día no es posible realizar un estudio lingüístico sin emplear un corpus. De hecho, los numerosos estudios mencionados hasta ahora se apoyan en corpus de datos extraídos de interacciones digitales auténticas. Los análisis son c*orpus-based* o *corpus-driven*, en función de si las categorías están previamente determinadas y enmarcadas en una opción teórica o si emergen del análisis y dan sustento a la construcción de una teoría (Collins 2019).

Los datos extraídos de la comunicación en línea se han visto a menudo como datos fáciles de recopilar. Todo está disponible ante nuestras pantallas: un conjunto casi ilimitado de palabras e imágenes en una enorme variedad de contextos. Aunque este acceso es notable y sin precedentes para los analistas del lenguaje, la gran cantidad de canales, el cambio constante de las fuentes de datos y las cuestiones éticas de lo que es apropiado recopilar y analizar constituyen muchas veces un muro infranqueable.

[17] Información disponible en: https://stoponsexgroom.com/acerca-de/proyecto/

En primer lugar, cabe considerar qué elementos es posible extraer y utilizar, ya que las tecnologías móviles y el rápido desarrollo tecnológico crean contextos en los que las personas interactúan con una gran variedad de modos de comunicación *online* y *offline* difíciles de delimitar. En segundo lugar, es necesario reflexionar sobre la cantidad datos que sirve recoger para llevar a cabo un estudio de discurso digital. Dada la abundancia de datos disponibles en línea, los investigadores pueden recopilar corpus muy grandes para realizar investigaciones empíricas que documenten la evolución de la lengua, por ejemplo, la innovación léxica. No obstante, no todos los estudios requieren grandes cantidades de datos. Por ejemplo, las investigaciones sobre aspectos discursivos y multimodales de la comunicación digital en las que se adoptan métodos cualitativos, suelen basarse en muestras más pequeñas de datos.

En los últimos años, con el desarrollo de la Lingüística de corpus (LC), se ha impuesto una tendencia a recopilar textos naturales y completos, tratando de alcanzar una relativa extensión y diversidad. Entre las ventajas de utilizar corpus (Parodi 2010; Rojo 2021), se han señalado las siguientes: permiten una adecuada representación del discurso en muestras amplias y representativas de textos originales; posibilitan la exploración de textos etiquetados y no etiquetados; por medio del procesamiento (semi)automático de los textos, es posible realizar análisis más amplios y detallados; ofrecen mayor fiabilidad en los análisis mixtos o híbridos (Alcántara-Plá 2020), esto es, cuantitativos y cualitativos; los resultados son acumulativos y confrontables con posteriores investigaciones.

Son muchos los aspectos sobre los que se interrogan los investigadores del discurso digital en español y en otras lenguas cuando se disponen a analizar la lengua en estos entornos. Buena parte del debate gira hoy en día alrededor de la constitución, almacenamiento y accesibilidad de los corpus de datos extraídos de las web y de las redes y sobre los métodos de PLN aplicados al tratamiento y al análisis de esos datos (Vandekerckhove *et al.* 2019).

Las cuestiones que se plantean en este apartado hacen referencia tanto a la recopilación de datos como a su publicación, lo cual lleva a preguntarse también si deben ser de acceso abierto y qué aspectos éticos deben tenerse en cuenta. Respecto a la recopilación, también es importante considerar la extensión y la representatividad, si se trata de corpus de datos únicamente textuales o de datos multimodales y, si decidimos anotarlos para los fines de nuestra investigación, qué aspectos se pueden anotar (género de discurso, elementos lingüísticos, metadatos) y de qué manera. En este caso también, se señalan problemas relacionados con el muestreo, la distribución y el archivo a largo

plazo de datos extraídos, por ejemplo, de redes sociales, sabiendo que pueden desaparecer de un momento a otro.

Por último, no hay que olvidar que la mayoría de nuestras interacciones en línea explotan múltiples modos comunicativos para crear significado. Nuestra comunicación a través de plataformas y aplicaciones de mensajería incluye numerosos modos semióticos, de ahí que la manera de entender un corpus de textos digitales obligue a conceptualizaciones y análisis multidimensionales. Especialmente, es necesario reflexionar en torno a aspectos como el contexto de publicación de los mensajes que se insertan en contextos semióticos complejos constituidos por fotografías, vídeos, etiquetas o enlaces, que en muchos casos determinan el sentido del enunciado verbal con el que «conviven».

Para los analistas del discurso formados en lingüística, descubrir cómo tratar los datos no lingüísticos puede suponer un desafío conceptual y práctico (Duque 2020). Como señala Page (2019), el estudio de *selfies* de grupo o de historias destacadas en Snapchat, junto a los vídeos de TikTok, son solo algunos ejemplos de la amplia gama de recursos semióticos que pueden analizarse pero que no cuentan aún con métodos adecuados. Además de los recursos visuales, auditivos, hápticos y audiovisuales que intervienen en la comunicación en línea, hay que considerar que se trata de datos efímeros. Véase el caso de las *historias* de Snapchat, que están a disposición de los usuarios únicamente durante veinticuatro horas.

2.1. Orden en el caos: extensión y representatividad de los corpus

Crear un corpus de interacciones comunicativas que se originan en internet puede parecer una tarea sencilla, en la medida en que los datos ya están digitalizados, son accesibles fácilmente si se cuenta con una buena conexión y, si hablamos de datos lingüísticos escritos, no requieren una transcripción para su análisis. Sin embargo, hay una serie de aspectos que pueden ser problemáticos si no se toman las decisiones adecuadas. Cuando se elabora un corpus de este tipo, surgen incompatibilidades entre los métodos asentados de la lingüística de corpus cuando, por ejemplo, se lleva a cabo la lematización y anotación lingüística de los textos extraídos de una o más plataformas o sistemas. Según Beiβwenger y Storrer (2008), y como plantean también Alcántara-Plá *et al.* (2018), es necesario interrogarse sobre si cabe mantener la peculiar ortografía de los textos que se intercambian en chats, foros, redes o wasaps, caracterizados, como se ha dicho, por numerosos rasgos coloquializadores, o bien

adaptarla a la norma de cada lengua para poder realizar en esos textos análisis sobre datos etiquetados morfosintácticamente.

Otra cuestión abierta es la de la extensión y representatividad de las muestras de discurso digital. En principio, extraer miles de tuits o publicaciones de Facebook o Instagram en poco tiempo es sencillo. Pero, ¿es necesario recoger más o menos mensajes? En el espacio de pocas horas o incluso minutos, es posible que en una plataforma pueda acumularse un gran número de publicaciones, comentarios o respuestas, sobre todo si el tema que se discute es polémico o el contenido publicado se vuelve viral. En este sentido, la facilidad de acceso y la acumulación de datos en poco tiempo parecen justificar que para realizar análisis de discurso digital sean necesarios corpus muy amplios.

Este es el caso de análisis recientes sobre tendencias de consumo de contenidos dentro del llamado *social media analytics* y de proyectos de detección automática de emociones y sentimientos, como los que mencionaba previamente, para los que se necesitan millones de datos, en particular, si el objetivo es alimentar modelos de aprendizaje automático.

Asimismo, los análisis que buscan documentar fenómenos de variación lingüística requieren corpus extensos para poder extraer resultados válidos y generalizables, mientras que los que combinan métodos cuantitativos y cualitativos para explorar determinados aspectos de la interacción en entornos digitales, o bien que exploran cualitativamente las estrategias pragmáticas de la comunicación digital, suelen basarse en corpus controlados o muestras más pequeñas.

En el ámbito del español, buena parte de las investigaciones se ha apoyado en corpus de distintas dimensiones en función de los objetivos planteados (Pano Alamán y Moya Muñoz 2016; Pano Alamán 2020). Y es que, como sucede respecto a los corpus de textos escritos u orales producidos en otros contextos, los corpus de discurso digital deben responder a un diseño previo, diseño que se establece en función de «los objetivos que se pretenden alcanzar, pero también, de otros factores, como, por ejemplo, el tiempo, el personal o la financiación disponibles» (Rojo 2021; v. Collins 2019). Que hayan sido creados de forma manual o bien automáticamente a partir de la extracción de mensajes de chat, blogs, foros, comentarios en la prensa, SMS, reseñas digitales, redes o WhatsApp, su extensión depende en gran parte de los objetivos del proyecto y de los recursos de los que se dispone.

En cuanto a la representatividad de los datos es necesario pensar, siempre en función de los objetivos de análisis, cuál es la población cuyo comportamiento comunicativo interesa explorar y cuáles pueden ser los datos posibles

generados en un entorno particular. Como es sabido, la muestra es una parte de la población, cuyos datos se recopilan y analizan para proporcionar información que puede aplicarse después a toda la población. Cuando la muestra es representativa, refleja los rasgos de la población con mayor precisión, de ahí que la cantidad y tipología de datos necesarios para obtener una muestra adecuada dependerá de la población que se esté investigando, el enfoque y el objetivo de investigación. Por este motivo, conocer y extraer los datos sobre la identidad de los hablantes o los metadatos sobre la ubicación geográfica o el momento de envío o de publicación de un mensaje puede ser muy útil para establecer una muestra representativa[18].

Como recuerdan Estrada Arráez y De Benito Moreno (2016), esto es esencial en los estudios lingüísticos en los que se utiliza Twitter/X, donde recopilar datos y descargar los metadatos de los mensajes que se extraen automáticamente ha sido hasta ahora bastante sencillo. Como demuestra precisamente el caso del microblog[19], la evolución tecnológica y las decisiones que toman los propietarios de las compañías respecto al acceso por parte de terceros a los datos publicados en las redes y sistemas que controlan tiene un gran impacto sobre el diseño de investigación. Por este motivo, el muestreo de datos puede ser uno de los aspectos del proceso de investigación que más diferencia el análisis del discurso digital de formas anteriores de análisis del discurso centradas en la interacción social.

Normalmente, los estudios adoptan soluciones flexibles para establecer una proporción adecuada del corpus que conduzca a ciertas proyecciones, renunciando en algunos casos a realizar generalizaciones (Vásquez Roca 2022). Por otro lado, un corpus no es una única instancia comunicativa y tampoco cuenta con un cierre de ningún tipo, sino que presenta una organización predeterminada en torno a categorías identificables para la descripción de un determinado fenómeno lingüístico. Por ejemplo, en el análisis de los cumplidos dirigidos en Twitter/X a los deportistas españoles durante los Juegos Olímpicos de Río 2016, Hernández Toribio y Mariottini (2020) recopilan un corpus de 500 tuits

[18] Ahora bien, a propósito de la noción de *web as a corpus*, Kilgarriff y Greffenstette se interrogan sobre este concepto, afirmando que: «Representativeness begs the question 'representative of what?' Outside very narrow, specialized domains, we do not know with any precision what existing corpora might be representative of» (2003: 340).

[19] Me refiero a la decisión del propietario, Musk, de eliminar el acceso a las apps de desarrolladores que ofrecían servicios asociados a la API del microblog o servicios alternativos a la app de la compañía. Esto ha limitado la posibilidad de obtener datos internos de la red para investigar. Información disponible en: https://www.eldiario.es/tecnologia/elon-musk-carga-cazadores-bots-twitter-deja-ciegos-justo-elecciones_1_10326741.html

extraídos de los perfiles de 14 deportistas españoles y publicados en agosto de 2016. En este caso, el muestreo se realiza a partir de un subgrupo de usuarios, recopilando una muestra de mensajes relacionados con el fenómeno pragmático (cumplidos) que se está analizando. Empleando también un corpus cerrado de tuits se han indagado aspectos lingüísticos muy distintos, como la preferencia por la coordinación o la subordinación por parte de los usuarios (Recio Diego y Tomé Cornejo 2017). En estos y otros estudios, se realiza una compilación de datos «idealmente representativa de uno o varios medios virtuales» (De Benito Moreno y Estrada Arraéz 2016, 10). Estos enfoques llevan a circunscribir los datos potenciales que podrían incluirse en el análisis, pero permiten considerar hasta qué punto los hallazgos son generalizables a una población más extensa.

En general, si se plantean las preguntas de investigación, los recursos disponibles y la población que interesa investigar desde un principio, puede ser útil realizar un estudio piloto para calibrar la cantidad y los métodos de recopilación de datos (Pihlaja 2022). En relación con este aspecto, es posible ampliar el concepto de muestra o población a través del concepto de comunidad, que pone el foco no tanto en quienes interactúan publicando mensajes individuales o hilos de discusión, sino más bien en el resultado de la interacción continua entre miembros de comunidades de intereses o de práctica. Si, por un lado, la idea de comunidad como grupo de usuarios con intereses compartidos es susceptible de enriquecer la noción de población, por otro, plantea desafíos para el muestreo de datos. Para obviar este problema, se propone que el investigador participe en una observación o participación a largo plazo en la comunidad sobre la que investiga, utilizando métodos como la etnografía en línea centrada en el discurso (Androutsopoulos 2008). Esta es la solución que adopta Noblía (2018) para su análisis de los modos como estrategia de atenuación en un grupo de WhatsApp en un contexto profesional. En todo caso, hay que tener en cuenta que, cuando las interacciones abarcan escalas de tiempo prolongadas y una variedad de plataformas diferentes, tanto públicas como privadas, la observación y la recogida de los datos relevantes para identificar a la comunidad en cuestión se hacen más complejas.

La constitución de un corpus de discurso digital requiere también comprender cómo son los contextos digitales de los que se extraen datos lingüísticos. Es curioso que en estudios actuales sobre interacciones en redes se dedique amplio espacio a describir los casos de estudio, la muestra considerada, la extensión del corpus y los procedimientos de creación del corpus, mientras se limitan o se prescinde de las explicaciones sobre los parámetros tecnológicos y

sociosituacionales que caracterizan el entorno digital analizado. Algunos de los parámetros propios de ciertas aplicaciones y plataformas, como su carácter privado, pueden dificultar la recopilación de datos. Este es un factor que caracteriza, por ejemplo, el correo electrónico o la mensajería instantánea. En cambio, los espacios de interacción pública, como los comentarios a las noticias de la prensa digital o los intercambios en las redes sociales, pueden ser extraídos de forma relativamente más fácil (Mancera Rueda y Pano Alamán 2014a), como veremos.

Más allá de la indicación de estos parámetros en el estudio que se lleva a cabo, es conveniente prestar especial atención también a cómo los hablantes entienden y usan estas tecnologías y plataformas a lo largo del tiempo. Este dato, que es posible recopilar por medio de entrevistas semiestructuradas y cuestionarios (Noblía 2018; Cantamutto y Vela Delfa 2020), puede ser esencial no solo para decidir cómo diseñar la investigación sino también para dar respuesta a nuestras preguntas.

2.2. Unir esfuerzos: corpus accesibles para la investigación

En estudios previos (Pano Alamán y Moya Muñoz 2015, 2016) sobre la evolución de las investigaciones en torno a la comunicación digital en español y la constitución de corpus en este campo, señalábamos que la multiplicación creciente de corpus asociados a análisis específicos sobre algunos entornos y con distintos métodos de recopilación y anotación estaba llevando a una disparidad de muestras a las que es difícil acceder o que difícilmente pueden compararse.

Si bien es posible empezar a hablar de una disciplina como la del Análisis de la comunicación digital que, en el ámbito del español, cuenta ya con congresos específicos, paneles temáticos en coloquios y proyectos de investigación, no existen todavía métodos compartidos por los investigadores para la creación, recopilación, anotación y puesta disposición de recursos y corpus[20]. En este campo parece necesario unir esfuerzos, y es que la constitución de corpus de discurso digital «exigirá la continua colaboración de distintos investigadores» (De Benito Moreno 2022: 489), amén del apoyo institucional.

[20] El hecho de que muchos de los datos recopilados dentro de proyectos o de estudios concretos no estén disponibles en corpus o bases de datos provoca el riesgo de que se pierdan y supone que otros investigadores no puedan acceder a ellos para replicar un estudio.

Existen, no obstante, iniciativas que van en esta dirección, como CODICE[21] (Vela Delfa y Cantamutto 2015), repositorio abierto y colaborativo de interacciones comunicativas digitales de distintas variedades del español. Las muestras se han preparado en base a un modelo de descripción de los datos a partir de su propuesta de propiedades del discurso digital (Cantamutto y Vela Delfa 2016), basada en los modos de realización, los modos de enunciación y las relaciones interpersonales, analizados en SMS, correo electrónico y WhatsApp. Otros proyectos que permiten acceder a datos de discurso digital y realizar búsquedas en línea son MEsA y MesA 2.0 (Fuentes Rodríguez 2021) y la Plataforma MarcoPolo[22], creada para analizar el discurso político electoral español (Alcántara-Plá *et al.* 2018).

El primero se presenta como una herramienta útil para «todos aquellos lingüistas o investigadores que quieran acercarse a la comunicación verbal en Internet, poniendo especial foco en los textos escritos». El corpus creado en el marco de los dos proyectos recoge materiales publicados en blogs, foros, sitios web, WhatsApp, Facebook, Instagram, Twitter/X y YouTube, transcritos siguiendo el protocolo descrito en la *Guía del usuario* disponible en la web del proyecto[23].

El segundo, la plataforma Marco Polo, da acceso a las coocurrencias de las palabras más frecuentes de un extenso corpus de tuits anotados morfológicamente y ofrece la posibilidad de buscar palabras y comparar su frecuencia de uso mediante gráficos. Dentro de este proyecto se descargaron los tuits y retuits que se publicaron entre octubre de 2015 y junio de 2016 en las cuentas oficiales de cinco partidos políticos españoles y de sus correspondientes líderes en ese periodo. Sumando el número de mensajes publicados por los partidos, por una parte, y por los políticos, por otra parte, se ha podido garantizar la representatividad del corpus y equilibrar las diferencias entre los dos grupos de perfiles.

De naturaleza distinta es el corpus WebLesp, creado por Piccioni y Pontrandolfo (2021). Accesible en línea[24], reúne textos de blogs, sitios web y artículos de prensa digital representativos de la comunicación digital especializada en los ámbitos de las ciencias ambientales, el derecho, la economía y la medicina. El corpus, etiquetado y lematizado, puede ser consultado a través de la

21 https://codice.aplicacionesonline.com.ar/

22 http://www.worldslab.eu/marcopolo/

23 *Guía de usuario* disponible en http://www.grupoapl.es/images/archivos/MESA/Gu%C3%ADa_2.0.pdf

24 https://corpora.unich.it/sito/corpus-weblesp-es.html

plataforma *Sketch Engine*, que permite obtener listados de frecuencia, colocaciones, concordancias y palabras clave que facilitan el estudio de los rasgos divulgativos de estos cibergéneros.

A pesar de las diferencias que presentan respecto a los corpus que se acaban de mencionar, señalo también el *Corpus del Español: Web/Dialects*[25] que contiene alrededor de dos mil millones de palabras en español, extraídas de dos millones de sitios web de 21 países diferentes de habla hispana; y la familia de corpus *esTenTen*, que se ha descrito en parte en la sección 1.2. Como se decía, *esTenTen* hace referencia al conjunto de corpus de datos lingüísticos extraídos de la web y disponibles en más de 40 idiomas, que han sido procesados con un tamaño objetivo de más de diez mil millones de palabras[26]. La plataforma de pago *Sketch Engine* da acceso a estos corpus, en concreto, a *esTenTen18*, con 16900 millones de palabras y con una clasificación de los temas para los dominios web más extensos; y a *esTenTen11*, con 9500 millones de palabras. Ambos incluyen textos extraídos de la web en español europeo, web en español americano y de la Wikipedia en español. El etiquetado y la lematización de partes del discurso de estos corpus se han realizado mediante el analizador *FreeLing* con configuración en español.

Si bien es posible considerar estos últimos macrocorpus como recursos de los que dispone el investigador para analizar el español en la web, es necesario establecer una distinción entre los corpus creados específicamente para analizar los rasgos del español en entornos digitales (CODICE, Marco Polo, WebLesp o MESA) y los corpus de referencia que contienen únicamente datos extraídos de la web o una parte de datos extraídos de la web. En este último grupo puede incluirse, por ejemplo, el CORPES XXI de la Real Academia Española, que cuenta con más de 365000 documentos. Como se declara en la página de presentación, «Los textos procedentes de libros suponen más de 186 millones de formas […]. Algo más de ocho millones más provienen de blogs, entrevistas digitales, redes sociales y miscelánea»[27].

La inclusión de textos extraídos de distintos cibergéneros confirma que los contenidos generados hoy en día en la red son un recurso fundamental para analizar distintos aspectos del español actual. Volviendo, en cambio, a los corpus que se focalizan específicamente en la comunicación digital y en las prácticas discursivas surgidas en la red, es necesario seguir ampliando los

[25] https://www.corpusdelespanol.org/web-dial/

[26] Información disponible en: https://www.sketchengine.eu/estenten-spanish-corpus/

[27] https://www.rae.es/banco-de-datos/corpes-xxi

recursos puestos a disposición para la investigación en este campo, y contribuir a la reflexión sobre las decisiones metodológicas adoptadas.

2.3. Algunos pocos ingredientes en la fase de extracción de datos

Desde un punto de vista metodológico, la reflexión actual se centra principalmente en las fases de recogida y fijación de muestras de discurso digital, el almacenamiento, el etiquetado o la anotación, además de la publicación y difusión de los corpus y otros recursos con fines académicos o para su divulgación (webs de proyectos, medios de comunicación).

En lo que respecta a la extracción y recolección de datos, no es fácil establecer fórmulas más o menos permanentes y que puedan aplicarse en todos los casos[28]: en primer lugar, por la constante transformación de los dispositivos y de las aplicaciones; en segundo lugar, por las diferencias en términos técnicos y de interfaz que presentan los sistemas y las plataformas en los que se apoyan el correo electrónico, los blogs, los foros, la mensajería instantánea y las redes sociales. Esto ha llevado a adoptar métodos y técnicas de recogida complementarios (extracción automática, semiautomática y manual), y en ocasiones, a acotar las muestras, especialmente si el entorno que se analiza combina distintos modos semióticos.

Los primeros estudios contaban con muestras de datos pequeñas (microcorpus) a menudo extraídas manualmente a través de funciones como *copia-pega* o descargando los datos en formato html o xml de una web o de una plataforma (Mancera Rueda y Pano Alamán 2013; Pano Alamán y Muñoz Moya 2016). En años recientes, se ha empezado a combinar varios métodos, como la extracción automática de los datos lingüísticos, y manual, por ejemplo, a través de capturas de pantalla, de datos multimodales (emojis, fotografías, infografías, vídeos, enlaces).

Cuando se quieren extraer grandes cantidades de datos de la web se suelen utilizar los llamados raspadores web (*web scraper*). Para proyectos más específicos, gracias a la existencia de Interfaces de programación de aplicaciones (APIs en inglés), es posible acceder automáticamente a los datos lingüísticos que se generan en distintas plataformas, mediante *software* que interrogan la

[28] Como se desprende del hilo de tuits publicado por @SoyMmadrigal el 10 de junio de 2023 a propósito de la extracción de contenidos políticos en TikTok, las plataformas son «territorios salvajes» para la extracción y análisis de datos. Hilo disponible en: https://twitter.com/SoyMmadrigal/status/1667481128498806784?t=fxR6dq20n8GXpHjqDgpRiw&s=03

API de la plataforma en cuestión o a través de *script* programados para tal fin. Esto permite constituir tanto macrocorpus como microcorpus que permiten realizar investigaciones con muestras más representativas, en función de los objetivos que se plantean, además de incluir metadatos útiles.

En este sentido, cabe recordar el caso de Twitter/X, que cuenta con el mayor número de muestras de discurso digital en redes debido en parte a sus características y al –hasta ahora– fácil acceso a sus datos a través de su API. La extracción automática de tuits es relativamente sencilla con determinados *script* creados con lenguajes de programación como Python[29].

La alternativa a este método es utilizar uno de los muchos programas gratuitos o de pago, disponibles en línea o descargables en el propio ordenador, diseñados con ese objetivo. Sin afán de exhaustividad, menciono aquí algunos de los más utilizados para extraer datos de Twitter/X: Twlets.com[30], Tweet Archivist[31], TwDocs[32], TAGS[33] o Deep Talk[34], que se basa en modelos de inteligencia artificial. Véanse también los archivos en línea de TwitterStream Grab[35], donde se almacenan muestras de tuits publicados mensualmente desde 2014 (De Benito Moreno 2022). Otros programas como ExportComments[36], IntaLoadGram[37], Sudota[38], Next Analytics[39], FireAnt[40] o DataOryx[41] permiten descargar no solo datos del microblog sino también de Instagram, Facebook, YouTube, TikTok, Twitch o Reddit.

Buena parte de estos servicios recuperan y guardan los mensajes extraídos en documentos .xls (Excel) que ordenan los metadatos y los datos en columnas

29 Sobre la posibilidad de crear *script* con lenguajes de programación como R o Python para extraer tuits automáticamente y exportarlos en formatos manejables, véanse De Benito Moreno y Estrada Arráez (2018), Alcántara Plá *et al.* (2018) y De Benito Moreno (2022).

30 https://twlets.com/

31 https://www.tweetarchivist.com/

32 https://www.twdocs.com/

33 https://tags.hawksey.info/

34 https://www.deep-talk.ai/

35 https://archive.org/details/twitterstream

36 https://exportcomments.com/

37 https://instaloadgram.com/

38 https://sudota.com/

39 https://nextanalytics.com/

40 https://www.laurenceanthony.net/software/fireant/

41 https://dataoryx.com/

de fácil consulta y uso para el análisis. Por ejemplo, el contenido de un mensaje, generalmente el texto y los emojis, etiquetas y enlaces que contiene, se recogen en una misma columna. Ese mismo contenido puede exportarse a un archivo .txt para ser analizado después con sistemas de gestión y análisis de corpus como *Sketch Engine*, *AntConc* o *LancsBox*.

La desventaja de este sistema es que los datos se extraen en formato de texto dejando de lado los contenidos multimodales como, por ejemplo, las imágenes o los vídeos que lleva incrustados. En este caso, algunas herramientas descargan el URL de cada mensaje (*URL status*), facilitando el acceso al contenido original publicado en la plataforma. Asimismo, si lo que se busca es analizar las interacciones que se producen en un entorno determinado será importante descargar no solo los mensajes que constituyen intervenciones iniciativas (publicación, post, vídeo, comentario, tiktok), sino también los que se publican como intervenciones reactivas, de forma automática –algunas herramientas lo prevén– o manual, modificando la opción de visualización de esos datos, de modo que la estructura del intercambio sea visible.

Cuando se acude a este tipo de servicios, especialmente si se quieren utilizar dentro de proyectos de investigación a medio o largo plazo, hay que tener en cuenta dos aspectos esenciales. El primero es que estos programas aparecen y desaparecen con mucha rapidez debido a cuestiones de rentabilidad económica de las empresas que están detrás, o a las decisiones de las compañías propietarias de las plataformas respecto al acceso a *sus* datos por parte de terceros, como apuntaba en la nota 149.

Este es el caso de TweetReach[42] o de Vicinitas[43], un servicio de extracción de datos de Twitter/X que ya no está disponible. El servicio en línea permitía extraer gratuitamente hasta 3200 tuits a través del sistema de búsqueda *Real-Time Tweets*. Igualmente, ofrecía un servicio de pago con reducción para investigaciones académicas, *Historical Tweets*, para obtener tuits publicados en fechas o periodos de tiempo concretos, incluso lejanos en el tiempo respecto al momento de la extracción. Esta opción era muy útil para rastrear, mediante *hashtags*, los mensajes relacionados con acontecimientos, eventos y campañas desarrolladas en un periodo más o menos largo de tiempo. El servicio cerró en abril de 2023 coincidiendo con la nueva política de acceso trámite API de Twitter[44].

[42] Como se explica en este artículo: https://fedica.com/blog/tweet-reach-shut-down-whats-the-alternative/

[43] https://www.vicinitas.io/

[44] Información disponible en: https://twitter.com/XDevelopers/status/1641222782594990080

El segundo aspecto, mencionado ya brevemente, se relaciona con las limitaciones que las plataformas imponen a terceros y que afectan a las preguntas de investigación. Por ejemplo, respecto al número máximo de datos que es posible descargar en periodos de tiempo predeterminados (al día, al mes), ya se emplee la versión gratuita, la de prueba o la de pago; a la decisión de trabajar sobre una plataforma en detrimento de otras; y al tipo de búsqueda que se lleva a cabo para recopilar los datos. La búsqueda puede hacerse por nombre de usuario, de perfil o de cuenta en algunos casos, o bien por *hashtag* o incluso por palabras clave, dando más opciones para investigar flujos de discurso en línea organizados temáticamente (Androutsopoulos 2013, 496)[45].

Un último aspecto a tener en cuenta sobre este método es que, si bien esta es una forma conveniente de recopilar datos de redes sociales, los resultados basados en la API pueden filtrarse mediante algoritmos, lo que implica que el investigador no tenga el control de los datos que descarga, ya que algunos pueden haber sido excluidos (Eynon y Schroeder 2016). Estas herramientas, sobre todo las gratuitas, no suelen recuperar todo lo que nos sirve, puesto que la API estándar que interrogan recupera los datos en función de la relevancia y no de la integridad.

Como se ha dicho, no todos los análisis necesitan grandes cantidades de datos y no todos los investigadores tienen las habilidades de programación o las posibilidades de acceder a las herramientas descritas. En estos casos, en función de los objetivos del estudio y del método cualitativo o mixto que se adopte es posible recuperar cantidades más pequeñas de forma manual, por ejemplo, mediante pantallazos o *screenshot* de los mensajes que nos interesan, que pueden guardarse en formato imagen o bien copiarse en procesadores de texto. Esta técnica captura y mantiene el formato original de los mensajes así como los metadatos que podrían perderse utilizando raspadores de web o programas de extracción, sin embargo, puede ser necesario obtener muchas más capturas para poder identificar los vínculos de unos mensajes con otros. En todo caso, puede ser útil si, por ejemplo, interesa extraer contenido multimodal de una plataforma, como pueden ser los memes. En estos casos, es posible combinar capturas de pantalla con aplicaciones de raspado en el navegador web.

Asimismo, al guardar los datos, es esencial utilizar estrategias de etiquetado y numeración claras para almacenarlos. Kaczmirek *et al.* (2014) utilizan

[45] Véase la descripción de esta metodología en Pano Alamán (2023), donde se lleva a cabo un análisis basado en corpus de tuits publicados en torno al cambio climático.

MongoDB, una base de datos en la nube que archiva contenido extraído de Twitter/X, así como Microsoft SQL Server, para almacenar datos de Facebook.

Hasta ahora he mencionado las posibilidades de recoger datos publicados generalmente en plataformas en que los mensajes son públicos (blogs, ciertos comentarios a las noticias, Twitter/X, YouTube) o semipúblicos (Facebook, Instagram, TikTok)[46]. ¿Qué sucede, en cambio, si queremos analizar la comunicación digital en espacios privados, asociados a una dirección o a un contacto telefónico, como el correo electrónico o la mensajería instantánea?

Cantamutto y Vela Delfa (2023) ofrecen un estado de la cuestión sobre las diferentes técnicas de recolección de muestras de conversaciones en WhatsApp. El hecho de que sean interacciones digitales privadas, caracterizadas cada vez más por la multimodalidad, dificulta aún más su recopilación. Las autoras constatan que se da una «debilidad metodológica» en trabajos sobre mensajería privada en los que las muestras, de tamaño limitado, se construyen generalmente por conveniencia. Así, los corpus de WhatsApp suelen estar formados por muestras pequeñas de interacciones de redes de familiares y amigos, o bien de contactos con jóvenes en instituciones educativas.

Antes de exportar las conversaciones que se desarrollan en los espacios de interacción entre dos personas o entre varias dentro de grupos, es habitual ponerse en contacto con los participantes para obtener su autorización[47], o bien para solicitar su participación si se pretende generar datos para descargarlos después, como en el estudio de Vázquez Cano *et al.* (2015), quienes analizan 417 conversaciones entre estudiantes de centros educativos españoles que participaron voluntariamente en el proyecto. En otros estudios sobre WhatsApp (Alcántara-Plá 2014; Sampietro 2016) se adoptan diferentes métodos para la creación de corpus y de muestras por conveniencia. Ante la disparidad de métodos y en función de sus propias investigaciones, Cantamutto y Vela Delfa

[46] Recordemos que los límites entre público y privado son difusos en la red; además, esos conceptos están relacionados no solo con el acceso a los mensajes sino también con el tipo de contenido. Aunque algunos mensajes pueden recuperarse desde plataformas de acceso público, el contenido que vehiculan puede incluir información confidencial. Por ello, a pesar de que los datos sean públicos o semipúblicos, la información personal que contienen debe ser anonimizada.

[47] Los blogs, los comentarios en línea, los foros de debate o las plataformas de redes sociales son, como decía, entornos públicos o semipúblicos más accesibles que no requieren, en principio, contactar a los usuarios para poder extraer datos verbales y multimodales. No obstante, esto no exime de la práctica de anonimización de esos mismos datos y de la adopción de una ética investigadora, si se quieren publicar, como se verá en la última sección de este capítulo.

(2023) proporcionan indicaciones prácticas para recoger datos, establecer un contacto con los informantes, garantizar su privacidad y almacenar las muestras.

Dependiendo del sistema operativo del móvil que se utiliza, es posible llevar a cabo una exportación de los datos verbales y multimedia (emojis, imágenes, audios) de WhatsApp desde la misma aplicación[48]. Puede hacerlo el investigador o la investigadora a partir de sus contactos, o bien a partir de sus propios contactos (familiares y amigos) que actúan como «intermediarios» y que pueden enviar los datos al investigador a través de la aplicación de mensajería o por correo, con el riesgo, en este caso, de perder datos (Sampietro 2016). Esta opción obliga también a tener en cuenta la diversidad de dispositivos y sistemas operativos con los que cuentan los colaboradores, por lo que es necesario darles instrucciones precisas antes de proceder con la recogida de los datos (Cantamutto y Vela Delfa 2019). En ambos casos, cabe obtener la autorización de los contactos para poder descargar sus conversaciones.

A la hora de extraer conversaciones de WhatsApp, hay que tener muy en cuenta también las restricciones del sistema relativas al tamaño del archivo completo de datos y metadatos que se exporta, dependiente del peso de los documentos adjuntos, como las imágenes, los vídeos o los audios cada vez más habituales en este tipo de interacciones. Una opción para dar respuesta a esta limitación puede ser la de combinar la exportación automática de archivos de texto con métodos manuales, como las capturas de pantalla. Esto permite conservar los datos multimodales, incluida la estructura del intercambio y las marcas paratextuales, además de evitar la manipulación (Androutsopoulos 2013).

Dada la naturaleza efímera de estos contenidos, es esencial capturar y guardar el mayor número de datos relevantes para la investigación, en todos los formatos potencialmente útiles, apuntando la fecha en que se han recogido, puesto que lo que un día está accesible en línea es susceptible de desaparecer al día siguiente (Vásquez 2019).

Los procesos de extracción de muestras de lengua para investigaciones sobre WhatsApp y no solo (véase también el caso del correo electrónico) prevén que quien recoge o elicita las muestras pueda ser al mismo tiempo un participante en la interacción, de ahí que la adopción de técnicas etnográficas aplicadas ya en el estudio de los chats (Mayans 2002), como la observación-participante, sean fructíferas en el estudio de WhatsApp y de aplicaciones similares (Vela Delfa y

[48] Desde el menú desplegable situado en la parte superior derecha de la aplicación, que coincide con tres puntos, es posible seleccionar la opción «Más» y de ahí la opción «Exportar chat».

Cantamutto 2016). De este modo, el investigador evita el riesgo de invadir el espacio de participación y accede a ese espacio comunicativo directamente sin necesidad de que los participantes le envíen información. Asimismo, los datos extraídos automáticamente o exportados del sistema pueden complementarse con datos obtenidos de los usuarios a través de cuestionarios, entrevistas y hábitos de test sociales, que permiten analizar de forma más completa ciertos usos y comportamientos comunicativos en el entorno que se investiga.

En definitiva, las plataformas y las aplicaciones establecen términos y condiciones diferentes de acceso que inciden en los métodos de recopilación de las muestras. Se pueden recuperar conjuntos de datos más completos y potencialmente menos filtrados, pero esto puede ser costoso o requerir habilidades de programación o herramientas de *web scraping* que no garantizan una extracción de datos libre de ruido. Estos aspectos condicionarán el tipo de investigación que se puede llevar a cabo, por ello es conveniente que, en los apartados de metodología y constitución de los corpus consten las decisiones tomadas sobre estas cuestiones, esto es, las características de la plataforma o del sistema seleccionado, el método de extracción, las posibles limitaciones en el acceso y recogida de datos, las soluciones adoptadas y los procedimientos de preparación y gestión de los datos y de los metadatos obtenidos para el análisis. De este modo, aunque se trate de datos que se manejan de forma privada, será posible aportar recomendaciones e intercambiar información sobre buenas prácticas en este campo.

2.4. <Post>: la anotación de los datos de comunicación digital

Si bien, como se ha dicho, los análisis sobre interacciones en Twitter/X o WhatsApp proporcionan indicaciones útiles sobre cómo constituir corpus o sobre cómo adoptar cierta cautela a la hora de extraer los datos, son escasos aún los planteamientos relativos al etiquetado de los corpus y su anotación, dos aspectos que afectan a la posibilidad de crear recursos disponibles en línea que puedan ser interrogados por los investigadores gracias al empleo de estándares de representación de los datos.

Desde un punto de vista metodológico, la cuestión del etiquetado con etiquetadores morfológicos (*tagger*) y sintácticos (*parser*) se aplica también al discurso digital; este permite afinar los análisis lingüísticos basados en corpus y tratar los datos de forma automatizada. En el ámbito del español, Alcántara-Plá *et al.* (2018) indican que el corpus de tuits electorales recopilados para el proyecto Marco Polo ha sido etiquetado con el programa *Freeling*, que está integrado en *Sketch*

Engine[49], y que se aplica a la lematización y etiquetado morfosintáctico de corpus de textos escritos y orales. Mediante el programa es posible segmentar los tuits en palabras y etiquetarlas, utilizando expresiones regulares adaptadas al español.

Sin embargo, como bien señalan los investigadores, etiquetar muestras de discurso digital con Freeling o con herramientas similares no garantiza resultados completamente válidos. En la medida en que esta herramienta está entrenada sobre corpus de español normativo (Alcántara-Plá *et al.* 2018, 15), presenta limitaciones ante elementos como las menciones, los enlaces y las etiquetas, los vídeos, audios o emojis, y ante los rasgos coloquializadores presentes en los textos. Me he referido ya en el primer capítulo a la peculiar ortografía que se observa en algunos tipos de interacciones digitales, un aspecto que puede dificultar la anotación (Mancera Rueda y Pano Alamán 2014, 309).

Otro estudio, dedicado en este caso al análisis de sentimientos en reseñas de Amazon y tuits (Moreno Ortiz 2019), utiliza el programa *Lingmotif*, integrado por un *core sentiment lexicon* que incluye más de 200000 palabras y expresiones multipalabra etiquetadas con el *Penn Treebank tag set*[50] atendiendo a su forma y categoría gramatical. Para determinar su valencia y polaridad, emplea recursos lexicográficos como Wiktionary y otros creados específicamente para el análisis de sentimientos. El programa integra también un léxico de complemento (*plug in lexicon*) para tener en cuenta el «sentimiento» contenido en expresiones de dominios y campos semánticos específicos (Moreno Ortiz 2017). La anotación automática del léxico mediante este tipo de recursos enriquece enormemente el análisis de corpus que se lleva a cabo, aunque, como bien señala Moreno Ortiz (2019, 54), la desambiguación de algunas palabras y la asignación de significado a los numerosos datos multimodales –como el pulgar hacia arriba– presentes en textos digitales como las reseñas que analiza, son tareas difíciles de automatizar.

Si se trata de llevar a cabo análisis sofisticados de corpus de discurso digital es necesario reflexionar también sobre los procedimientos de etiquetado y anotación de este tipo de datos. Sobre todo hoy en día, cuando podemos llevar a cabo análisis asistidos por corpus con programas de *Qualitative Discourse Analysis* aplicados al estudio de la comunicación digital, como ATLAS.ti o NVivo (Paulus 2022).

Más allá de los etiquetadores morfosintácticos y de la anotación semántica, puede ser útil ampliar la perspectiva y considerar no solo el nivel microtextual

[49] Información disponible en: https://www.sketchengine.eu/spanish-freeling-part-of-speech-tagset/

[50] https://www.sketchengine.eu/spanish-treetagger-part-of-speech-tagset/

o lingüístico de los datos, sino también la macroestructura de los textos que se recopilan, teniendo en cuenta que los datos extraídos a partir de la web presentan un elevado grado de heterogeneidad, que no siempre están bien estructurados y que contienen ruido (Martínez Santiago *et al.* 2001).

Una reflexión interesante en este sentido se ha llevado a cabo en proyectos de creación de corpus de discurso digital en alemán, francés, italiano y esloveno, dentro del *Special Interest Group* (SIG) sobre *Computer-Mediated Communication* (CMC), centrado en «modelling user contributions (posts) to written CMC dialogues; modelling CMC document structures (macrostructures in forum threads, chat logfiles, Twitter timelines); developing perspectives for the representation of discourse in multimodal CMC».

El grupo asume que internet y las redes sociales han dado lugar a una amplia gama de nuevos *géneros* comunicativos, como los chats, los foros, los mensajes de texto (SMS, WhatsApp), las páginas de «conversación» en Wikipedia, los comentarios en blogs y los mensajes de las redes sociales y de entornos 3D multimodales. Para la representación y descripción de dichos géneros y de los rasgos estructurales y lingüísticos que presentan, los investigadores proponen la adopción del estándar de etiquetado digital de la TEI (*Text Encoding Initiative*)[51], que garantiza la interoperabilidad entre los recursos lingüísticos que se integran en los corpus de discurso digital, así como el análisis y la explotación automática de recursos de ese tipo en varios aspectos. Entre otros, señalan la creación de corpus comparables en diferentes idiomas, lo cual podría mejorar la base empírica para realizar investigaciones en distintas lenguas. Igualmente, incluir modelos de anotación multimodal permite describir y analizar el discurso digital en sus diferentes modos. El grupo trabaja, pues, en la adaptación de las directrices de la TEI para la representación de datos extraídos de distintos cibergéneros.

Beißwenger *et al.* (2012) proponen un modelo de etiquetado (*CMC-core schema*)[52] que se relaciona con los distintos *modos sociotécnicos* de comunicar

[51] Información disponible en: https://tei-c.org/Activities/SIG/CMC/; TEI es una iniciativa lanzada a finales de los años 1990, que propone un estándar para la representación de los textos (escritos y orales) en forma digital, dentro del amplio campo de las humanidades digitales. Se basa en el lenguaje de marcado XML (*Extensible Markup Language*). A diferencia de otros formatos textuales, como HTML, se trata de un etiquetado semántico y no presentacional que identifica y establece el «significado» de cada uno de los elementos que aparecen en un texto, asignándoles un atributo. La última versión de las guías directrices de la TEI (P5), actualizadas en abril de 2023, están disponibles en este enlace: https://tei-c.org/release/doc/tei-p5-doc/es/html/index.html

[52] Existen distintos esquemas de anotación aplicados a otros proyectos. Para una descripción de estas propuestas, véase la página: https://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication

en la red. Por ello establecen una diferencia entre el texto plano, que puede incluir enlaces (URL, *hashtag*) y el archivo de audio, fotografía y otros elementos *incrustados* (*embedded*) en el mensaje, asumiendo que este tipo de elementos son predominantes en el discurso digital actual y no pueden ignorarse en el proceso de anotación formal[53]. Por otra parte, distinguen entre dos niveles de anotación: el de la macroestructura del texto, que afecta a las unidades básicas de los textos resultantes de las actividades comunicativas de los usuarios en un entorno específico, así como de las rutinas del sistema; y el de la microestructura, que hace referencia al contenido verbal y multimodal de esas unidades. El primer nivel se relaciona con las unidades básicas de texto empleadas en una determinada interacción. La macroestructura hace referencia al orden y la presentación de los mensajes que se publican en una plataforma o sistema y que aparecen en el espacio de la pantalla como resultado de las actividades del usuario y de los procesos del sistema que emplea. En este caso, la macroestructura incluye no solo la unidad mensaje, sino también el archivo de registro (*logfile*) en el que las intervenciones de distintos hablantes aparecen en orden cronológico, y el hilo (*thread*), en el que se combinan la estructuración cronológica y temática. La unidad básica mensaje se anota mediante el elemento <post>[54] (Beißwenger y Lüngen 2020) que las directrices TEI para anotar textos escritos u orales no prevén[55]. El *post* puede designar una unidad monológica, aislada, o bien una unidad dialógica, si se trata de una respuesta o una reacción a otro *post*, aspecto que tiene en cuenta la dinámica conversacional observada en buena parte de los entornos digitales.

Estas propuestas pueden resultar eficaces para reflexionar sobre los elementos que se pueden anotar en los datos de un corpus de discurso digital en el nivel macroestructural. En particular, el elemento <post> facilita la anotación de muestras desde el punto de vista de la estructura de la interacción, utilizando

[53] En Beißwenger *et al.* (2012, 6), se consideran los textos de CMC que cumplen con los siguientes criterios: [they are] (i) based on the TCP/IP protocol suite for data exchange, (ii) dialogic (with all participating users being able to switch between the role of a recipient/reader and the role of a producer/author of messages), and (iii) based on writing as the main encoding medium for the users' dialogue contributions (that is, the verbal parts of the contributions must be encoded using writing, though they may also include graphics, embedded audio, or video files)».

[54] De acuerdo con estos autores, «None of the elements provided in TEI P5 can serve as an adequate model for the representation of posts [in CMC], even though several elements, at least at first glance, may appear to be a practical solution for the description» (Beißwenger y Lüngen 2020).

[55] Sí incluyen el elemento párrafo <p> o división , el de enunciado <u> (por *utterance*, en inglés) o habla <sp> (*speech*).

atributos propios de ese elemento en la TEI, como @who, @when y @from, que remiten al autor del mensaje, la fecha y hora de publicación y el remitente (Pano Alamán y Muñoz Moya 2015, 125).

En cualquier caso, dada la complejidad de los contextos digitales en los que interactuamos cotidianamente, parece necesario ampliar los descriptores relativos a la macroestructura, atendiendo, por ejemplo, a la temática o al objetivo comunicativo, en un específico entorno; y considerar las relaciones que pueden establecerse entre los *post* o los mensajes que se insertan en un determinado *thread* o hilo de discurso, para obtener una representación detallada de las dinámicas interactivas.

Respecto al primer punto, cabe considerar el contexto comunicativo y las propiedades enunciativas que definen cada situación, el tiempo o tiempos de la interacción o el tipo de relaciones interpersonales y sociales que se instauran en ese contexto. Describir los parámetros tecnológicos y sociosituacionales a los que hacíamos referencia en la sección 1 del primer capítulo, contribuye a enriquecer la representación de la macroestructura de los textos recogidos. Por ello, puede ser útil establecer un esquema abierto (Fig. 16) que recoge y describe los posibles parámetros asociados al tipo de plataforma o aplicación que se analiza y al tipo de situación comunicativa considerada:

Tabla 1. Factores o parámetros descriptivos

Tipo de factores	Característica	Definición
Factores del medio (aplicación/ plataforma)	Sincronicidad	Conexión al mismo tiempo entre los usuarios
	Transmisión del mensaje	La transmisión se realiza mensaje por mensaje o carácter por carácter
	Persistencia de la transcripción	El tiempo que se mantienen los mensajes almacenados
	Tamaño del mensaje	Número de caracteres que el sistema permite en un único mensaje
	Modos de comunicación	Modos de comunicación (texto, audio, vídeo, imágenes, etc.)
	Mensajes anónimos	Alguien con identidad desconocida puede contactar con otro usuario
	Mensajes privados	Un usuario puede contactar en privado con otro
	Filtración	Hay administradores que filtran los mensajes
	Citación	Los mensajes se pueden incrustar para facilitar la interacción
	Formato de los mensajes	Determina la presentación visual de los mensajes

Tipo de factores	Característica	Definición
Factores situacionales	Estructura de la participación	Número de participantes
	Características de los participantes	Características ideológicas o demográficas
	Propósito	Metas de la interacción
	Tópico	Tema
	Tono	Formal o informal
	Actividad	Acción llevada a cabo
	Normas	Prácticas establecidas aceptadas por el grupo
	Código	Lengua o variedad de lengua

Fuente: Pano Alamán y Moya Muñoz, 2015.

Respecto al segundo punto, cabe tener en cuenta que existen muy diversas tipologías de *post*: *mensajes* en *chats* o en *foros*, *correos electrónicos, publicaciones, tuits, entradas, comentarios, wasaps, tiktoks*, *vídeos en directo*, etc. La definición que se propone en el modelo <post> se centra en los procesos de producción y envío del mensaje.

En relación con el proceso de producción-envío-recepción del mensaje, asociado al tipo de plataforma que se empleaba para comunicar, Herring (2001) proponía hace años un esquema basado en dos tipos de transmisión: la transmisión *one-way*, que remite al envío de un mensaje cuando está completo y a la lectura del mismo tras su recepción por parte del destinatario (como en el correo electrónico o los foros); y la transmisión *two-way*, relacionada con la posibilidad de que el texto escrito aparezca gradualmente en la pantalla (como en los canales de chat), de tal manera que quienes participaban en la interacción podían seguir la elaboración de un mensaje y asistir a los cambios repentinos de tema o a las autocorrecciones (Pano Alamán 2008, 29-30). Con algunas diferencias determinadas por la interfaz, esta dinámica se observa también en WhatsApp, cuyo sistema informa de que el interlocutor está «escribiendo…» o «grabando audio…». De todos modos, es posible visualizar el mensaje únicamente cuando el emisor lo publica haciendo clic en *Enter* o activando el botón «Enviar». Desde la perspectiva del destinatario, un *post* es por tanto un texto que ha sido compuesto de antemano y que se presenta en bloque en su pantalla; las partes que lo conforman y los modos que se integran en él aparecen simultáneamente como un elemento visual que ocupa un espacio dentro de

la interfaz de la aplicación o de la plataforma que se está utilizando, en la pantalla del móvil, de la tableta o del ordenador.

El mensaje puede formar parte de un enunciado concluido que da la posibilidad de ser contestado o no, y que no requiere de otros mensajes del mismo emisor para completarse, o bien de un enunciado más extenso, de un discurso que se desarrolla por medio de varios mensajes del mismo emisor que se van acumulando en sucesión en la pantalla.

A partir de esta idea, el *post* puede definirse como una unidad comunicativa completada, textual, escrita o multimodal, elaborada en su totalidad por un emisor y enviada posteriormente por el mismo a un servidor en bloque. El receptor visualiza la unidad completa en la interfaz de la plataforma o del sistema en el que interactúa y en la pantalla del dispositivo que utiliza. Presenta distintas formas, como puede ser la nube o el bocadillo en WhatAspp o la línea o párrafo en los comentarios en la prensa o en YouTube.

Para completar esta caracterización, es necesario precisar que la unidad propuesta en este modelo no se corresponde con el concepto de enunciado. Especialmente en los sistemas de mensajería instantánea, pero también en Twitter/X o en los comentarios a las noticias, un enunciado verbal coherente desde el punto de vista temático y que persigue un mismo objetivo comunicativo puede estar formado por más de un *post* producido y publicado por una misma persona en secuencia, en momentos distintos de envío al sistema. Como se ha dicho también, el *post* puede ser una unidad textual aislada (piénsese en una entrada en un blog o en un correo que no requiere una respuesta), o bien formar parte de una secuencia de mensajes, correspondientes a las intervenciones de distintos hablantes en una interacción. Asimismo, puede constituir una reacción a un *post* precedente, como respuesta a una intervención ajena, por ejemplo, en los comentarios a una misma noticia en un diario digital o en una red social.

Si se asume que la comunicación digital es esencialmente dialógica, es posible considerar un *post* como una intervención iniciativa (entrada en un blog, noticia en un diario digital, publicación en Facebook), que puede ir seguida de otros *post* (comentarios, respuestas, citas, reacciones como *me gusta* o *compartir*, entre otras), como intervenciones reactivas. En este sentido, para poder representar y describir en profundidad el elemento <post>, así como las relaciones de dependencia que se establecen entre distintos mensajes de un usuario o de varios en un mismo contexto, parece adecuado incluir en la propuesta de etiquetado los atributos recogidos en la TEI que se relacionan con ese elemento, como son el autor (@who) del mensaje, o el tiempo y espacio de la interacción (@next o @prev).

En concreto, este atributo permite representar las dos dimensiones en las que se sitúa un mensaje, enunciadas en Beißwenger *et al.* (2012, 13): «the above/below dimension, which usually stands for a temporal before/after relation; the left/right dimension, in which one can use indentation to emphasize the topical affiliation of one message to a previous message».

En los corpus de CMC etiquetados por los miembros del SIG TEI CMC siguiendo el esquema CMC-Core, se emplean numerosos atributos de TEI P5 para caracterizar el <post>, como @who y @type, entre otros. Véase, por ejemplo, la Figura 17, que contiene una anotación adaptada a CMC-core de dos <post> extraídos de una interacción en WhatsApp, uno oral y otro escrito en el que se ha incluido un emoji (type), elaborados por dos interlocutores distintos (who):

```
<post mode="spoken" creation="human" synch="#t003" who="#A05"
    xml:id="m7"> Sagt Anne auch gerade. JA! Kann ich zustimmen. </post>
<post mode="written" creation="human" synch="#t003" who="#A02"
    xml:id="m8"> Da kostet ein Haarschnitt 50 € <figure type="emoji"
        creation="template">
        <desc type="meaning">face screaming in fear</desc>
        <desc type="unicode">U+1F631</desc></figure>
</post>
```

Figura 16. Anotación TEI de mensajes de WhatsApp en alemán (Beißwenger y Lüngen 2020, 18).

El esquema prevé también el atributo @mode, que anota la distinción entre publicaciones escritas y habladas, por ejemplo, en una secuencia de mensajes escritos y de audios en WhatsApp. Otros atributos útiles en este tipo de mensajes son: @replyTo, que se emplea para anotar a qué mensaje responde un determinado <post> o indicar que se trata de una «respuesta técnica», como activar un botón de respuesta[56]; e @indentLevel, atributo que marca la colocación del mensaje en cuestión dentro de la estructura de un hilo.

En cuanto a la dimensión microtextual, la anotación se focaliza en los rasgos propiamente lingüísticos de los mensajes del corpus. Sin embargo, en este caso, la diversidad de plataformas y propósitos de uso sobrepasan cualquier caracterización lingüística *a priori* de los datos, aunque los textos que componen nuestro corpus de discurso digital presenten rasgos comunes, como la tendencia a la coloquialización en los niveles ortográfico, morfosintáctico y léxico.

[56] Se relaciona más bien con el concepto de «reacción» mencionado más arriba, más que con respuestas interpretadas o inferidas y basadas en señales lingüísticas o marcadores.

Versiones anteriores del esquema CMC-core, aplicadas a proyectos como CoMeRe (2014) para el francés[57], consideran en el nivel micro de los mensajes tanto los elementos lingüísticos contenidos en los enunciados, como los datos relativos a las actividades que se llevan a cabo en la pantalla (*onscreen activities*), que pueden capturarse, como decía, mediante pantallazos, y anotarse utilizando la amplia gama de modelos proporcionados por TEI P5. El objetivo de esta propuesta es ofrecer a quien investiga sobre la comunicación digital en español un modelo estándar de anotación que pueda aplicarse a una amplia gama de cibergéneros. Dependiendo del grado de detalle con el que queramos anotar los datos de nuestro corpus tanto en los niveles macro como microtextual, será posible adaptar los elementos y atributos de TEI o proponer otros, de acuerdo con sus pautas de personalización[58].

3. Me temo que no puedo hacer eso: cuestiones éticas

Casi todos recordamos la escena final de *2001. Odisea en el espacio*, cuando Hal 9000, superordenador a bordo de la nave espacial *Discovery*, contesta de forma glacial a su interlocutor diciendo: «Lo siento, Dave, me temo que no puedo hacer eso». Hal parece tener una conciencia y, de hecho, los motivos que aduce para no permitir el acceso a la nave a Dave tienen que ver más bien con la voluntad de no abrir la puerta más que con la posibilidad de hacerlo. Retomo esa célebre frase para recordar que en la investigación sobre discurso digital y, en general en cualquier tipo de investigación, hay que considerar tanto lo que queremos hacer como lo que podemos o no podemos hacer.

El rápido y constante desarrollo sociotecnológico, la existencia de grandes cantidades de datos, textos, imágenes, vídeos y datos personales disponibles en la web (*Big Data*) obliga a replantearse las cuestiones éticas relacionadas con nuevas prácticas sociales y formas de participación inéditas hasta ahora. Algunas interacciones en línea se pueden observar, capturar y almacenar sin estar sujetas a la paradoja del observador, esto es, sin que los participantes sepan que el investigador está presente; o bien, este puede hacerse presente en algunos entornos, como se apuntaba a propósito de WhatsApp, para obtener datos adicionales que pueden enriquecer el análisis.

[57] Información sobre el proyecto *Communication Mediée par les Réseaux* (CoMeRe) disponible en: https://wiki.tei-c.org/index.php?title=SIG:CMC/CoMeRe_metadata_schema_draft_for_CMC_(2014)

[58] Disponibles en: https://tei-c.org/guidelines/customization/

Se ha dicho también que es aparentemente sencillo extraer datos de las plataformas y las aplicaciones, sin embargo, hay que tener en cuenta que esos datos no nos pertenecen, de ahí que sea necesario preguntarse si es ético utilizarlos sin el consentimiento de las personas que los crean y los publican en la red. Asimismo, aunque los contenidos son a menudo de fácil acceso, cabe considerar su naturaleza efímera y el hecho de que los algoritmos que determinan la interacción en muchas plataformas y aplicaciones hagan visibles solo algunos datos[59]. Esto incide en la toma de decisiones con respecto a los procedimientos de recopilación. Por ello, antes de recuperar los datos que interesa analizar, es necesario establecer ciertos criterios para su recopilación, por ejemplo, identificando las palabras clave, temas, *hashtags* o cuentas de usuario que pueden ser relevantes, el periodo de tiempo de extracción de los datos o si es necesario descargar y archivar metadatos asociados a la ubicación geográfica de un usuario o a su participación. Un análisis multimodal requerirá una muestra de datos diferente y, por lo tanto, también diferentes métodos de recopilación. Un plan de gestión de datos también debe incluir consideraciones con respecto a las cuestiones éticas y legales.

En este sentido, Thompson (2022) señala que el investigador debe reflexionar sobre lo que puede hacer en las distintas fases en las que lleva a cabo su estudio. El análisis de naturaleza etnográfica y multimodal que el autor realiza sobre las prácticas comunicativas de los usuarios en aplicaciones de encuentros como Tinder se apoya en entrevistas reflexivas y «de acción dialógica» con personas que se citan en línea. Esto prevé un procedimiento introspectivo en el que se pide a los participantes que verbalicen sus reflexiones respecto a la toma de decisiones sobre los encuentros en los que participan, en presencia del investigador. La entrevistas se realizan adaptando los criterios éticos de las ciencias sociales, en función de los datos de su contexto de investigación y de sus participantes. El análisis de los textos y otros elementos multimodales, junto al análisis de las respuestas de los usuarios, amplía el alcance del estudio, aunque esto implica una mayor intervención por parte del investigador y una mayor atención a los aspectos éticos.

Más allá de adherirse a las prácticas éticas regulatorias en este y otros ámbitos de investigación en ciencias sociales y humanas, cabe desarrollar una

[59] La implantación de una web cada vez más personalizada y localizada, adaptada a medida para cada uno de nosotros, obliga a lidiar con la realidad de que los datos que se presentan a los investigadores no son neutrales. No existe una versión neutral y estándar de los contenidos a los que accedemos, por lo que el muestreo de los datos en línea debe considerar no solo las elecciones explícitas de los investigadores, sino también las decisiones implícitas que subyacen a las tecnologías.

«alfabetización ética» (Spilioti y Tagg 2022), que lleve a reflexionar sobre las dimensiones éticas de cualquier proyecto, sobre todo si se trabaja con datos sensibles, por ejemplo, si consideramos las interacciones entre menores en TikTok o entre un médico y sus pacientes en redes como Facebook. Esto significa que los analistas del discurso digital, cada vez más presentes en proyectos financiados a nivel nacional y europeo, deben adherirse a los protocolos de ética nacionales e institucionales, a la vez que adoptan un método flexible de toma de decisiones que no se limite a las etapas iniciales del diseño de la investigación.

De acuerdo con las pautas éticas propuestas por la Asociación de Investigadores de Internet (AoIR por sus siglas en inglés, 2019)[60], es conveniente adoptar un enfoque amplio y plural de los procesos de reflexión ética atentos al contexto que se investiga, incluidas la lengua y la cultura en que se lleva a cabo el estudio, en lugar de prescribir enfoques que abarquen todas las situaciones posibles. Así, los principios generales pueden evaluarse y aplicarse en contextos de investigación particulares. En otras palabras, la ética de la investigación es «una cuestión de método» (Markham *et al.* 2018) que no se limita al diseño inicial del proyecto, sino que concierne también la definición del contexto de estudio, la identificación de los participantes, la recopilación y la anotación de los datos, su análisis, publicación y difusión.

Respecto a la cuestión compleja del tratamiento de los datos relativos a los participantes, se ha propuesto aplicar una «minimización de los datos» (Ess y Hård af Segerstad 2019), procedimiento previsto por el Reglamento General de Protección de Datos de la Unión Europea (RGPD 2016/679, 27 de abril de 2016). Esto implica recopilar solo los datos que se necesitan para dar respuesta a las preguntas de investigación, preguntas que deben definirse desde el principio para evitar recoger datos innecesarios.

Otra cuestión importante en este ámbito concierne la posibilidad de acceder y extraer los datos sobre la identidad de los hablantes. En lo que se refiere a la privacidad, Kendall (1999) fue una de las primeras investigadoras en poner de manifiesto cómo en el análisis de la comunicación digital cabe entender el concepto de «identidad» de otro modo, ya que la distinción entre público y privado se difumina. De hecho, en algunas plataformas en línea se acepta que haya un cierto nivel de «acecho encubierto», no declarado, de la identidad pública o «mostrada» de forma más o menos consciente por los internautas[61]. Y es que el

[60] Disponibles en https://aoir.org/reports/ethics3.pdf

[61] Es más, en la Web encontramos un continuo entre lo privado y lo público, pues las expectativas del usuario en cuanto a su privacidad y visibilidad no son las mismas en WhatsApp, Facebook, Twitter/X o YouTube.

carácter fugaz e intermitente de la participación en las redes sociales o en los comentarios en foros o en prensa, por ejemplo, puede hacer inviable que un investigador indique su presencia a todos los participantes (Barbosa y Milan 2019). Muy distinto es el caso de aquellos espacios en los que puede producirse un tipo de interacción susceptible de constituir un delito (Lorenzo-Dus 2022), en los que es necesario que el investigador no se identifique.

Desde un enfoque más amplio, De-Matteis (2014) proporciona una serie de pautas sobre el tratamiento ético de los datos considerando tres ejes relativos a la relación entre investigador e investigado (eje interaccional); la relación entre sujeto investigado, tema y contexto comunicativo (eje de privacidad); y la relación entre la durabilidad del dato, su indexación y la existencia de metadatos (eje de trazabilidad). No obstante, en la mayor parte de investigaciones se determina que el tipo de consentimiento necesario radica en la diferencia entre entornos públicos y privados (De Benito Moreno 2022), o, según la distinción de Page *et al.* (2014), entre públicos, semipúblicos, semiprivados y privados.

El principio general es que cuanto más público sea el entorno y más abierto sea el acceso a él, menos necesidad habrá de proteger la privacidad de los participantes. En todo caso, esto implica prestar atención a las características del entorno y al tipo de contenidos que se publican, teniendo en cuenta que será conveniente anonimizar los datos personales, como suele hacerse cuando se explora la comunicación en interacciones privadas (correo electrónico, WhatsApp) y (semi)públicas (foros, redes), en las que el acceso requiere, en muchos casos, una inscripción previa mediante la creación de un perfil o una cuenta, y establecer un contacto con otros usuarios para poder visualizar sus contenidos e interactuar con ellos. De hecho, tanto en estudios basados en la observación participante en redes públicas de acceso restringido como Facebook como en intercambios privados en WhatsApp, se solicita el consentimiento informado y se anonimizan los datos personales (Vela Delfa 2017; Mapelli 2019).

Otro tipo de proyecto en el que es viable utilizar los datos relativos a la identidad de los participantes es el que se ocupa de la comunicación en línea de quienes desempeñan una actividad pública (políticos, periodistas, personajes célebres, influyentes), cuya identidad es relevante para la interpretación de los contenidos publicados (Mancera Rueda y Pano Alamán 2013b; Gallardo Paúls y Enguix Oliver 2016).

La anonimización de los datos asociados a la información personal de los participantes no impide completamente el acceso a ese tipo de información. En efecto, en el contexto digital, donde todo queda registrado, es posible rastrear

mensajes, nombres de usuario, fotografías e incluso voces. Por ello, las pautas éticas contenidas en el RGPD (2016) y abordadas también en el documento de la AoIR recomiendan no solo evitar la mención a la información personal relativa a los participantes, sino también obtener su consentimiento informado. Como he señalado, existe el riesgo de que esos textos o esas imágenes puedan rastrearse después en el contexto original en línea, incluso en espacios privados con cifrado de extremo a extremo, accediendo así a datos inicialmente anonimizados (Barbosa y Milan 2019, 58). En algunos casos, el riesgo puede evitarse si se renuncia a las citas directas mediante una «fabricación ética» de los datos (Markham 2012) que prevé, por ejemplo, proteger la identidad de los participantes difuminando algunas palabras o manipulando las imágenes.

Por otro lado, cabe asumir que en un proyecto, tras obtener el consentimiento firmado por escrito al inicio de la investigación, es aconsejable obtener, incluso informalmente, un segundo consentimiento o autorización verbal relativo a la reproducción, en artículos, conferencias y otros trabajos, de extractos de texto o de imágenes extraídas o elicitadas.

Asimismo, cuando tratamos datos de comunicación digital, tanto si provienen de entornos públicos como privados, hay que tener en cuenta que una cosa es la autoría del mensaje y otra la propiedad de los datos publicados, que están en manos de la compañía que posee la plataforma o la aplicación. En ocasiones, ante posibles problemas de *copyright* de ciertos contenidos, se puede solicitar previamente por escrito a la compañía la autorización a utilizarlos para fines de investigación, además de citar la fuente de donde se han extraído.

En general, la mayor parte de los proyectos se apoya en corpus cerrados con objetivos específicos y prevé una difusión de los datos circunscrita al ámbito académico, adaptándose a los estándares éticos vigentes en la recolección y publicación, en función de los objetivos del proyecto. Por ejemplo, la plataforma Marco Polo que se ha mencionado previamente, hace referencia a la Ley de Protección de Datos y a la nueva normativa vigente, así como a las normas que impone Twitter/X sobre la redistribución de los tuits (Alcántara-Plá *et al.* 2018, 9). Algunas cuestiones a las que debemos dar respuesta pasan por definir los criterios de selección de los participantes; las formas de consentimiento informado adaptadas a las normativas, teniendo en cuenta los términos y condiciones de uso de las plataformas y aplicaciones; cómo garantizar la integridad de los datos en los procesos de extracción, almacenamiento y difusión a través de la nube o en servidores remotos. En este sentido, la legislación europea vincula hoy en día los proyectos basados en corpus de comunicación digital para los que se dispone ya de listas de buenas prácticas (Collins 2019).

Estas reflexiones y las propuestas enunciadas indican el camino para adoptar una ética investigadora en el ámbito del discurso digital en español y la creación de estándares para que su estudio se haga de manera responsable y respetuosa con los usuarios y, en general, con los hablantes. Sin embargo, resulta paradójico que las grandes compañías tecnológicas tengan acceso a nuestros datos personales, mientras que el acceso a la investigación sea cada vez más limitado como consecuencia de las políticas restrictivas de acceso a las APIs de la mayor parte de plataformas y aplicaciones.

El público no solo tiene derecho a saber qué datos suyos se registran y se archivan o con qué fines se emplean, sino que debe estar protegido legalmente ante el uso ilegítimo de sus datos, como se recordaba en Mancera Rueda y Pano Alamán (2020) en relación con el escándalo de Cambridge Analytica, cuando se supo que la empresa había recopilado datos personales de millones de usuarios de Facebook sin su consentimiento y con fines electorales no declarados.