El *big data* en los estudios del lenguaje

Javier Valenzuela
Universidad de Murcia
jvalen@um.es

Resumen

El presente trabajo examina las posibilidades que los acercamientos basados en los *big data* ofrecen a la investigación sobre el lenguaje. De manera resumida, los *big data* o "macrodatos" son los datos masivos que los usuarios generan en sus interacciones con el mundo digital y cuyo ingente volumen y naturaleza heterogénea precisa de un tratamiento especializado. El trabajo revisa de manera inicial las principales características de los *big data* para centrarse a continuación en los posibles problemas derivados del uso de *big data* en los análisis lingüísticos. La siguiente sección ofrece una revisión de estudios concretos que utilizan este acercamiento aplicándolo a la multimodalidad: un estudio del lenguaje que incluye no sólo el componente verbal sino aspectos multimodales como la gestualidad o la entonación. El trabajo concluye con una revisión de las ventajas y los problemas de la utilización de este tipo de datos.

Palabras clave: lingüística; análisis de corpus; *big data*; multimodalidad.

Abstract

This paper examines the possibilities that big data-based approaches offer to language research. In a nutshell, the term "big data" makes reference to the massive amount of data that users generate in their digital interactions and whose great volume and heterogeneous nature typically requires a specialized treatment. The chapter starts by reviewing the main characteristics of big data, and then focuses on the possible problems arising from the use of big data in linguistic analysis. The following section offers a review of specific studies that apply this big-data approach to the study of multimodality: an approach to language study that includes not only the verbal component but also multimodal aspects such as gestures or intonation. The paper concludes with a review of the advantages and problems of using this type of data.

Keywords: linguistics; corpus analysis; big data; multimodality.

1. Introducción

El término *big data*, traducido a veces como "macrodatos" o "datos masivos", aparece mencionado con gran frecuencia en los medios de comunicación; su uso está extendido por multitud de ámbitos que afectan nuestro día a día, a pesar de lo cual existe un cierto nivel de desconocimiento sobre su naturaleza y su tratamiento. Básicamente, se denomina *big data* a toda la información que se genera al utilizar los distintos dispositivos digitales (internet, teléfonos, tarjetas de crédito, etc.). Esta información es recopilada y utilizada por empresas (o incluso por agencias gubernamentales), que basan sus decisiones en el conocimiento aportado por este tipo de datos personales. Dada la extensión de su uso y

DOI 10.36950/elies.2022.45.8857
ISSN: 1139-8736 https://bop.unibe.ch/elies

su gran efectividad para determinadas predicciones, la pregunta obvia es hasta qué punto son útiles estos datos masivos para actividades de corte más académico, y más específicamente, cómo de valiosos son para los estudios del lenguaje. Este trabajo revisa el concepto y el tratamiento de los *big data*, discute su aplicación al ámbito de los estudios lingüísticos y finaliza con una descripción de los acercamientos que intentan aplicar de la manera más efectiva posible las principales características de los *big data* a las ciencias lingüísticas. Para ello nos centramos en los acercamientos que abordan el estudio del lenguaje desde el punto de vista de la multimodalidad, es decir, entendiendo la comunicación lingüística en toda su complejidad, prestando atención no solo a las palabras que decimos sino a su entonación, a los gestos que realizamos, a la información transmitida con los ojos y la mirada, las expresiones faciales o en general cualquier tipo de información física presente en el acto comunicativo. El trabajo discute algunas de las oportunidades así como de los desafíos inherentes a la aplicación del concepto de *big data* a los estudios del lenguaje.

2. *Big data*: algunos conceptos básicos

Sin lugar a dudas, el anglicismo *big data*, traducido aproximadamente como "grandes datos", "macrodatos" o "datos masivos", es uno de los términos más usados, buscados y estudiados del momento. Una búsqueda en Google devuelve al menos 120 millones de resultados. Al parecer, estamos en la era del *big data*, que cobra una importancia cada vez mayor en nuestras vidas. Pero ¿qué es exactamente el *big data*?

El término *big data* hace referencia al conjunto de datos que generamos en nuestra interacción con el mundo digital, que claramente abarca un porcentaje amplísimo de nuestra actividad cotidiana. Es nuestro uso de los distintos dispositivos electrónicos modernos (lo que podemos denominar nuestra "interacción digital") lo que sirve como fuente de estos datos. Así, todos los aparatos electrónicos conectados a la red recopilan información puntual y precisa sobre el uso que hacemos de ellos. Por ejemplo, los servidores de correo electrónico recogen estadísticas de cuáles son los momentos en los que escribimos un correo (o un número mayor de ellos), a quién le escribimos, sobre qué temas escribimos (el asunto de la privacidad es un tema espinoso, que trataremos más abajo) y qué programa utilizamos para ello. Pero no solo nuestros correos electrónicos: nuestras búsquedas en la red (por ejemplo, en Google) quedan registradas y generan información sobre cuándo y qué buscamos (consultas de opiniones sobre un producto, una película, un restaurante, un hotel, un concierto, un libro, la búsqueda de un determinado tema, etc., información que luego sirve para devolvernos publicidad dirigida a nuestros intereses específicos), desde qué dispositivo –móvil, tableta, ordenador, portátil o incluso en qué lugar estamos al realizar la búsqueda–; el uso del GPS en nuestros móviles (el uso de Google Maps, por ejemplo) o en los navegadores de nuestros coches genera información detallada espacial y temporal sobre nuestros desplazamientos; el pago con tarjeta de crédito informa de cuánto hemos gastado, cuándo, dónde y en qué; el pago de facturas informa sobre nuestro consumo y nivel de gasto, así como las transacciones bancarias por internet; las llamadas de teléfono informan de a qué horas llamamos con mayor frecuencia, cuánto duran esas llamadas o con quién nos comunicamos; qué películas vemos en Netflix, HBO, AmazonPrime, Movistar+, Disney+, Apple TV... La lista es realmente casi interminable. Y, por supuesto, no podemos olvidar nuestro uso de las redes sociales. Las redes sociales son un fenómeno absolutamente desbordante, que

nos permite contactar con otros usuarios para comunicarnos con ellos, informarles de nuestros gustos, hacer amistades, compartir experiencias, ligar, jugar, encontrar trabajo, enterarnos de opiniones de un tema, escuchar música, buscar alojamiento, compartir aficiones... Las redes sociales están por todos sitios y su número sigue creciendo, tanto en diversidad como en número de usuarios. Algunas de ellas son archiconocidas y omnipresentes (Facebook tiene 2800 millones de usuarios; recordemos que la población mundial se estima en 7500 millones de habitantes), otras redes son más especializadas. Además de Facebook podemos mencionar WhatsApp, YouTube, Twitter, Instagram, TikTok, Telegram, Vimeo, Linkedln, Pinterest, Tuenti, Snapchat, Tinder, Yelp, Google Maps, Waze, WeChat, Skype, Fortnite, World of Warcraft, Tumblr, Reddit, Viber, Line, Spotify, Google +, Flickr, Slideshare, Soundcloud, Badoo, TripAdvisor, AirBnB, Foursquare, Quora... además de las usadas en China y menos explotadas en Occidente. Nuestro uso de estas redes sociales también genera una gran cantidad de información, específica de cada caso. El volumen de datos que se genera es tan absolutamente ingente que resulta difícil de conceptualizar, de ahí la denominación de esta información como *big data*. Este volumen de información supera la capacidad del *software* convencional para poder capturarla, administrarla o procesarla en un tiempo razonable. De todas maneras, lo que hace especialmente complicado el tratamiento de estos datos no es solo su volumen, sino su velocidad de crecimiento (se generan datos de manera continua a cada momento, véase Figura 1) así como su heterogeneidad (hay datos de muy distinta clase). Es por ello por lo que los estudios sobre *big data* suelen caracterizar estos datos con una serie de "V"s: Volumen, Velocidad, Variedad, Visualización, Valor, Viralidad y Veracidad.

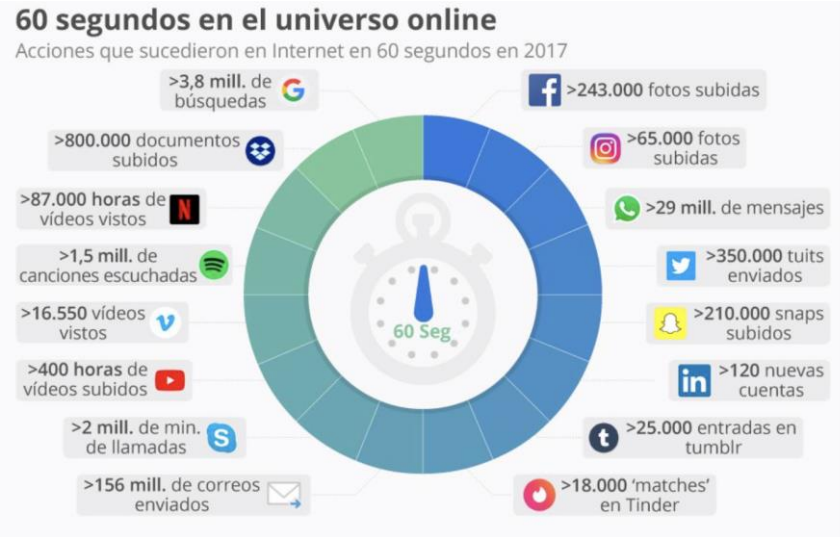


Figura 1. Datos generados en un minuto en el universo digital (https://es.statista.com/)

Y todo esto ocurre cuando todavía no ha llegado "El Internet de las Cosas", que conectará a la red de internet los objetos con los que interactuamos diariamente. Esta conexión convertirá nuestras casas en hogares inteligentes (la famosa nevera que te avisa sobre el estado de sus alimentos o la falta de algunos, pero también termostatos inteligentes que permitan optimizar el consumo de energía, o avisos sobre tus medicamentos caducados). Este nuevo estadio de información implica que ya no habrá nunca más objetos fuera de *stock* al ir a comprar; tampoco se perderán las cosas (imaginemos las llaves del coche o

de casa conectadas a internet y geolocalizadas en todo momento). Las aplicaciones de esta revolución en marcha se extienden a prácticamente todos los ámbitos de nuestra vida: la medicina y la salud, el transporte, los procesos industriales, el consumo y la economía de mercado y un larguísimo etcétera que probablemente no somos capaces de imaginar aún. Obviamente, esta revolución multiplicará las posibilidades (y los problemas) de los *big data* en un grado difícil de prever. Un cálculo conservador cifraba en 26000 millones los dispositivos conectados en 2020. A las enormes dificultades derivadas de la gestión de un tamaño tan ingente de datos se suma la variedad de estos. Aunque es complicado establecer categorías fijas, de manera aproximada los datos recogidos se pueden clasificar en tres tipos:

(i) Datos estructurados: son datos que tienen un formato y una estructura fija, como la que podríamos encontrar en una base de datos. Estos datos son fácilmente explotables, puesto que es sencillo encontrar la información buscada (véase a la Figura 2):

	nombre	color	edad	altura	peso	puntuacion
1:	Paco	Rojo	24	182	74.8	83
2:	Juan	Green	30	170	70.1	500
3:	Andres	Amarillo	41	169	60.0	20
4:	Natalia	Green	22	183	75.0	865
5:	Vanesa	Verde	31	178	83.9	221
6:	Miriam	Rojo	35	172	76.2	413
7:	Juan	Amarillo	22	164	68.0	902

Figura 2. Datos estructurados

(ii) Datos no estructurados: estos datos no responden a un patrón concreto; pueden ser los movimientos de ratón en una página web, el recorrido captado por el GPS en nuestros desplazamientos, datos de audio con grabaciones de voz, incluso los datos textuales (comentarios escritos por los usuarios en las redes sociales), etc. tal y como aparece en la Figura 3. Claramente, los métodos de recuperación de la información que pueden aplicarse a estos datos son mucho más complicados; cada categoría dentro de ellos requiere de un procedimiento distinto de tratamiento y de procesado para poder extraer la información requerida.

Figura 3. Ejemplos de datos no estructurados (ratón, audio, texto y GPS)

(iii) Datos semiestructurados: estos datos combinan una estructura más o menos fija con datos abiertos (véase Figura 4).

Figura 4. Ejemplo de datos semiestructurados

De todas las características de los *big data*, dos de ellas, su gran volumen y su variedad, son las principales responsables de la dificultad tanto de su almacenamiento como de su tratamiento y gestión (son precisamente estas dos características las más interesantes para los estudios de la lengua, como veremos más adelante). De esta manera, han surgido compañías que ofrecen esos servicios especializados (por ejemplo, Spark o Hadoop) tanto para el almacenamiento de estos datos (que suele realizarse en "granjas de servidores", lugares especializados que reúnen multitud de ordenadores que son capaces de almacenar enormes cantidades de información), como para su tratamiento.

El tratamiento de los datos merece también un comentario especial. Han surgido una gran cantidad de técnicas de explotación dedicadas a su procesamiento. Suelen ser acercamientos estadísticos, que detectan patrones generales subyacentes a esa gran cantidad de datos que son muy difíciles (por no decir imposibles) de observar a simple vista. Proliferan los cursos especializados en este tipo de procesamiento, que aumenta su paleta de herramientas de manera progresiva: sin ninguna intención de ser exhaustivo, se pueden mencionar los análisis de regresión, análisis de series temporales, A/B Testing (*split testing*), algoritmos genéticos o aprendizaje por máquina (*machine learning*). Más relacionado con el lenguaje está también el "Análisis de Sentimientos" (*Sentiment Analysis*), que permite extraer automáticamente la opinión de los consumidores sobre un producto (una película, un electrodoméstico, un restaurante...) de sus comentarios en las redes sociales. Ahora bien, en realidad, la gran mayoría de los datos masivos se procesan utilizando los modelos conocidos como "redes neuronales" artificiales (también conocidas como "redes conexionistas"). De hecho, para muchos autores, esta nueva ola de datos masivos es la que ha permitido el gran auge de las redes neuronales, cuyo uso se inició en los años 80 (culminando en la obra clásica de Rumelhart y McLelland 1986), y que han alcanzado su mayor grado de extensión y desarrollo en estos momentos. Las redes neuronales son programas de ordenador que utilizan una arquitectura distinta de los programas más clásicos o "algorítmicos". En estos programas algorítmicos, la información se trata de manera secuencial, como se hace típicamente en lenguajes de programación como Python, C++, Java o Javascript. De esta manera, la información va siguiendo una serie de "pasos" secuenciales, con determinados saltos condicionales ("si se cumple esta condición, haz tal cosa; si no, sigue por la siguiente instrucción"). Sin embargo, las redes neuronales funcionan de manera muy distinta: estos programas están formados por una serie de "nodos", cuyo funcionamiento está inspirado en el comportamiento de las neuronas del cerebro. De esta manera, cada nodo puede recibir distintos grados de "activación"; esta activación se pasa en paralelo a todos los otros nodos con los que esté conectado. Existen distintas "capas" de nodos; en las redes multicapa, representadas en la Figura 5, al menos existe una capa de entrada (*input*), una capa de salida o respuesta y una o más capas "intermedias". Uno de los modos de funcionamiento más populares requiere dos estadios. En el primero, o fase de entrenamiento, las redes intentan conectar, por medio de un proceso de prueba y error, un estado de entrada (el *input*) a una respuesta deseada (el *output*) y lo hacen modificando poco a poco la fuerza con la que un nodo se conecta al siguiente (lo que se conoce como los "pesos" de la red), hasta que se alcanza el resultado deseado. Una vez la red ha sido "entrenada" y consigue el resultado deseado con el set de datos de entrenamiento, esa red ha "aprendido" y es capaz de encontrar la solución correcta a casos nuevos, que no formaban parte de su set de entrenamiento, es decir, es capaz de generalizar, que es lo que hace tan útiles y poderosas a estas redes. Por ejemplo, se alimenta una red con una serie de fotos de lunares (que sería el *input*), unos de ellos cancerígenos y otros benignos

(esta clasificación sería el *output* buscado). Una vez que la red ha sido entrenada de manera adecuada, y ha alcanzado la configuración de pesos o conexiones que permite distinguir entre ambos tipos de lunares, se le pueden dar una nueva serie de datos (lunares que no ha visto nunca) y será capaz de predecir con un altísimo nivel de confianza si los lunares son cancerígenos o no.

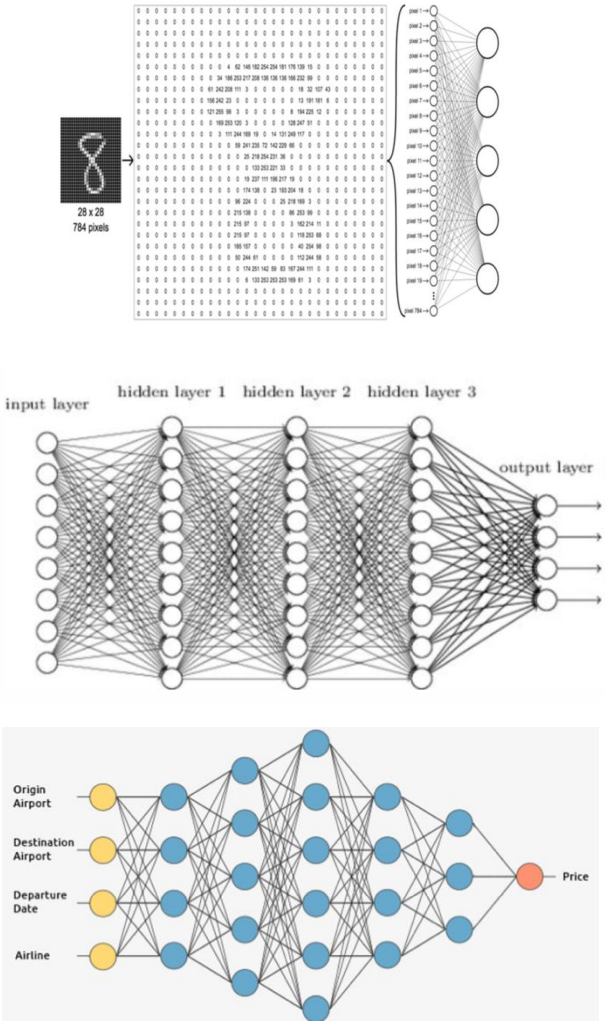


Figura 5. Ejemplos de redes neuronales multicapa

Las redes neuronales son herramientas muy efectivas; una vez entrenadas con una gran cantidad de datos, son capaces de generalizar con un alto grado de éxito. Sin embargo, tienen un problema: las capas intermedias (también conocidas como "ocultas" o *hidden layers*) no permiten saber exactamente cómo se produce el proceso de generalización. Por ejemplo, en nuestro caso de selección de lunares cancerígenos frente a benignos, es posible que la red acierte, pero no se sabe bien cómo realiza su elección, ni en qué tipo de parámetros se basa. En este ejemplo concreto, no parece un problema muy grave, pero imaginemos una red cuyo fin sea decidir sobre la concesión o no de préstamos bancarios: tras ser entrenada con multitud de casos basados en una gran cantidad de parámetros distintos, aprende a distinguir entre personas que han sido capaces de devolver un préstamo bancario y aquellas que no lo han sido. Una vez el programa ha sido entrenado para distinguir entre estos casos reales, puede usarse con nuevos casos para predecir si alguien tiene una mayor probabilidad de devolver o no devolver un préstamo; en caso

negativo, el banco puede denegar entonces la solicitud del cliente. Pero si le preguntan al banco por las razones del rechazo, no serán capaces de explicitarlas: la información específica que ha usado la red para tomar la decisión está distribuida en los pesos de conexiones de la capa oculta y no hay manera de saber cuáles han sido las razones concretas. Lo mismo sucede con un programa que diga a qué prisioneros se debe conceder libertad condicional o no según su experiencia con casos previos. Si el programa la aprueba o la deniega no hay manera de saber por qué lo ha hecho. Además de los problemas éticos, de este tipo y otros (se ha demostrado que los programas reproducen los prejuicios y sesgos implícitos en los datos con los que han sido entrenados, por ejemplo), esto obviamente va a ser uno de los obstáculos para utilizar el tratamiento de datos masivos con redes neuronales como herramienta para averiguar el funcionamiento de sistemas complejos. En cualquier caso, y como acabamos de decir, el uso de los *big data* es un tema altamente controvertido, por otras muchas razones adicionales. Ya hemos mencionado el espinoso problema de la privacidad. En 2016 surgió una gran controversia al descubrirse que agencias del gobierno de Estados Unidos (la NSA o *National Security Agenc*y o el FBI) habían accedido a millones de correos electrónicos de compañías como Yahoo, por ejemplo, para escanear su contenido en busca de material comprometido, con el objeto de combatir el terrorismo. Resulta ilusorio pensar que este tipo de prácticas han cesado; claramente, el poder de los *big data* puede ser utilizado para asuntos de gran utilidad, como este que mencionamos sobre actividades terroristas o peligrosas. Lo mismo podríamos decir de información relacionada con nuestra salud, que puede servir para diseñar programas sanitarios destinados a atajar determinados problemas. Sin embargo, a muchas personas les resulta inaceptable la utilización de datos personales por parte de agencias que no siempre exhiben el grado de transparencia adecuado. El gobierno de China recopila todo tipo de información sobre sus ciudadanos y ha llegado a establecer un "sistema de crédito social", que consiste en la monitorización continua del comportamiento de sus ciudadanos, puntuándolo y estableciendo un sistema de "recompensa" o "castigo". Se llega a extremos que resultan muy llamativos en las democracias occidentales: por ejemplo, las cámaras de reconocimiento facial pueden detectar a personas que crucen una calle con el semáforo en rojo, que pueden ser introducidas en una "lista negra" si son reconocidas en cinco ocasiones distintas en esta conducta incívica. Por supuesto, esto es un ejemplo extremo: es mucho más sencillo controlar violaciones de pago de impuestos, búsquedas consideradas inadecuadas en servidores, o comentarios en redes sociales, apoyo a causas críticas con el gobierno y un larguísimo etcétera. Estas recompensas o castigos pueden tener consecuencias como la facilidad o dificultad para conseguir un billete de avión o alojamiento en hoteles de prestigio, o el ingreso en un colegio u otro de tus hijos. Muchos advierten de que esta situación es muy similar al Gran Hermano de Orwell; toda esta información recopilada por el sistema está basada en estos datos masivos recopilados de manera digital. En Occidente no existe un programa explícito de manejo de los datos de esta manera, pero obviamente, la privacidad personal ha disminuido e incluso, según algunos, esta es ya un concepto ilusorio, perteneciente a un tiempo pasado.

3. El *big data* y los estudios lingüísticos

En principio, se tendería a pensar que esta revolución y las oportunidades que ofrece habrían sido bienvenidas sin ningún tipo de reticencia por parte de los lingüistas. No es

el caso, sin embargo; tal y como hemos mencionado, las dos características más relevantes de los *big data* (es decir, su volumen y su variedad) ofrecen por igual tanto oportunidades como desafíos. Examinemos ambas características por turnos.

3.1. El volumen de datos

Para empezar, en lo referente al volumen, algunos lingüistas han expresado un cierto escepticismo frente a estos estudios de grandes datos. Sus reservas vienen de varias fuentes. Por un lado, se argumenta que los lingüistas llevan ya un tiempo trabajando con *big data*, al menos, en lo referente al volumen de información. Mientras que en los años 1960-1970, corpus como el Brown o el LOB tenían un millón de palabras, en los años 80, el corpus de Birmingham/Cobuild contaba ya con 20 millones de palabras; así mismo, en los años 90, el British National Corpus contaba con 100 millones de palabras, lo que durante mucho tiempo fue considerado un estándar. Ya en el siglo XXI, The Bank of English cuenta con 645 millones de palabras y la familia de corpus TenTen ofrece tamaños mucho mayores aún: el TenTen inglés, por ejemplo, cuenta con 19.000 millones de palabras. Existen, además, otro tipo de críticas relacionadas con el volumen de datos. Esto es así porque en realidad no existe un acuerdo generalizado sobre la importancia que tiene el tamaño de un corpus. El asunto de cómo de grande debe ser un corpus es un tema muy controvertido. Existen opiniones absolutamente encontradas, como la de Krishnamurthy (2001) quien expresa claramente su opinión en el título de su artículo ("Size matters"), o la de Geoffrey Leech (1991: 10), quien opina que el tamaño no es tan importante ("Size is not all-important"). La clave de este desacuerdo se basa en la función que debe desempeñar un corpus o dicho de otro modo, la respuesta a la pregunta de hasta qué punto es posible generalizar, a partir de los datos encontrados en un corpus dado, acerca del funcionamiento general de la lengua. John Sinclair, una de las figuras fundacionales de los estudios de corpus lingüísticos (véase García-Miguel 2022), advertía de que era preciso distinguir entre "archivos" o "bases de datos lingüísticos" y "corpus" (Sinclair 1991). Una lista de palabras, un archivo de texto cualquiera, una recopilación de citas, por ejemplo, no formarían un corpus. La información textual de internet, según este autor, tampoco conformaría un corpus. Por ejemplo, los datos de frecuencia de uso en internet pueden ser muy engañosos: si buscamos en Google la frase *A long time ago in a galaxy far far away*, nos devuelve más de 800.000 resultados. Esta alta frecuencia no es representativa del uso de esta frase en el lenguaje en general, sino que se debe más bien a su inclusión en multitud de páginas que hacen referencia a su uso en la franquicia de la Guerra de las Galaxias. De una opinión parecida es Andrew Hardie, otra figura señera en la lingüística de corpus (autor de uno de los programas de procesamiento de corpus más completos y populares, CQPWeb), que, de hecho, se muestra muy crítico de entrada con los acercamientos de *big data*. Según este autor, algunos estudios basados en *big data* son apenas "pseudoinvestigación"; llega a decir que "the prospect of a *big data* revolution in linguistics is fundamentally illusory" ['la posibilidad de una revolución de *big data* en la lingüística es fundamentalmente ilusoria'] (Hardie 2010: 23). De manera que la razón por la que el tamaño de un corpus puede ser un factor secundario es esta: existe un acuerdo generalizado en que un corpus no es una mera colección de textos, sino que se construye deliberada y cuidadosamente para que pueda servir como ejemplo representativo de una lengua, y para que los resultados extraídos de sus textos puedan servir de ejemplo de cómo funciona la lengua en general, es decir, puedan "generalizarse" y no tomarse como una característica especial del corpus del que se han obtenido los datos. Es decir, un

corpus debe ser "representativo" y "equilibrado", con una composición cuidadosamente escogida. El problema añadido es que estos dos conceptos no han alcanzado una definición aceptada por todos y son todavía muy controvertidos. Según Leech (1991), la suposición de representatividad debe ser más bien aceptada como un "acto de fe"; autores como Tognini-Bonelli (2001) también han expresado su escepticismo, ya que no tenemos de momento manera de asegurarla o de evaluarla de manera objetiva. Discusiones de este tema pueden consultarse en Biber (1993) o Atkins et al. (1992).

En realidad, parece que el tamaño ideal de un corpus depende, en primer lugar, de su propósito, es decir, de cuál sea la pregunta de investigación que se desea responder. No es lo mismo estudiar aspectos muy generales del lenguaje que un fenómeno muy poco frecuente, que, posiblemente, no aparezca o no aparezca (o, al menos, no con la frecuencia deseada) en un corpus de tamaño medio.

En cualquier caso, y una vez presentadas las objeciones, hay que decir que también existen muchos argumentos a favor de corpus lo más grandes posibles. Un mayor tamaño facilita de manera obvia el estudio de fenómenos lingüísticos más raros, menos frecuentes, que no ocurren en corpus pequeños con la frecuencia necesaria. Igualmente, contar con un número suficiente de datos permite el estudio de un fenómeno teniendo en cuenta distintas variables; al usar distintas variables, el número de casos para el estudio debe aumentar. Por poner un ejemplo muy sencillo: si pensamos que hombres y mujeres exhiben una diferencia en la expresión de un fenómeno lingüístico, debemos recopilar un número de casos que será dividido en dos grupos según la variable "sexo", lo que requiere el doble de casos que si no hiciéramos esa distinción. Si aumentamos el número de variables (que es potencialmente muy amplio: registro, contexto lingüístico, procedencia geográfica, frecuencia léxica, número de vecinos semánticos, valencia, y un larguísimo etcétera), se empieza a comprender por qué puede ser necesario contar con un número de ejemplos lo más grande posible.

3.2. La variedad de los datos

La otra característica de los *big data* que hemos destacado como especialmente relevante para los estudios lingüísticos es su gran variedad. Ya hemos mencionado que existen muchos tipos de datos en el *big data* y también que existe un número potencialmente muy grande (podríamos incluso decir que un número indefinidamente grande) de variables que pueden ser relevantes en el estudio lingüístico. Por ejemplo, con mucha frecuencia, los datos lingüísticos (esto es, las palabras) se anotan con información morfológica (según el tipo de morfema, por ejemplo), sintáctica (si una palabra es nombre, verbo o adjetivo) o gramatical (anotándose la función de una palabra, si es sujeto, objeto o modificador verbal). Potencialmente, hay muchas más capas de anotaciones posibles: anotaciones prosódicas, gestuales, discursivas, pragmáticas, geográficas (usando el GPS, por ejemplo), entre otras. Pensemos por un momento en la investigación sociolingüística: incluir información del hablante (su género, su edad) o del momento en que se dijo algo o el lugar geográfico, por ejemplo, son todos aspectos absolutamente importantes para una caracterización sociolingüística de las expresiones. Muy relacionado con este tema encontramos las distintas "capas" de información que se distinguen en los corpus lingüísticos. Los datos asociados con un corpus lingüístico se pueden dividir en tres tipos.

En primer lugar, tenemos los datos primarios, que son el material lingüístico en sí, es decir, las expresiones lingüísticas escritas o dichas por los hablantes. En segundo lugar,

están los metadatos. En esta segunda capa se incluye información referida a los detalles de la captura o toma de los datos: dónde se llevó a cabo, en qué momento, de dónde se extrajeron los datos –periódico, entrevista de televisión, novela…–. En el último lugar se encuentra la capa de anotación. Esto son datos añadidos por el analista o por el sistema y corresponden a información morfosintáctica, semántica, gestual, prosódica, o de cualquier otro tipo que podamos imaginar.
Esta última capa es la que ofrece mayores oportunidades de expansión y enriquecimiento, ya que es absolutamente abierta. Unos datos primarios (por ejemplo, un texto) se pueden anotar con sucesivas capas de información, que permitirán, por un lado, búsquedas en el corpus mucho más refinadas y enfocadas, y por otro, la realización de estudios cuantitativos mucho más sofisticados, que exploren correlaciones completas entre los distintos tipos de información presente en las capas de anotación.

Sin embargo, es precisamente esta promesa de enriquecimiento de la información la que trae aparejada nuevos problemas, que son críticos para llegar a un aprovechamiento real de los "grandes datos" en los estudios lingüísticos. Para empezar, la adición de determinadas etiquetas o anotaciones puede (y, de hecho, suele) estar asociada a unos supuestos teóricos concretos; es difícil encontrar etiquetas que estén "libres de teoría", lo que limita la utilidad de ese etiquetado a los investigadores que siguen determinada teoría. Pero, al margen de este problema, el verdadero cuello de botella tiene que ver con la manera en que se añaden esas etiquetas que describen los aspectos suplementarios. La adición de etiquetas puede ser un proceso manual, automático o semiautomático. El proceso manual es el que permite un ajuste más preciso y razonado de la etiqueta que se añade; determinados tipos de anotación pueden necesitar de manera más acusada el juicio humano, como podría ser el caso de anotaciones de tipo semántico o pragmático. Sin embargo, esto conlleva un claro coste temporal: un humano puede anotar únicamente una cantidad de información determinada y, cuando estamos hablando de corpus de miles de millones de palabras que aumentan a una gran velocidad, lo más probable es que únicamente un porcentaje muy limitado del tamaño real pueda ser anotado de esta manera. Es decir, podemos tener Volumen o Variedad, pero no ambas cosas al mismo tiempo. A este problema se añade el que la anotación humana trae también aparejada problemas relativos a la "subjetividad" del analista. Esto quiere decir que se requiere una validación adicional: los mismos datos deben ser anotados por distintos anotadores y se debe efectuar una comparación de las anotaciones para ver si se alcanza el necesario nivel de "acuerdo entre jueces", para lo que existen herramientas estadísticas que miden el nivel de acuerdo deseado (por ejemplo, la K de Kendall).

Queda entonces claro que el verdadero desafío consiste en encontrar herramientas de anotación que sean o bien completamente automáticas, o al menos, semiautomáticas, y que permitan, por tanto, a los anotadores humanos optimizar el proceso de anotación y aumentar de manera sensible el número de casos anotados. De otra manera, aunque seamos capaces de satisfacer la primera V, de "Volumen", por muy grande que sea el volumen de datos primarios, no tendremos la V de "Variedad" más que en un grupo reducido de casos. En las siguientes secciones vamos a examinar este problema con más detalle centrando nuestra atención en un tipo especial de corpus que intenta incluir información lingüística de distintos tipos: los corpus multimodales.

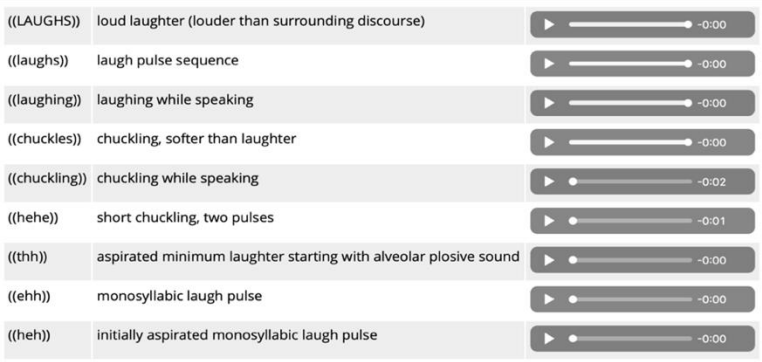
4. Un ejemplo de variedad en los datos: los corpus multimodales

La cantidad y variedad de información que está presente en la comunicación lingüística va más allá de lo puramente verbal, es decir, del simple reconocimiento de una lista de palabras (véase la Figura 6). Un alto porcentaje de la información que utilizamos de manera usual en la comunicación se recibe de manera visual, no auditiva; es el caso de los gestos, las posturas corporales, las expresiones faciales, la mirada... Por supuesto, también hay información adicional más allá del reconocimiento léxico en la señal sonora; existen toda una serie de matices de entonación que no siempre se tienen en cuenta en los análisis lingüísticos. Pensemos en un intercambio tan sencillo como alguien que dice la palabra *hola*. Dependiendo de la entonación, de la expresión facial o de la postura corporal, podemos entenderla de muchísimas maneras distintas: por ejemplo, podemos imaginar un *hola* que signifique 'qué sorpresa verte aquí' (asociado con una entonación y una expresión facial determinada), o que signifique '¿hay alguien aquí?' (dicho con un volumen y una entonación concreta), o 'cómo me sorprende que digas esto', o simplemente un saludo neutro. De todos estos parámetros, el estudio de los gestos está especialmente avanzado y se ha reconocido ya su ubicuidad: al parecer, no existe ningún lenguaje humano en el que no se gesticule al hablar. Estos gestos a veces refuerzan la información presente en el mensaje verbal (actúan de manera redundante); a veces ofrecen información nueva y lo complementan; y a veces ofrecen pistas de cómo interpretar de manera correcta un mensaje con varias interpretaciones posibles. Existe toda una nueva corriente en los estudios de lenguaje dirigida a extraer información multimodal de los eventos comunicativos e integrarla en modelos más complejos del lenguaje. Este movimiento de comunicación multimodal está en plena evolución y sus fronteras están todavía por definir. Áreas como la gestualidad (y la entonación) están más desarrolladas, pero existen muchas otras áreas, más minoritarias, que están desarrollándose de manera paralela. Entre otros ejemplos, Keevalik y Ogden (2020) ofrecen una revisión de sonidos usados en la comunicación (extraídos del inglés y el estonio) y que no forman parte del repertorio fonético clásico de una lengua.

Figura 6. Ejemplos de multimodalidad en la comunicación lingüística

Un ejemplo concreto del nuevo tratamiento con datos muy variados y no incluidos en los análisis más tradicionales es el proyecto BabyCASE, que podemos usar como ejemplo para comprender tanto la utilidad y el valor explicativo de este tipo de información, como las dificultades inherentes a su uso y tratamiento. Este proyecto contiene 764 ejemplos transcritos de comportamiento no verbal, con rasgos como movimientos de cabeza (asentimientos o negaciones), sonrisas, dirección de la mirada, inclinación del cuerpo hacia delante o hacia atrás, imitaciones, risas y diversos movimientos de las manos (Brunner y Diemer 2018). Entre otros aspectos, este proyecto ha estudiado cómo se relacionan aspectos no verbales con eventos de *code-switching*: cambios de idioma de los hablantes, que pasan a utilizar en un momento dado y puntual un código distinto en una

interacción. En la Figura 7 se puede ver así las distintas clasificaciones de tipos de risas, cómo se relacionan con eventos de *code-switching*, y cuál es la posición de esa risa en el contexto en el que se cambia de idioma.



((LAUGHS))	loud laughter (louder than surrounding discourse)	-0:00
((laughs))	laugh pulse sequence	-0:00
((laughing))	laughing while speaking	-0:00
((chuckles))	chuckling, softer than laughter	-0:00
((chuckling))	chuckling while speaking	-0:02
((hehe))	short chuckling, two pulses	-0:01
((thh))	aspirated minimum laughter starting with alveolar plosive sound	-0:00
((ehh))	monosyllabic laugh pulse	-0:00
((heh))	initially aspirated monosyllabic laugh pulse	-0:00

Code-switching and contextual laughter in BabyCASE	**n**	**% of total**
Contextual laughter (5 preceding or following intonation units)	60	56.07%
No contextual laughter	47	43.12%
Code-switches total	107	100.00%

Position of contextual laughter in BabyCASE	**n**	**% of total**
Contextual laughter (5 preceding intonation units)	20	33.33%
Contextual laughter (5 following intonation units)	24	40.00%
Contextual laughter (5 preceding AND 5 following intonation units)	16	26.67%
Code-switches with contextual laughter	60	100.00%

Figura 7. Ejemplos de análisis multimodal: función de distintos tipos de risas en el code-switching (Brunner y Diemer 2018)

Este estudio, por lo tanto, cumpliría una de las dos características principales del *big data* (la variedad), pero ofrece un número extremadamente limitado de ejemplos analizados. Podemos comprender, al examinar este estudio, la dificultad de extender este tipo de análisis, absolutamente dependiente de la anotación humana, a la enorme cantidad de datos disponibles. Este es, de hecho, el principal problema de los corpus multimodales. Hace ahora unos 10 años, en Knight (2011) se comentaba que apenas existían corpus multimodales cuyo tamaño fuera más allá de unos pocos miles de palabras. Y se mencionaba como ejemplo el AMI corpus con "an impressive 100 hours of video". Estos datos sirven como referencia para la siguiente sección, en la que presentamos un corpus multimodal que puede ser tenido en cuenta como integrante real del movimiento *big data*, debido tanto a la variedad de información ofrecida como a su gran tamaño.

5. The NewsScape corpus y el Laboratorio Red Hen

Presentamos en este apartado la investigación llevada a cabo por el Laboratorio Red Hen (https://redhenlab.org), un consorcio internacional que agrupa a distintas universidades y grupos de investigación, y que está dedicado al estudio del lenguaje multimodal. El principal recurso del Red Hen Lab es una base de datos multimodal, conocida como la *NewsScape Library of International Television News* (a partir de ahora *NewsScape*), alojada en la biblioteca de la University of California, Los Ángeles (UCLA). Esta base

de datos consiste en la recopilación de un gran número de horas de programas de televisión, más de 500.000. Estos programas se graban junto con los subtítulos de lo que se dice en cada momento. De manera crucial, estos subtítulos se alinean de manera adicional (mediante un procedimiento conocido como *forced-alignment*): un súper ordenador de altas prestaciones es el responsable de incluir una marca temporal (cada segundo) en los subtítulos. Esto permite que el material textual que describe lo que se está diciendo pueda ser tratado como cualquier otro corpus textual, con las mismas herramientas de búsqueda. En este caso, sin embargo, los resultados de una búsqueda textual incorporan esta marca temporal, de manera que se puede buscar una palabra, una frase o una expresión gramatical compleja y los resultados incluyen un enlace al momento exacto en que se dijo dicha expresión en los archivos de vídeo. Las implicaciones de esto para el estudio multimodal son enormes. Podemos no solo saber cuántas veces, cuándo y quién ha dicho una determinada expresión, sino examinar a los hablantes en el momento de decirla, teniendo por lo tanto a nuestra disposición información multimodal como rasgos prosódicos, gestuales o de cualquier otro tipo. El corpus textual formado por esos subtítulos asociados a las 500.000 horas (aproximadas, puesto que el corpus va creciendo a un ritmo aproximado de 150 horas por día) alcanza un tamaño de unos 4.000 millones de palabras, lo que permite clasificar el estudio basado en *NewsScape* como "de *big data*".

Figura 8. The Red Hen Lab (https:/www.redhenlab.org)

5.1. Formato de los datos en Red Hen: metadatos y anotaciones

Como hemos dicho anteriormente, se ha reservado la etiqueta de "metadatos" para el etiquetado que es añadido de manera más o menos automática y que responde a información sobre los detalles de la captura de los datos en cuestión: el origen de los datos, esto es, de qué fuente se extrajeron (si era un periódico, una novela o una conversación telefónica), cuándo se realizó la toma de datos, etc. En *NewsScape* existen una serie de metadatos adaptados a su idiosincrasia. Como todos los datos se extraen de programas de televisión, los datos primarios vienen anotados de manera automática con información sobre de qué cadena de TV se extraen los datos, el momento temporal (año,

mes, día, hora, segundo). En la Tabla 1 se muestra una selección de los metadatos más básicos en NewsScape (para más información, se puede visitar la dirección www.redhenlab.org).

TOP	contiene la marca de tiempo de inicio y el nombre del archivo
COL	contiene el nombre de la colección
UID	una identificación única para la colección
PID	el episodio del programa (EP) o el ID del programa (SH) (cuando esté disponible)
AQD	el momento de la adquisición
DUR	la duración de la grabación en horas: minutos: segundos, centésimas de segundo
VID	el tamaño de la imagen del vídeo comprimido y del vídeo original
TTL	el título del evento si corresponde, o la serie, si contiene caracteres no ascii
URL	la fuente web si corresponde
TTS	el tipo de transcripción si corresponde
SRC	el lugar de grabación
CMT	un comentario agregado por la persona que programa la grabación
LAN	código de idioma ISO de tres letras (ISO 639-2T)
TTP	la página de teletexto
HED	el encabezado si está disponible, generalmente con información resumida sobre el contenido
OBT	la hora de transmisión original, cuando difiere de la hora de transmisión local "OBT | Estimado" se utiliza en archivos digitalizados cuando se desconoce la hora de transmisión precisa
LBT	la hora de transmisión local, con zona horaria
END	La marca de tiempo de finalización se deriva de la hora de inicio más la duración del vídeo. Le sigue una repetición del nombre del archivo.

Tabla 1. Metadatos en NewsScape

En teoría, hemos mencionado una distinción existente entre "metadatos" (datos relativos a la recogida) y "anotaciones" (datos añadidos por anotadores humanos). Sin embargo, estas dos categorías con frecuencia se solapan, de manera que se distinguen únicamente dos capas: la primaria, de los datos en sí (texto, audio) y los metadatos. Podemos dividir los metadatos añadidos o secundarios entre los automáticos y los añadidos por anotadores humanos. En Red Hen se intenta utilizar herramientas de anotación automática a disposición de la comunidad académica; por ejemplo, programas de *software* existentes en el mercado que anotan de manera automática las "partes de la oración" de todas las palabras de un corpus. En otras ocasiones, se añaden anotaciones provenientes de equipos de anotadores humanos (véase Tabla 2).

FRM_01	marcos lingüísticos (*frames*) de la FrameNet 1.5 por medio del software Semafor 3.0-alpha4
GES_02	gestos de tiempos etiquetados manualmente por el grupo de investigación del Daedalus Lab
GES_03	gestos etiquetados manualmente con ELAN
NER_03	reconocimiento de nombres de entidades (usando el Stanford NER tagger 3)
POS_01	categorías gramaticales inglesas con dependencias, usando MBSP 1.4
POS_02	categorías gramaticales inglesas usando el Stanford POS tagger 3.4
POS_03	categorías gramaticales alemanas usando Pattern.de
POS_04	categorías gramaticales francesas usando Pattern.f
POS_05	categorías gramaticales españolas usando pattern.es
SEG	fronteras entre historias por Weixin Li, UCLA
SEG_00	fronteras de anuncios, usando información de CCExtractor 0.74
SEG_01	detección de anuncios por Weixin Li
SEG_02	Fronteras entre historias por Rongda Zhu, UIUC
SMT_01	Análisis de sentimientos usando Pattern 2.6
SMT_02	Análisis de sentimientos usando SentiWordNet 3.0
DEU_01	Traducción automática de alemán a inglés

Tabla 2. Algunas etiquetas de metadatos añadidas en NewsScape

Uno de los principales objetivos de Red Hen es desarrollar herramientas que permitan automatizar el proceso de etiquetado manual, que, como hemos dicho, es el verdadero cuello de botella de los corpus multimodales. En la actualidad, existen diversos procedimientos de etiquetado (que van del etiquetado manual, realizado por grandes equipos de investigadores, o de etiquetados semiautomáticos, en los que se utiliza un *software* especial para acelerar o facilitar el proceso de anotación), hasta distintas propuestas para un etiquetado automático. A continuación explicamos algunos de estos procesos:

(i) Etiquetado manual

El etiquetado manual se lleva a cabo por parte de equipos dedicados a anotar un determinado comportamiento multimodal. Por ejemplo, el equipo del Daedalus Lab, de la Universidad de Murcia, ha ido recopilando información y tiene una base de datos de unos 10.000 clips anotados gestualmente. Esta base de datos está dedicada a estudiar los gestos espaciales que realizamos al articular expresiones temporales; por ejemplo, el gesto lateral realizado mientras decimos "from beginning to end" (Pagán et al. 2020; Valenzuela et al. 2020). Los clips en los que aparecen este tipo de expresiones han sido clasificados según la visibilidad de los gestos manuales, según el eje utilizado en el patrón gestual, su direccionalidad, la mano utilizada y una variedad de aspectos adicionales (véase también Alcaraz y Valenzuela 2021 para una relación entre la distancia temporal presente en la expresión lingüística –*distant future* vs *near future*– y la distancia física del gesto que acompaña a estas expresiones). La base de datos de la que hablamos ha sido

recopilada de manera manual en un periodo de unos 5-7 años. Es por ello por lo que es necesario intentar acelerar el proceso y aumentar el número de casos anotados.

(ii) Etiquetado manual ayudado por software: el Red Hen Rapid Annotator

Gracias a la colaboración con el lingüista computacional Peter Uhrig, así como por la participación del Red Hen Lab en el proyecto Google Summer of Code, se ha diseñado un *software* que optimiza el tiempo de anotación humana. Al anotador se le presentan los clips de manera automática con una serie de menús con opciones y sus elecciones pasan a grabarse de manera automática en la base de datos, acelerando de esta manera el tiempo de anotación en un factor de x10. La Figura 9 ofrece un ejemplo de este programa para anotar características de una imagen.

Figura 9. Red Hen Rapid Annotator

(iii) Aprendizaje por máquina: visión computacional

Se han realizado también experimentos en los que los clips de vídeo han sido procesados por medio de un algoritmo de visión computacional, que de nuevo optimiza de manera clara el tiempo dedicado a la anotación manual. Por ejemplo, los anotadores humanos deben descartar aquellos clips de vídeo en los que se pronuncia una expresión dada, pero no aparece ningún humano en pantalla (lo que se conoce como "voz en *off*": una persona habla describiendo atascos de tráfico, por ejemplo, pero las imágenes no muestran a esta persona hablando, sino imágenes de tráfico con carreteras con largas colas). El programa desarrollado por Sergiy Turchyn (Turchyn et al. 2018) era capaz de detectar en un clip la presencia o ausencia en la pantalla de un hablante humano, descartando de manera automática los casos de "voz en *off*" y agilizando de manera muy clara el tiempo de anotación del equipo humano. Un estudio piloto en la Universidad de Navarra comprobó que el uso de este software reducía a la mitad (de 1:07:55 a 00:33:47 de media) el tiempo necesario para llevar a cabo la anotación manual.

(iv) Open Pose

En último lugar, en estos momentos el equipo del Red Hen Lab está explorando la utilización del *software* OpenPose (Cao et al. 2021). Este programa de visión computacional identifica puntos del cuerpo humano, especialmente las articulaciones y otros puntos movibles del cuerpo, así como los ojos, o las cejas, y les asigna un número, tal y como se ve en la Figura 10.

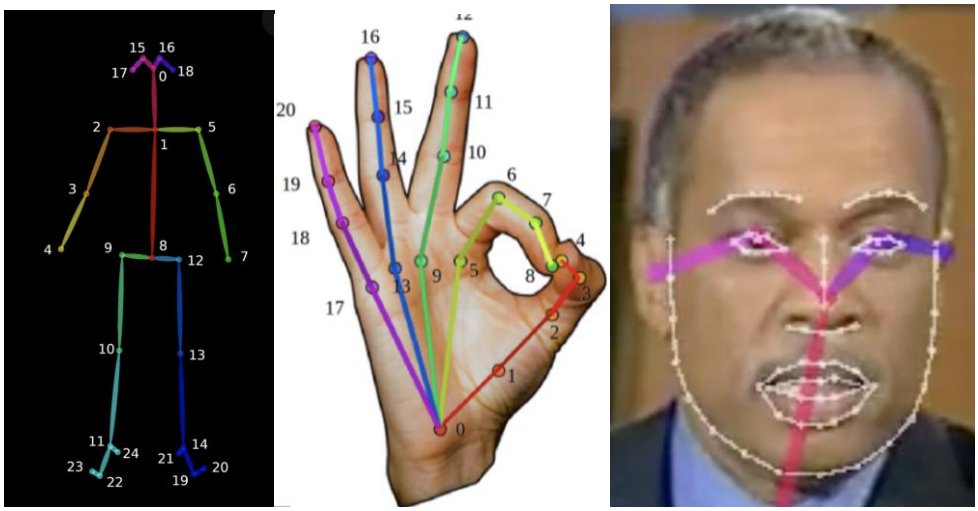


Figura 10. Algunos puntos reconocidos por el programa OpenPose

Los clips de vídeo que resultan de interés para el programa de investigación son de esta manera procesados por el programa OpenPose, que permite la visualización con la información superpuesta o simplemente con los puntos aislados (véase Figura 11).

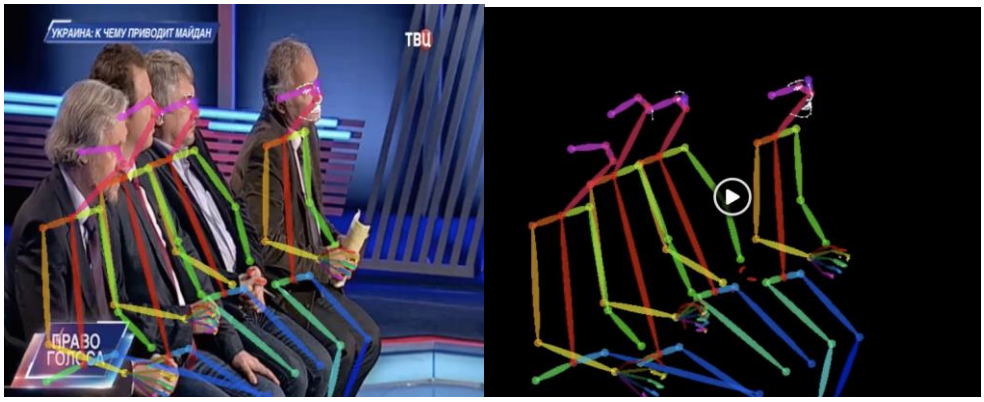


Figura 11. Modos de visualización de la información en OpenPose

A partir de este análisis, el programa genera archivos que van indicando en qué posición espacial (es decir, cuáles son las coordenadas *x* e *y*) se encuentra el punto correspondiente a una determinada parte del cuerpo en un *frame* de un vídeo. Esto quiere decir que es posible realizar un seguimiento del desplazamiento espacial de, digamos, el punto 4, correspondiente a la muñeca izquierda, y comprobar su trayectoria a medida que transcurre el tiempo del clip. Obviamente, esto genera una ingente cantidad de datos (recordemos que un segundo de vídeo contiene unos 30 frames), que tienen que ser sometidos a un cuidadoso análisis por medio de programas que lidien con el "ruido" que

el programa inevitablemente genera (el movimiento de un punto puede verse momentáneamente interrumpido por un obstáculo visual, puede haber varias personas hablando al mismo tiempo, puede haber cambios de cámara que requieran recalcular las posiciones relativas de los puntos, y un largo etcétera). En cualquier caso y una vez superadas estas dificultades técnicas, cruzando estos datos de movimiento detectado automáticamente con los datos de audio de cuándo se empieza a decir algo, estamos en disposición de identificar automáticamente gestos, de manera que podríamos ser capaces de realizar búsquedas como "encuentra a alguien diciendo *en los meses venideros* mientras hace un gesto con la mano derecha de izquierda a derecha", o incluso de buscar gestos de manera automática y estudiar qué tipo de expresiones lingüísticas son las que se asocian con su realización. En estos momentos, el grupo Daedalus Lab está trabajando en este problema, que permitiría dar el salto a un acercamiento verdaderamente de *big data* a los datos multimodales, puesto que la mayor parte del procesamiento de anotación gestual pasaría a automatizarse por medio del algoritmo de OpenPose. La unión de este tipo de información con un análisis prosódico de la señal sonora por medio de programas tipo Praat (Boersma y Weenik 2021) puede ser realmente la puerta que por fin permita llevar el análisis multimodal al siguiente estadio, iniciando la etapa de *big data* en lingüística en su sentido más completo.

6. Conclusiones

A lo largo de este trabajo hemos visto cómo los paralelismos del *big data* en su uso más extendido y generalizado y los *big data* en Lingüística son más bien aproximados. De hecho, sus propósitos son distintos: mientras que el *big data* en general es una empresa de tipo eminentemente práctica, un problema de ingeniería; el objetivo de los *big data* en lingüística es muy distinto: la lingüística es una disciplina académica cuyo propósito es entender los mecanismos subyacentes a la comunicación lingüística, y el uso de *big data* debe servir siempre a ese objetivo de comprensión del fenómeno. De hecho, existen áreas de la lingüística en las que el uso del *big data* en su versión ingenieril ha resultado de gran utilidad: es el caso de los sistemas de reconocimiento de habla, o los sistemas de traducción automática, que han mejorado su funcionamiento de manera notable en los últimos años gracias al entrenamiento de programas con grandes datos (aunque esto no haya llevado aparejado un incremento en la comprensión de cómo se produce el fenómeno).

Claramente, la utilidad de los *big data* para la empresa lingüística es indudable, siempre que se sea consciente de su papel, sus ventajas y sus problemas. En este sentido, como cualquier otra herramienta, su utilidad depende en gran medida del uso que se le quiera dar. Pero, al margen de estas reticencias y precauciones, resulta evidente que el uso de los *big data* puede contribuir de manera decidida a la construcción de modelos mucho más complejos y completos del lenguaje. Como viene siendo la costumbre en los últimos años, los datos provenientes de este tipo de fuente deberán entrar en la "rueda" del resto de herramientas metodológicas que conforman la ciencia cognitiva (métodos conductuales –véase Igoa (2022)– , de seguimiento ocular –*eye-tracking*, véase Álvarez García (2022)– o provenientes de la neuropsicología), contribuyendo de esta manera a afianzar la noción de "evidencia convergente" ya establecida en las ciencias cognitivas: cuando difererentes tipos de evidencia empírica recogida desde diferentes perspectivas apuntan a un mismo tipo de explicación.

7. Agradecimientos

Este trabajo ha sido realizado gracias al apoyo del Ministerio de Ciencia, Innovación y Universidades, Agencia Estatal de Investigación y fondos FEDER/UE funds (grant number PGC2018-1551 097658- B-100).

8. Referencias

Alcaraz Carrión, Daniel; Valenzuela, Javier. 2021. Distant time, distant gesture: speech and gesture correlate to express temporal distance. *Semiotica* 241. DOI: 10.1515/sem-2019-0120

Álvarez García, Esther. 2022. Lo que esconden tus ojos: la metodología eye-tracking aplicada al estudio del lenguaje. *Estudios de Lingüística del Español* 45: 205-239.

Atkins, Sue; Clear, Jeremy; Ostler, Nicholas. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7.1: 1-16.

Biber, Douglas. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8.4: 243-257.

Boersma, Paul; Weenink, David. 2021. *Praat: doing phonetics by computer* [Computer program]. Version 6.1.50.

Brunner, Marie-Louise; Diemer, Stefan. 2018. "You are struggling forwards, and you don't know, and then you … you do code-switching…" – Code-switching in ELF Skype conversations. *Journal of English as a Lingua Franca* 7.1: 59-88.

Cao, Zhe; Hidalgo, Gines; Simon, Tomas; Wei, Shih-En; Sheikh, Yaser. 2021. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE *Transactions on Pattern Analysis and Machine Intelligence* 43 .1: 172-186.

García-Miguel, José M. 2022. Lingüística de corpus: de los datos textuales a la teoría lingüística. *Estudios de Lingüística del Español* 45: 11-42:

Hardie, Andrew. 2010. *Big data* in language studies: from cargo-cult science to phantom revolution. Conferencia plenaria en el 7 Congreso de AELINCO 2015, Universidad de Valladolid (https://docplayer.es/4700819-Aelinco-2015-book-of-abstracts.html).

Keevallik, Leelo; Ogden, Richard. 2020. Sounds on the Margins of Language at the Heart of Interaction. *Research on Language and Social Interaction* 53.1: 1-18. DOI: 10.1080/08351813.2020.1712961

Knight, Dawn. 2010. The future of multimodal corpora. *Revista Brasileira de Linguística* Aplicada 11.2: 391-415.

Krishnamurthy, Ramesh. 2001. Size Matters: creating Dictionaries from the World's Largest Corpus. *8th Annual KOTESOL Conference Proceedings.* Taegu: KOTESOL: 169-180.

Igoa, José Manuel. Las tareas conductuales en la investigación sobre el procesamiento del lenguaje. *Estudios de Lingüística del Español* 45: 133-158.

Leech, Geoffrey. 1991. The state of the art in corpus linguistics. En K. Aijmer y B. Altenberg, eds. *English Corpus Linguistics*, Londres: Longman, pp. 8-29.

Olza, Inés; Valenzuela, Javier; Pagán-Cánovas, Cristobal. 2017. Automatic visual analysis and gesture recognition: Two preliminary pilots. Universidad de Navarra: Instituto Cultura Sociedad.

Pagán Cánovas Cristóbal; Valenzuela Javier; Alcaraz Carrión Daniel; Olza Inés; Ramscar Michael. 2020. Quantifying the speech-gesture relation with massive multimodal datasets: Informativity in time expressions. *PLOS ONE* 15.6: e0233892.

Rumelhart, David E.; McClelland, James L.; PDP Research Group. 1986. *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1*. Cambridge, MA: MIT Press.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam y Philadelphia: Benjamins.

Turchyn Sergiy; Olza Moreno, Inés; Pagán Cánovas, Cristóbal; Steen, Francis F; Turner Mark; Valenzuela, Javier; Ray, Soumya. 2018. Gesture Annotation with a Visual Search Engine for Multimodal Communication Research. En *The Thirtieth AAAI Conference on Innovative Applications of Artificial Intelligence* (IAAI-18) [Internet]. 2018. Disponible en: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewFile/16703/16398

Valenzuela, Javier; Pagán-Cánovas, Cristóbal; Olza, Inés; Alcaraz, Daniel. 2020. Gesturing in the wild: spontaneous gestures co-occurring with temporal demarcative expressions provide evidence for a flexible mental timeline. *Review of Cognitive Linguistics* 18.2: 289-316. DOI: 10.1075/rcl.00061.val.