ISSN: 2386-9658 https://doi.org/10.21071/arf.v37i.18687

# ***MSD-Manuals DE-ES*. Un corpus de comunicación especializada mediada médica en el Parallel Corpus of German and Spanish (*PaGeS*)**[1]

## (MSD-Manuals ***DE-ES. A corpus of specialised medical communication as part of the Parallel Corpus of German and Spanish (PaGeS)***)

MARÍA TERESA SÁNCHEZ NIETO
https://orcid.org/0000-0002-0378-3720
mariateresa.sanchez.nieto@uva.es
*Universidad de Valladolid* https://ror.org/01fvbaw18

Fecha de recepción: 29 de septiembre de 2025
Fecha de aceptación: 16 de enero de 2026

**Resumen:** Este trabajo presenta el corpus *MSD Manuals DE-ES*, un suplemento del corpus *PaGeS* en el proyecto PaCorES, centrado en la comunicación especializada en medicina. El corpus se basa en las versiones profesional y para el público general de los Manuales Merck/MSD, reconocidos por su fiabilidad y accesibilidad. El objetivo principal es documentar el proceso de compilación y alineación del corpus, que, en su versión 1.0 incluye más de 500.000 bisegmentos y 17,5 millones de palabras, distribuidas en dos subcorpus (*MSD Manual DE-ES home* y *MSD Manual DE-ES profesional*) que reflejan registros comunicativos distintos: experto-lego y experto-experto. La metodología empleada combina herramientas del proyecto MTUOC y scripts desarrollados con ayuda de modelos de lenguaje como Copilot y LeChat, lo que permite automatizar tareas complejas como la extracción de texto, la alineación y la gestión de metadatos. Se detallan los retos técnicos, como la limpieza de datos y la segmentación, y se reflexiona sobre el potencial de la inteligencia artificial en la lingüística de corpus. El corpus *MSD-Manuals DE-ES* sirve a la investigación en traducción especializada,

[1] Este estudio forma parte del proyecto *Corpus paralelos online del español: Una herramienta multifuncional para la traducción, el aprendizaje de lenguas y la investigación lingüística (PaCorES)*, financiado por el Ministerio de Ciencia e Innovación (PID2021-125313OB-I00, 2022-2026). La autora agradece a la Direction of Marketing and Strategic Partnerships for the Merck and MSD Manuals la comunicación mantenida para establecer las condiciones de reutilización del contenido procedente de las distintas ediciones y versiones de los Manuales MSD utilizados en los corpus a los que se refiere este trabajo.

lexicografía y enseñanza de lenguas con fines específicos y contribuye a paliar la escasez de recursos paralelos en el ámbito biosanitario para el par de lenguas alemán-español. Su integración en *PaGeS* refuerza la oferta de corpus accesibles en línea para la comunidad investigadora.

**Palabras clave:** Corpus paralelo. Proyecto PaCorES. Compilación del corpus. Alemán. Español. PaGeS.

**Abstract:** This paper introduces the *MSD Manuals DE-ES* corpus, a bilingual German-Spanish supplement to the *PaGeS* corpus within the PaCorES project, focused on specialised medical communication. The corpus draws on both the professional and general-public versions of the Merck/MSD Manuals, renowned for their reliability and accessibility. Its primary aim is to document the compilation and alignment process of the corpus, which comprises over 500,000 bi-segments and 17.5 million words, distributed across two subcorpora (*MSD Manual DE-ES home* and *MSD Manual DE-ES professional*) that reflect distinct communicative registers: expert-to-layperson and expert-to-expert. The methodology combines tools from the MTUOC project with scripts developed using language models such as Copilot and LeChat, enabling the automation of complex tasks including text extraction, alignment, and metadata management. Technical challenges such as data cleaning and segmentation are addressed, and the potential of artificial intelligence in corpus linguistics is explored. The *MSD-Manuals DE-ES* corpus serves as a valuable resource for research in specialised translation, lexicography, and language teaching for specific purposes, while also helping to address the scarcity of parallel resources in the biomedical domain for the German-Spanish language pair. Its integration into *PaGeS* enhances the availability of freely accessible corpora for the research community.

**Keywords:** Parallel corpus. PaCorES project. Corpus compilation. German. Spanish. PaGeS.

## 1. Contextualización y objetivo

El objetivo de este trabajo es documentar adecuadamente para la comunidad investigadora el corpus *MSD Manuals DE-ES*, un corpus bilingüe español/alemán paralelizado que representa el discurso científico y el lenguaje especializado de la medicina en los registros divulgativo y profesional, así como el proceso de compilación, con el fin de asegurar su replicabilidad.

El corpus *MSD-Manuals DE-ES* se ha compilado para formar parte del proyecto "PaCorES. Corpus paralelos del español". PaCorES es un conjunto de corpus paralelos en los que el español es la lengua central y que pueden consultarse libremente en internet (Doval Reixa y Sánchez Nieto 2026: 2-3). Actualmente, a través de la web de PaCorES (https://www.pacores.eu/) se puede acceder a los corpus *PaGeS* (DE-ES), *PaEnS* (EN-ES), *PaChES* (ZH-ES) y *PaFrES* (FR-ES). Los corpus del proyecto tienen una estructura común: constan de un corpus nuclear y un número variable de suplementos.

El corpus nuclear consta de textos originales españoles de ficción y no ficción y sus traducciones publicadas en la lengua extranjera, así como de textos originales de ficción y no ficción escritos en la lengua extranjera y sus traducciones publicadas en español. Una vez alineados, los textos se han sometido a un proceso de revisión manual. Los corpus tienen una variación léxica importante y un amplio potencial para la investigación Doval Reixa y Sánchez Nieto (2026: 7-11) detallan criterios de compilación y diseño de los corpus nucleares, las operaciones de preprocesado, anotación, segmentado y alineación y las características del sistema de búsqueda y visualización de ejemplos.

Por otra parte, junto a los corpus nucleares, el usuario puede seleccionar en la interfaz de los corpus de PaCorES una serie de corpus paralelos de diferente naturaleza y tamaño: los denominados "suplementos". Estos corpus generalmente cubren usos más específicos de la lengua y son de especial interés desde el punto de vista lexicográfico, de la traducción especializada y del estudio de las lenguas con fines específicos. Doval Reixa y Sánchez Nieto (2026: 11) los definen y caracterizan como sigue:

> [a] separate group of data materials with an adequate level of quality and textual variety, in many cases already aligned, even though we may not always be able to verify the original language, or the alignment may not have undergone manual review.

Los suplementos que existen actualmente en cada uno de los corpus de PaCorES, así como la variedad funcional o tipo de discurso (v. Hoffmann 2007) que cubren, están reflejados en la Tabla 1:

| Suplemento | Corpus en los que se incluye | Tipo de discurso |
|---|---|---|
| Europarl v7[2] | *PaGeS*, *PaEnS* | Administrativo |
| Ted-Talks[3] | *PaGeS*, *PaEnS* | Científico-divulgativo |
| Global Voices[4] | *PaEnS* | Periodístico |
| Open Subtitles | *PaGeS*, *PaEnS* | No específico, si bien refleja las restricciones del medio "subtítulo" |
| Parte del United Nations Parallel Corpus[5] | *PaCheS* | Administrativo |
| Alineaciones de ficción en terceras lenguas | *PaGeS*, *PaFreS* | Artístico-literario |
| Wikimatrix[6] | *PaCheS* | Científico-académico (artículos de Wikipedia) |
| Paracrawl[7] | *PaCheS* | No específico (sitios web multilingües) |

Tabla 1. Suplementos de los corpus de PaCorES

En este contexto, el corpus *MSD-Manuals DE-ES* es un suplemento compilado para integrarse en el corpus *PaGeS*, el primero de los corpus de PaCorES, creado en 2015 (Doval 2017), que pasa así a contar con cinco suplementos[8].

En lo que sigue, tras un breve estado de la cuestión (v. apartado 2), se detallará la naturaleza del material incluido en el suplemento MSD-Manuals y se justificará

[2] V. detalles en http://www.statmt.org/europarl. La versión incluida en los corpus de PaCorES es el corpus CoStEP (Graën / Batinic / Volk 2014), una versión limpiada y corregida de Europarl v7, así como los metadatos de CoStEP, disponibles en http://pub.cl.uzh.ch/purl/costep/

[3] V. detalles en https://www.ted.com/talks. La versión de los ted-Talks incluida en los corpus de PaCorES es la disponible en el Web Inventory of Transcribed and Translated Talks https://wit3.fbk.eu/ (Cettolo / Girardi / Federico 2012).

[4] V. detalles en https://globalvoices.org/.

[5] V. detalles en https://shorturl.at/eyDS9.

[6] V. detalles en https://arxiv.org/pdf/1907.05791.pdf.

[7] V. detalles en https://aclanthology.org/2020.acl-main.417.pdf.

[8] PaGeS, con su corpus nuclear y sus suplementos, incluido *MSD-Manuals DE-ES*, está accesible en https://www.corpuspages.eu/corpus/search/search. V. Ilustración 12.

su interés para los corpus de PaCorES (v. apartado 3). A continuación, se dará cuenta de los procesos de compilación y anotación extralingüística del corpus *MSD-Manuals DE-ES* (v. apartado 4). Finalmente, en las conclusiones se detallarán las cifras del corpus y se reflexionará sobre el proceso seguido (v. apartado 5).

## 2. Estado de la cuestión: la falta de recursos paralelos especializados en biomedicina en el par de lenguas ES/DE

En general, se presupone una baja demanda de traducción en el sector biomédico entre las lenguas alemana y española, al ser el inglés la lengua que cataliza este tipo de traducción especializada hoy en día. Los dos últimos estudios de mercado de la traducción en España no revelan datos específicos sobre la demanda de traducciones en el sector biomédico en el par de lenguas español y alemán. Sin embargo, en esos estudios sí hay datos parciales para la traducción alemán/español, por un lado, y para la traducción biomédica, por otro, que sintetizamos a continuación.

En primer lugar, Rico Pérez y García Aragón (2016: 32) recogían que el 43 % de las empresas encuestadas ofrecía servicios en la combinación alemán/español. En segundo lugar, estos autores (*ibid*, 37) reflejan la demanda de traducción biomédica que existía en 2014-15 (si bien sin especificar por pares de lenguas): entre un 1 % y un 20 % de las empresas tenían un 15 % de facturación en traducciones biomédicas; porcentaje que descendía al 15 % para el 21-40 % de las empresas y en torno al 7-8 % para el 41-80 % de las empresas. En tercer lugar, en el estudio de ANETI y ASPROSET sobre el mercado de los servicios lingüísticos en España (ANETI y ASPROSET 2024), se menciona un género de traducción biomédica como ejemplo del grado 3 de traducción: "traducción personal con revisión personal". Este grado es el que demanda una calidad más alta del servicio:

> **Traducción personal con revisión personal**. Esta es una modalidad donde se busca la máxima calidad de la traducción y, por tanto, se requiere de un traductor especializado en la materia, más un revisor posterior de conocimientos superiores Esta modalidad se aplica, por ejemplo, en traducciones de contratos jurídicos o en la traducción de un prospecto de un medicamento (Rico Pérez y García Aragón 2016: 136)

Sin embargo, Beyerlein-Büchlein (2011: 255) insiste en la demanda de traducción que genera la "creciente economización de la asistencia sanitaria, la

técnica médica y la industria farmacéutica", que lanza al mercado "productos innovadores que aúnan los resultados técnicos y biológicos de la investigación". Y refiriéndose precisamente al ámbito químico-farmacéutico, Mayor Serrano *et al.* (2011: 159) recalcan la importancia del alemán y el volumen de traducción que genera, dado que, desde principios de la década de 2010, "Alemania se ha consolidado como uno de los principales centros de investigación, desarrollo y producción de medicamentos a nivel mundial, y gran parte de las principales empresas europeas del ramo están ubicadas en los países de habla alemana".

Finalmente, siguen apareciendo publicaciones con materiales para la enseñanza de este tipo de traducción en el par alemán/español, como García Jiménez y Varela Salinas (2021) o Ramírez Almansa (2019), lo que indica el potencial no solo didáctico, sino también profesional de los textos de este ámbito en el par de lenguas alemán-español.

Así pues, podemos concluir que la traducción biosanitaria en el par alemán-español se practica y que se sigue investigando en torno a su enseñanza. Cruzando estas informaciones con las obtenidas de los estudios de mercado, podemos aventurar que una pequeña parte de los servicios en el par de lenguas DE-ES ofrecidos en España lo sean de textos del ámbito de la biomedicina y que esos servicios se entienden como de alta calidad y especialización.

Para hacer frente a esta demanda, así como para su enseñanza, hacen falta recursos. El recurso más destacado aparecido en los últimos años es el *Wörterbuch Medizin*, (Navarro 2025b). En la presentación, el autor detalla unos antecedentes que apuntan a los escasos recursos lexicográficos para el par de lenguas alemán/español en el ámbito de la medicina:

> La producción editorial para otras parejas lingüísticas que no sean inglés-español ofrece en nuestros países, desde la II Guerra Mundial en adelante, un aspecto desolador que me atrevo a calificar, sin temor a exagerar, de auténtico páramo lexicográfico. Este panorama desolador resulta muy llamativo para la pareja alemán-español, por contraste con la importancia del papel central desempeñado por los países germánicos en la medicina y otras ciencias biosanitarias del siglo XX (Navarro 2025b: s.p.).

El *Wörterbuch Medizin* ha venido a paliar estas necesidades documentales básicas. El autor y su equipo de colaboradores recogen —tal y como se deduce del apartado "Bibliografía" de su presentación— el legado de diccionarios previos y las aportaciones de glosarios especializados publicados por expertos en diferentes medios dedicados a la traducción especializada o a la comunicación en

biomedicina (*Lebende Sprachen*, *Panace@*, entre otros). Sin embargo, como el propio autor reconoce (Navarro 2025b: s.p.), *Wörterbuch Medizin* es "un gran diccionario médico alemán-español de corte clásico: "con simples equivalencias sin apenas acompañamiento crítico, explicativo ni enciclopédico", que se entiende como paso necesario y previo a un posible diccionario de dudas y dificultades específicas de este par de lenguas, al estilo del *Libro rojo* (Navarro 2025a) para el inglés y el español.

Si dejamos de lado los glosarios especializados, los diccionarios en papel y los diccionarios en línea —con *Wörterbuch Medizin* a la cabeza—, así como los recursos recogidos en la bibliografía de este último, el par alemán/español también adolece de otros recursos especializados en biomedicina dirigidos al traductor y consultables en línea, como p. ej. corpus, ya sean monolingües o bilingües. Una búsqueda documental en la base de datos especializada BITRA[9] con las palabras clave CORPUS, MEDICINA, ALEMÁN no arroja resultados; tampoco lo hace la búsqueda con las palabras clave CORPUS, MEDICINA y las cadenas "German" o "Deutsch" en el campo "título"; con las palabras clave CORPUS, MEDICINA y la cadena "alemán" en el título obtenemos un artículo: Balbuena Torezano (2019), si bien este no remite a un corpus disponible en línea, sino a una creado *ad hoc* para un estudio específico.

Una búsqueda en el repositorio de corpus paralelos descargables OPUS (https://opus.nlpl.eu/) para el par de lenguas alemán-español[10] permite localizar el corpus paralelo multilingüe EMEA, del que se puede extraer el par alemán-español. Este corpus consta de una alineación de documentos PDF provenientes de la Agencia Europea del Medicamento (European Medicines Agency), y contiene respuestas a unos formularios que respondieron profesionales de la medicina. La extracción del par de lenguas alemán-español contiene 681 147 unidades de traducción (bisegmentos) (ELRC 2020). Este corpus es consultable hoy por hoy a través de SketchEngine (EMEA v.03) como subcorpus del Open Parallel Corpus (OPUS) y también puede descargarse del repositorio Clarin[11]. Aparte de este recurso, no existe ningún otro consultable directamente en línea para el par de lenguas alemán-español.

[9] V. https://aplicacionesua.cpd.ua.es/tra_int/usu/buscar.asp.
[10] Los recursos existentes pueden consultarse en esta tabla: https://opus.nlpl.eu/results/de&es/corpus-result-table.
[11] V. https://inventory.clarin.gr/corpus/747.

## 3. Los Manuales MSD como base textual del Corpus *MSD-Manuals DE-ES*

Los *Merck Manuals* (o *Manuales MSD*, como se conocen fuera de Norteamérica) son obras de referencia médica que se publicaron por primera vez en 1899 como una guía de bolsillo para médicos y farmacéuticos. Con el tiempo, se han transformado en un recurso en línea con versiones tanto para facultativos y otros profesionales de la salud como para el público no especializado (Bullers 2017: 369; Tomes 2021: 1). La versión dedicada a este último o "versión para el público general" (*The Merck Manual of Medical Information: Home edition*) apareció en 1997[12]. En 2014 Merck & Co. deciden publicar estos manuales exclusivamente en línea, traducidos (inicialmente al español, y, desde finales de 2016, al alemán, chino, coreano, francés, italiano, japonés, portugués y ruso; actualmente también al hindi, suajili e indonesio) con acceso gratuito. Según Bullers (2017: 370), este giro editorial radica en la iniciativa "Conocimiento médico global", con la que Merck & Co. quieren conseguir que "3000 millones de profesionales y pacientes en todos los continentes tengan acceso a la mejor información médica actual" (Merck & Co. 2025: s.p.)

Entre los objetivos de la publicación, los editores destacan "posibilitar decisiones más informadas, mejorar las relaciones entre los pacientes y los profesionales, y mejorar los resultados de la atención médica en todo el mundo" (Merck & Co. 2025). Actualmente, el sitio web de los manuales reúne dos versiones, marcadas por un código de color: rojo para la versión para el público general y azul para la versión para profesionales, entre las que el usuario puede cambiar mediante los botones respectivos situados en la esquina superior izquierda de la página. El contenido principal de los manuales son los artículos temáticos que constituyeron la base de las primeras ediciones (Bullers 2017: 370). Los artículos constan de una introducción y apartados dedicados a las causas (*etiología*, en la versión profesional), valoración (*evaluación* en la versión profesional), tratamiento y conceptos clave. La versión profesional contiene, además, un apartado dedicado a la fisiopatología. Los artículos incluyen también guías de referencia para valores normales de laboratorio, tablas, tests y estudios de caso (para interaccionar con el contenido), así como contenido multimedia: gráficos, animaciones y vídeos; estos últimos muestran técnicas procedimentales y de inspección (370).

[12] La edición original en inglés denomina a este público *consumers*. La modulación subyacente a esta equivalencia terminológica presente en la versión inglesa frente a la española o la alemana (ingl. *consumers* vs. esp. *público general*, al. *Patienten*) no pasa inadvertida. Sin embargo, la discusión de sus implicaciones histórico-culturales excede los límites de este artículo. Véase, a este respecto, Tomes (2016).

Los artículos de la actual versión profesional son, como en sus inicios, una fuente "escrita por médicos para médicos" (Tomes 2021: 2). Por el contrario, los de la versión para el público general "are written at about an eighth-grade level", y, sin embargo, "include more detail than might be found on other consulter health resources" (Bullers 2017: 370). Tomes (2021: 2-3) considera que el caso de la evolución del *Merck Manual* "from doctor's handbook to patient's bible" es un caso paradigmático para estudiar "la revolución en la información para la salud pública acaecida a finales del siglo XX y sus implicaciones para la relación médico-paciente"[13] (mi traducción), así como para entender cómo el acceso a fuentes especializadas "ha servido para marcar la línea divisoria experto/lego en la difusión del conocimiento científico" (mi traducción).

Bullers (2017: 371) valora actualmente los manuales como "a useful ready reference tool" que puede recomendarse con confianza, basándose en tres hechos principalmente: a) los artículos están revisados por un comité independiente de profesionales médicos; b) los autores aparecen declarados junto con sus datos profesionales identificativos, y c) las fechas de publicación y última revisión o edición también están explícitas (*id*).

La fiabilidad de los manuales y la variación diafásica que representan —uso especializado de la lengua en el ámbito médico en situaciones de comunicación simétricas (experto-experto) y asimétricas (experto-lego) y ello tanto en español como en alemán— los convierten en una fuente ideal para elaborar un suplemento para *PaGeS*.

En los siguientes epígrafes explicaremos el proceso de selección del material y de compilación del suplemento *MSD-Manuals DE-ES* a partir de las versiones alemana y española de los Manuales MSD disponibles en línea (*Manual MSD. Versión para profesionales; Manual MSD. Versión para público general; MSD Manual. Ausgabe für Patienten; MSD Manual. Profi-Ausgabe)*.

## 4. Proceso de compilación del corpus *MSD-Manuals DE-ES*: fases, pasos y resultados

La compilación se divide en dos grandes fases.

La primera fase comprende los pasos que permiten la obtención de texto paralelo de la web y su alineación (v. pasos 1 a 4 de la Ilustración 1) y reproduce en parte la metodología esbozada por Oliver González (2023a, 2024e: 35-37) para el entrenamiento de motores de traducción automática. Estos pasos están descritos en los epígrafes 4.1 a 4.4. Nos hemos servido de varios *scripts* del proyecto MTUOC (Oliver González 2023b, 2023d, 2023c, 2024d), que

[13] Traducción propia.

comentaremos y vincularemos, así como otros creados *ad hoc* para este proyecto (Oliver González 2024a, 2024c).

La segunda fase, que comprende las operaciones de limpieza, numeración y creación de las bases relacionales texto-metadatos (v. pasos 5 a 9 de la Ilustración 1), despliega la metodología específica de este proyecto. Los pasos están descritos en los epígrafes 4.5 a 4.9 y conducen a la obtención de dos archivos con la estructura que requieren los corpus del proyecto PaCorES:

1. un archivo .tsv (con datos separados por tabuladores) con tres columnas: ID (una cadena alfanumérica con la estructura nnnnn-nnnnnnn que identifica cada segmento); L1 (texto en lengua extranjera); ES (texto en español);
2. un archivo .tsv con la anotación documental, esto es, con los metadatos de los textos que conforman el corpus.

Para automatizar ciertas tareas lingüísticas necesarias en esta segunda fase, hemos utilizado los bots conversacionales Copilot (Microsoft) y LeChat (Mistral.ai), aplicaciones que, como otras de su clase, descansan sobre modelos de lenguaje grandes (LLM, por sus siglas en inglés), para crear y adaptar pequeños *scripts*. Estos programas están escritos en el lenguaje de Python, ampliamente utilizado por su legibilidad y versatilidad. Hemos probado y ajustado estos *scripts* hasta que realizaron satisfactoriamente las tareas requeridas. Los programas se ejecutan en Linux, un sistema operativo de código abierto, y para ello hemos utilizado específicamente la distribución Ubuntu 22.04.5 LTS[14].

[14] Disponible en línea en ubuntu.com.

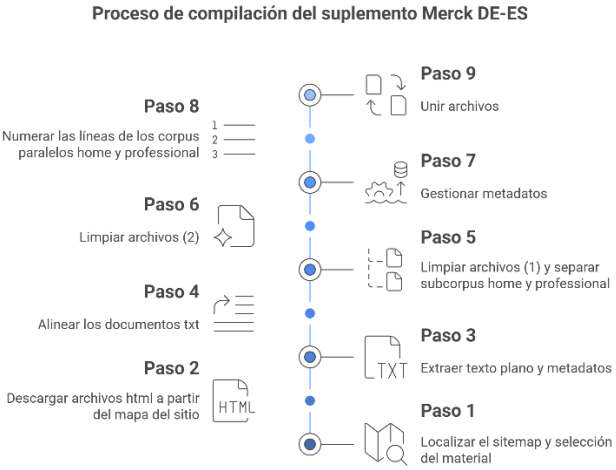



Ilustración 1. Pasos para la compilación del corpus *MSD-Manuals DE-ES*. Elaboración propia con napkin.ai

*4.1 Localización del* sitemap *y selección del material*

El *sitemap* es un archivo que contiene muchos de los enlaces de un determinado sitio web. El algoritmo incluido en el *script MTUOC-sitemap.py* (Oliver González 2023d) permite obtener el mapa de un sitio web (*sitemap*) introduciendo en el parámetro -u la *URL* del sitio:

python3 MTUOC-sitemap.py -u https://www.merckmanuals.com

Tras ejecutarlo, el programa devuelve un archivo de texto, en nuestro caso, sitemap-meckmanuals_com.txt, con todas las direcciones que componen el sitio.

Este primer mapa descargado nos permitió observar a) que, de entre las *URL* captadas por el programa, un buen número de ellas contenían la cadena "/multimedia/", que llevaba a páginas con imágenes o vídeos y poco contenido textual; b) que se habían descargado vínculos a textos en el resto de los idiomas que contempla la página (vietnamita, inglés, italiano...); c) que en el caso del español, la página www.merckmanuals.com contenía cinco *locales* o variedades /es-ar/, /es-do/, /es-cr/, /es-cl/ y /es-ec/ con el mismo número de enlaces y los mismos contenidos; d) que las *URL* reflejaban efectivamente la adscripción de los contenidos respectivos a una las dos versiones existentes en los subdirectorios /heim/ u /hogar/, /profi/ o /profesional/, como reflejan las siguientes capturas del archivo inicial de *sitemap*, como ejemplifican la Ilustración 2 y la Ilustración 3):

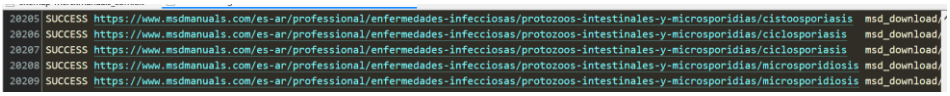



Ilustración 2. Vínculos a artículos temáticos
de la versión española de la edición profesional del manual MSD

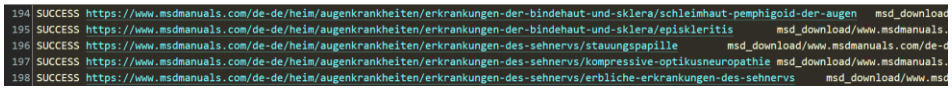



Ilustración 3. Vínculos a artículos temáticos de la versión alemana
de la edición para el público general del manual MSD

Sin embargo, posteriormente supimos que Merck & Co. publica los propios *sitemaps* de sus versiones, separados por lenguas. En el caso de DE y ES: msdmanuals.com/de/sitemap.ashx y msdmanuals.com/es/sitemap.ashx.

Pueden descargarse ejecutando el comando *wget*, por ejemplo: wget https://www.msdmanuals.com/es/sitemap.ashx

*4.2 Descargar archivos html a partir del* sitemap

Para este paso, se utiliza el *script MTUOC-download-from-sitemap.py* (Oliver González 2023b) Sin embargo, en nuestro caso utilizamos una versión modificada de dicho script, adaptada a la estructura de la página de los manuales: *downloadMSDhtmls.py* (Oliver González 2024a). Este último *script* a) evita descargar las páginas que remiten a archivos multimedia (esto es, que contienen la cadena "multimedia"; b) puede modificarse para seleccionar las lenguas cuyas páginas se desea descargar (en nuestro caso, DE, ES y EN); c) guarda cada página con un nombre de archivo que corresponde a la última parte de la *URL* de la página que contiene el artículo correspondiente y d) añade el prefijo *home-* o *professional-* al nombre del archivo, en función del subdirectorio al que pertenezca la *URL* (v. Ilustración 4)[15]. Los archivos se descargan en carpetas por lenguas: html-de, html-es y html-en. Resulta interesante que los archivos .html de las tres carpetas tienen el nombre en inglés. Esto facilitará el proceso de alineación.

[15] Creamos una copia del *script* así modificado y la denominamos *downloadMSDhtmls-de-y-es.py*.

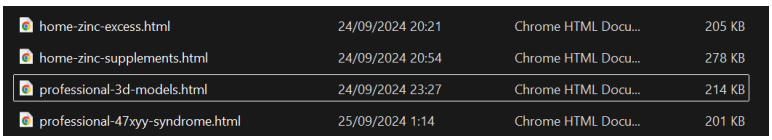



Ilustración 4. Archivos .html con prefijo identificador de la versión (profesional o para el público general) en la carpeta de descarga de archivos .html

*4.3 Extracción de texto plano y metadatos*

Para este cometido recurrimos al *script Merck2Text_info.py*[16] (Oliver González 2024c), una adaptación de *MTUOC-downloadedweb2text.py* (Oliver González 2023e). Ambos *scripts* normalizan los archivos .html, esto es, los convierten a texto plano; pero *Merck2Text_info.py*, además, extrae los metadatos autor, *URL*, fecha de última revisión, fecha de última modificación, nombre del archivo (parte de la *URL*). En la orden, tenemos que escribir a) el directorio de entrada, esto es la carpeta en la que están los archivos .html (en nuestro caso, html-de, html-es); b) el directorio de salida, esto es, la carpeta en la que vayamos a guardar los archivos de texto (en nuestro caso, merck-text-de, merck-text-es; c) el nombre de archivo en el que el programa tiene que escribir los metadatos (merck-metadata-de.txt, merck-metadata-es.txt), como se observa en las siguientes líneas de comando:

```
python3 Merck2Text_info.py html-de merck-text-de merck-metadata-de.txt
python3 Merck2Text_info.py html-es merck-text-es merck-metadata-es.txt
```

*4.4 Alineación de documentos txt*

El proceso de alineación de los documentos .txt de los directorios merck-text-de y merck-text-es lo llevamos a cabo en tres fases:

a) segmentación de los archivos .txt mediante el *script MTUOC-segmenterDIR.py* (Oliver González 2023c), que toma los archivos contenidos en un directorio, genera versiones segmentadas y las coloca en un nuevo directorio. Utilizamos el siguiente comando[17]:

```
python3 MTUOC-segmenterDIR.py -i merck-text-de -o merck-text-seg-de -s segment.srx -l German -p
```

[16] Los nombres de algunos scripts y de las carpetas mencionados de aquí en adelante contienen la denominación "Merck" en referencia a la corporación responsable de la edición de los Manuales MSD.

[17] Para este paso y para los que siguen, ejemplificaremos escribiendo solo el comando para la segmentación de los archivos del directorio alemán. Para que el comando sea más legible, simplificamos la ruta al directorio.

El parámetro -i especifica el directorio de entrada (el que contiene los archivos .txt), -o especifica el directorio de salida (el que contiene los archivos .txt segmentados); -s indica el archivo que contiene las reglas de segmentación, -l indica la lengua en la que están los archivos; -p indica que queremos que nos coloque una marca de párrafo.

b) Creación del archivo *batch*. A continuación, generamos un archivo *batch* como paso previo a la alineación. Este archivo organiza los pares de archivos segmentados que van a ser alineados e identifica aquellos que no podrán alinearse (porque solo existen en una de las dos carpetas); para cada par "archivo origen-archivo meta", el archivo *batch* contiene una línea con las rutas respectivas a cada archivo del par, así como la ruta al archivo de salida con la alineación de ambos. Para crear el archivo *batch*, nos servimos del programa *MTUOC-create-batchfile.py* (Oliver González 2023f), con el comando:

python3 MTUOC-create-batchfile.py --dirSL merck-text-seg-de --dirTL merck-text-seg-es --dirALI merck-text-ali --batchfile batch-merckDeEs.txt

Los parámetros --dirSL y --dirTL especifican, respectivamente, los directorios donde se encuentran los que consideramos textos origen y textos meta. El parámetro --dirALI especifica el directorio donde se pondrán los archivos alineados, el parámetro --batchfile sirve para indicar el nombre que ha de tener el archivo *batch* que estamos creando.

c) Ejecución de la alineación. La alineación propiamente dicha se realizó con el programa *Hunalign* (Varga *et al.* 2008). Este programa, así como otros auxiliares desarrollados dentro del proyecto MTUOC, están disponibles para su descarga en la dirección https://github.com/mtuoc/MTUOC-aligner/tree/main. *Hunalign* utiliza diccionarios bilingües para mejorar la precisión de la alineación. Por eso, tal y como se propone en el punto 1.1 de Oliver González (2024e), creamos un diccionario de alineación para el par DE-ES (de-es_MUSEdic.txt), que incluimos en la línea de comando, así como el archivo *batch* creado previamente, esto es, batch-merckDeEs.txt

./hunalign -text -utf -realign -batch de-es_MUSEdic.txt batch-merckDeEs.txt

Tras esta operación comprobamos que el directorio de alineaciones merck-text-ali tenía 6.373 archivos, pero el archivo batch-merckDeEs tenía 6378 líneas, lo cual indica que al menos cinco archivos de los inicialmente emparejados finalmente no pudieron alinearse.

*4.5 Limpieza de archivos 1 y separación de subcorpus (home, professional)*

Tras una primera inspección visual del contenido de los archivos alineados, se observó que había algunos elementos que era necesario eliminar, como la

cadena <p>/t<p> (correspondiente a las marcas de párrafo introducidas voluntariamente en el proceso de alineación).

Nuestra aproximación consistió en crear una carpeta de pruebas, denominada merck-text-ali-clean-pruebas y colocar en ella todos los archivos alineados cuyo nombre comenzaba con la letra "a". En esa carpeta colocamos también el *script clean_texts.py*, resultado de pedir a Copilot que generara un código para limpiar los elementos descritos arriba. Fue necesario probar el código en varias rondas y solicitar a la IA que lo refinara hasta obtener resultados satisfactorios. En este proceso de prueba y error, aprendimos que es importante incluir en el código un paso que especifique que los archivos originales se han de mantener y los modificados se han de colocar en una subcarpeta: así siempre tendremos el material en el estadio previo a la modificación, por si algo sale mal.

A continuación, separamos los archivos alineados (el corpus bilingüe en bruto) en los dos subcorpus que han de representar los dos registros comunicativos en comunicación biosanitaria en español y en alemán: merck-de-es-home y merck-de-es-professional. Situándonos en el directorio padre que contiene el subdirectorio merck-text-ali-clean, creamos primero subdirectorios para el corpus *home* y el *professional*:

mkdir merck-text-ali-clean-home

mkdir merck-text-ali-clean-professional

Acto seguido, pasamos los archivos respectivos a cada subdirectorio:

cp merck-text-ali-clean/ali-home-*.txt merck-text-ali-clean-home/

cp merck-text-ali-clean/ali-professional-*.txt merck-text-ali-clean-professional/

*4.6 Limpieza de archivos 2*

Fue necesario un segundo proceso de limpieza, en este caso no de elementos dentro de archivos, sino de un conjunto de archivos. Hasta este momento, no habíamos detectado una serie de archivos que contenían nombres propios en su denominación (p. ej. "ali-home-aaron-denise.txt"). Inspeccionamos el contenido de esos archivos y solicitamos a Mistral.ai que nos escribiera un código que leyera el contenido de los archivos buscando esos contenidos típicos (v. "líneas específicas a buscar" en la Ilustración 5) y aquellos que localizara, los moviera a un nuevo subdirectorio específico, denominado "autores". El *script* lo denominamos *mover_archivos_por_contenido.py*.

```
File  Edit  Format  Run  Options  Window  Help
import os
import shutil

def mover_archivos_con_lineas_especificas(ruta_carpeta):
    # Crear la subcarpeta si no existe
    subcarpeta = os.path.join(ruta_carpeta, 'archivos_autores')
    if not os.path.exists(subcarpeta):
        os.makedirs(subcarpeta)

    # Líneas específicas a buscar
    lineas_a_buscar = [
        "Fachgebiete und Expertise\tEspecialidades y experiencia\t0.2625",
        "Verbindungen\tAfiliaciones\t0.3",
        "Kapitel und Kommentare für Manuals\tCapítulos y comentarios del manual\t0.3",
        "Kapitel\tCapítulos\t0.225"
    ]

    # Recorrer todos los archivos en la carpeta
    for archivo in os.listdir(ruta_carpeta):
        ruta_archivo = os.path.join(ruta_carpeta, archivo)
        if os.path.isfile(ruta_archivo):
            try:
                # Leer el contenido del archivo
                with open(ruta_archivo, 'r', encoding='utf-8') as f:
                    contenido = f.read()

                # Verificar si alguna de las líneas específicas está en el contenido
                if any(linea in contenido for linea in lineas_a_buscar):
                    # Mover el archivo a la subcarpeta
                    shutil.move(ruta_archivo, os.path.join(subcarpeta, archivo))
            except Exception as e:
                print(f"Error al procesar el archivo {archivo}: {e}")

# Ejemplo de uso
ruta_carpeta = 'merk-text-ali-clean-professional'
mover_archivos_con_lineas_especificas(ruta_carpeta)
```

Ilustración 5. *Script mover_archivo_por_contenido.py*

Este *script* lo ejecutamos tanto en la carpeta del subcorpus *home* como en la del subcorpus *professional*. A posteriori se borraron manualmente dos archivos cuyo nombre comenzaba por "about" y que no contenían artículos en sí.

*4.7 Gestionar metadatos*

Este paso comprendió dos operaciones. En primer lugar, fue necesario comprobar si existen metadatos para todos los archivos y, al darse el caso contrario, hacer los filtrados correspondientes. En segundo lugar, fue necesario asegurar un sistema que permita relacionar cada línea de cada archivo de alineación con los metadatos del archivo al que corresponda.

4.7.1 Filtrado de archivos sin metadatos relacionados

Una inspección de los archivos de metadatos de los artículos de las versiones alemana y española (metadata-de.txt, merck-metadata-es.txt) extraídos de los .hmtl en el paso comentado en 4.3 permitió constatar que había 533 archivos para los que no existían metadatos. Se decidió no incluir estos archivos en los subcorpus, dado que la documentación bibliográfica de cada uno de los resultados de las consultas es uno de los rasgos distintivos de los corpus de PaCorES. Así, creamos con Mistral.ai un *script* que detectara aquellos archivos de las carpetas de cada uno de los subcorpus (merck-text-ali-clean-home y merck-

text-ali-clean-professional) para los que sí existían metadatos en los archivos mencionados (comparando los nombres de los archivos) y los pasara a unas nuevas carpetas (merck-text-ali-clean-home-metadata y merck-text-ali-clean-professional-metadata). El *script* se denomina *filter_files_by_metadata.py* (v. Ilustración 6).

```
File  Edit  Format  Run  Options  Window  Help
import os
import shutil

# Ruta a la carpeta con los archivos alineados
alineados_dir = '/home/maite_sanchez/AOLIVERG-downloadMSDhtmls-main/merk-text-ali-clean-professional'

# Ruta a los archivos de metadatos
metadata_de_file = 'merk-metadata-de.txt'
metadata_es_file = 'merk-metadata-es.txt'

# Ruta a la nueva carpeta donde se copiarán los archivos filtrados
output_dir = '/home/maite_sanchez/AOLIVERG-downloadMSDhtmls-main/merk-text-ali-clean-profesional-metadata'

# Crear la carpeta de salida si no existe
os.makedirs(output_dir, exist_ok=True)

# Leer los nombres de los archivos en los archivos de metadatos
with open(metadata_de_file, 'r', encoding='utf-8') as file:
    lineas_metadata_de = file.readlines()

with open(metadata_es_file, 'r', encoding='utf-8') as file:
    lineas_metadata_es = file.readlines()

# Extraer los nombres de los archivos de los metadatos, excluyendo el prefijo y la extensión
nombres_metadata_de = {linea.split('\t')[0].replace('.html', '').strip() for linea in lineas_metadata_de}
nombres_metadata_es = {linea.split('\t')[0].replace('.html', '').strip() for linea in lineas_metadata_es}

# Unir los nombres de los archivos de los metadatos
nombres_metadata = nombres_metadata_de.union(nombres_metadata_es)

# Leer los nombres de los archivos en la carpeta de archivos alineados
archivos_alineados = os.listdir(alineados_dir)

# Copiar los archivos que coinciden con los nombres de los metadatos
for archivo in archivos_alineados:
    nombre_archivo_sin_prefijo = archivo.replace('ali-', '').replace('.txt', '')
    if nombre_archivo_sin_prefijo in nombres_metadata:
        src = os.path.join(alineados_dir, archivo)
        dst = os.path.join(output_dir, archivo)
        shutil.copy2(src, dst)

print("Archivos copiados a merk-text-ali-clean-professional-metadata con éxito.")
```

Ilustración 6. *Script filter_files_by_metadata.py*

A continuación, creamos otro *script* que filtrara los metadatos de la selección de archivos que acabábamos de hacer para cada subcorpus (*filter_metadata_for_corpus.py*) y los guardara en documentos separados por idioma y lengua para cada subcorpus: merck-metadata-de-select-home.txt, merck-metadata-es-select-home.txt, merck-metadata-de-select-professional.txt y merck-metadata-es-select-professional.txt.

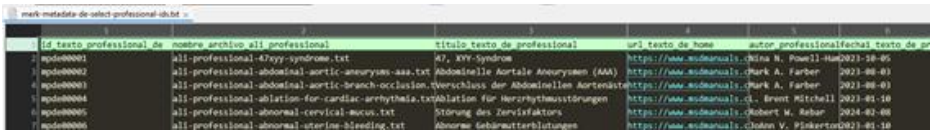
### 4.7.2 Creación de las claves para relacionar los metadatos con los archivos de texto

Al igual que ocurre en el resto de corpus de PaCorES, cada ejemplo extraído del corpus *MSD-Manuals DE-ES* debe contar con una referencia bibliográfica. En este caso, hemos de garantizar que el usuario pueda saber a qué artículo de la

versión española y de la versión alemana y a qué edición (para profesionales o para el público general) corresponde la concordancia recuperada.

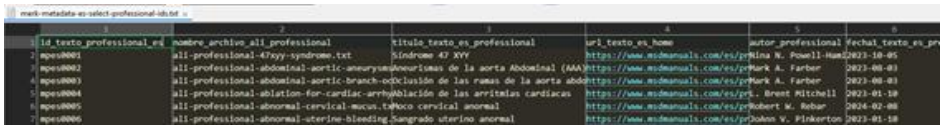
Para asegurar la recuperación de esta información según el modelo de los corpus de PaCorES, decidimos crear dos elementos relacionales.

Elemento1: clave de archivo. Generamos una nueva versión de los archivos de metadatos seleccionados (v. 4.7.1) a la que añadimos una nueva columna (la 1 en las dos ilustraciones inferiores) para incluir un identificador para cada texto español y alemán del subcorpus. A los nombres de los archivos que contenían esta versión de los metadatos les añadimos la coletilla "ids", p. ej.: merk-metadata-de-select-professional-ids.txt. La clave de archivo identifica el archivo como perteneciente al corpus MSD Manuals (m), a la edición profesional o *home* (mp o mh), la versión alemana o española de cada una de las ediciones (mpes, mpde, mhde, mhes) y, finalmente, el número de archivo (mpes00001).



| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| id_texto_professional_de | nombre_archivo_all_professional | titulo_texto_de_professional | url_texto_de_home | autor_professional | fechal_texto_de_pr |
| mpde00001 | all-professional-47xyy-syndrome.txt | 47, XYY-Syndrom | https://www.msdmanuals.c | Nina N. Powell-Hun | 2023-10-05 |
| mpde00002 | all-professional-abdominal-aortic-aneurysms-aaa.txt | Abdominelle Aortale Aneurysmen (AAA) | https://www.msdmanuals.c | Mark A. Farber | 2023-08-03 |
| mpde00003 | all-professional-abdominal-aortic-branch-occlusion.t | Verschluss der Abdominellen Aortenäste | https://www.msdmanuals.c | Mark A. Farber | 2023-08-03 |
| mpde00004 | all-professional-ablation-for-cardiac-arrhythmia.txt | Ablation für Herzrhythmusstörungen | https://www.msdmanuals.c | L. Brent Mitchell | 2023-01-10 |
| mpde00005 | all-professional-abnormal-cervical-mucus.txt | Störung des Zervixfaktors | https://www.msdmanuals.c | Robert W. Rebar | 2024-02-08 |
| mpde00006 | all-professional-abnormal-uterine-bleeding.txt | Abnorme Gebärmutterblutungen | https://www.msdmanuals.c | John V. Pinkerton | 2023-01-10 |

Ilustración 7. Ejemplificación de los metadatos de los textos alemanes del corpus *MSD-Manual DE-ES professional* con identificador de texto en columna 1



| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| id_texto_professional_es | nombre_archivo_all_professional | titulo_texto_es_professional | url_texto_es_home | autor_professional | fechal_texto_es_pr |
| mpes0001 | all-professional-47xyy-syndrome.txt | Síndrome 47 XYY | https://www.msdmanuals.com/es/pr | Nina N. Powell-Hun | 2023-10-05 |
| mpes0002 | all-professional-abdominal-aortic-aneurysms | Aneurismas de la aorta Abdominal (AAA) | https://www.msdmanuals.com/es/pr | Mark A. Farber | 2023-08-03 |
| mpes0003 | all-professional-abdominal-aortic-branch-oc | Oclusión de las ramas de la aorta abdo | https://www.msdmanuals.com/es/pr | Mark A. Farber | 2023-08-03 |
| mpes0004 | all-professional-ablation-for-cardiac-arrhy | Ablación de las arritmias cardiacas | https://www.msdmanuals.com/es/pr | L. Brent Mitchell | 2023-01-10 |
| mpes0005 | all-professional-abnormal-cervical-mucus.tx | Moco cervical anormal | https://www.msdmanuals.com/es/pr | Robert W. Rebar | 2024-02-08 |
| mpes0006 | all-professional-abnormal-uterine-bleeding. | Sangrado uterino anormal | https://www.msdmanuals.com/es/pr | John V. Pinkerton | 2023-01-10 |

Ilustración 8. Ejemplificación de los metadatos de los textos españoles del corpus *MSD-Manual DE-ES professional* con identificador de texto en columna 1

Elemento 2: marcador de capítulo. Para asegurarnos de que, una vez unidos todos los archivos del corpus *home* o del profesional en un único archivo respectivo, pudiera distinguirse dónde empiezan y acaban los bisegmentos de cada archivo de alineación, decidimos colocar el título de texto (metadato de la columna 3 en la Ilustración 8 y en la Ilustración 9) como marcador de capítulo, colocándolo en la primera línea de cada archivo y entre corchetes, tal y como se observa en la Ilustración 9. Para ello creamos el *script reorganizar_titulos_chapters.py*.

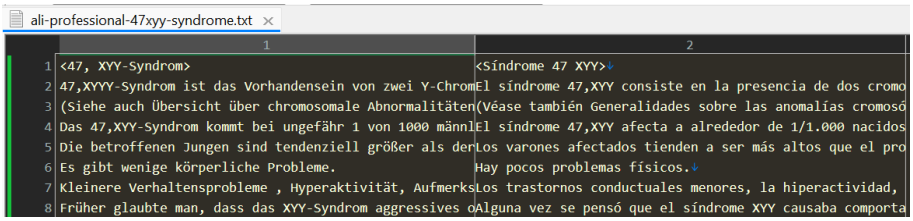



Ilustración 9. Nombre de archivo como marca de capítulo entre corchetes

En la práctica, gracias a este segundo elemento, al unir los dos corpus obtendremos dos grandes "libros" (el subcorpus *home* y el *profesional*) con tantos capítulos como archivos tiene cada uno de estos corpus. Además, una vez numerados los segmentos (v. 4.8) y tras la indexación y publicación del corpus en la web de *PaGeS*, el número de segmento correspondiente al título del capítulo cumple una función doble: 1) relaciona cada resultado de una búsqueda con el «libro» o versión del manual de la que proviene (*Manual MSD versión para el público general* o *Manual MSD versión para profesionales*); 2) como hipervínculo, da acceso a una segunda pantalla que permite ampliar contexto y que muestra el resto de metadatos con los que el usuario puede reconstruir la referencia bibliográfica de ese resultado concreto: autor del capítulo, títulos alemán y español del capítulo en las respectivas ediciones del manual y las URL que llevan al capítulo en las ediciones alemana y española (v. Ilustración 10).


Ilustración 10. Ejemplo de referencia bibliográfica
para concordancia en el suplemento *MSD-Manual DE-ES home*

*4.8. Numeración de los bisegmentos*

La numeración de las líneas de los archivos alineados consiste en añadir un identificador (ID) con la estructura bimembre nnnnn-nnnnnnn. La primera parte del ID identifica el subcorpus y la segunda el número de segmento. Este ID se añade como si de un prefijo se tratara, a cada línea de archivo y se separa mediante un tabulador (\t) del resto de la línea. Cada línea del archivo recibe un identificador único y consecutivo. De esta manera, las líneas pueden desordenarse y volverse a ordenar por la columna que contiene este ID y siempre puede recuperarse el contexto previo y posterior de cada una de las líneas.

Se tomaron las siguientes decisiones para conformar la ID de cada línea de archivo:

1) Primer grupo de la ID, de 5 dígitos (identificador de subcorpus, v. paso 7 en la Ilustración 1). Siguiendo la lógica de la numeración interna de *PaGeS*, estos primeros cinco dígitos identifican el corpus al que pertenece el bisegmento: 02 identifica que se trata de un suplemento; 3 es el número que identifica el contenido del suplemento como perteneciente al dominio de la medicina; 01 identifica el subcorpus *home* y 02 el subcorpus *professional* dentro de *MSD-Manuals DE-ES*. Así, 02301 identifica las líneas de los archivos ali-home* y 02302 las de los archivos ali-professional*.

2) Segundo grupo de la ID, de 7 dígitos (v. paso 8 en Ilustración 1). Se creó con ayuda de Mistral.ai un *script* para iterar la adición de 7 dígitos en todos los archivos de la carpeta y, a la vez, asegurar que los identificadores fueran consecutivos tanto dentro del archivo como entre archivos, esto es:

- el segundo segmento del primer archivo debería recibir el ID 02301-0000002, el tercero el ID 02301-0000003;
- si el ID de la última línea del primer archivo es ID 02301-0000010, entonces el ID de la primera línea del segundo archivo del directorio debería ser 02301-0000011.

El *script* que creamos lo denominamos *add_ids_copiar_archivos.py*, pues, además de numerar, crea unan copia de los archivos numerados en una nueva carpeta, manteniendo la versión de los archivos sin numerar (v. Ilustración 11).

```
File  Edit  Format  Run  Options  Window  Help
import os

# Initialize counter for IDs
counter = 1

# Directory containing the files
directory = "merk-text-ali-clean-home-metadata"

# Directory to save modified files
new_directory = "merk-text-ali-clean-home-metadata-id"

# Create the new directory if it doesn't exist
if not os.path.exists(new_directory):
    os.makedirs(new_directory)

# Iterate over each file in alphabetical order
for filename in sorted(os.listdir(directory)):
    filepath = os.path.join(directory, filename)

    # Read the content of the file
    with open(filepath, 'r') as file:
        lines = file.readlines()

    # Add ID to each line and update the counter
    new_lines = []
    for line in lines:
        id_str = f"02301-{counter:07d}"
        new_line = f"{id_str}\t{line}"
        new_lines.append(new_line)
        counter += 1

    # Define the new filepath in the new directory
    new_filepath = os.path.join(new_directory, filename)

    # Write the modified content to the new file
    with open(new_filepath, 'w') as file:
        file.writelines(new_lines)
```

Ilustración 11. *Script add_ids_copiar_archivos.py*

*4.9 Unir archivos*

Finalmente unimos todos los archivos de cada subcorpus en un solo archivo, obteniendo los archivos merck-professional-corpus-de-es.txt y merck-home-corpus-de-es.txt.

*4.10 Retos y resultados*

En el momento de abordar la revisión del alineado (para la cual es imprescindible tener numerados los segmentos), pudimos observar que la metodología de compilación seguida y expuesta en este trabajo presenta al menos dos retos importantes, derivados del origen del material que integra la base textual del corpus. Por un lado, en próximas compilaciones similares (p. ej. para los pares EN-ES o EN-FR), ha de incluirse en la primera fase de limpieza un paso para filtrar segmentos con caracteres exclusivamente numéricos procedentes de tablas, así como los que contienen datos de las referencias bibliográficas especializadas presentes en muchos de los capítulos de las ediciones de la versión para profesionales. Ambos tipos de elementos no revisten interés para el usuario

del corpus. También convendrá repasar cuidadosamente los resultados de la segmentación previa al alineado y modificar, en su caso, las reglas de segmentación a la vista de patrones de errores de segmentación detectados[18].

Así, tras limpiar los corpus de bisegmentos espurios como los mencionados, ha sido necesario volver a separar los archivos de alineación en archivos monolingües a partir de las columnas de texto alemán y texto español y proceder a una segunda alineación, en este caso llevada a cabo por miembros del equipo de PaCorES con el algoritmo Bertalign (Liu y Zhu 2023). Esta nueva alineación fue sometida a un segundo proceso de revisión, en el que dividimos los segmentos DE o ES de más de 400 caracteres y limpiamos más segmentos sin interés.

Los corpus obtenidos, que constituyen la versión 01 del suplemento *MSD-Manuals DE-ES*, presentan los siguientes datos generales:

| lengua | caracteres | palabras | tokens | MSTTRATIO | bisegmentos |
|---|---|---|---|---|---|
| alemán | 22.738.873 | 3.437.601 | 3.872.045 | 0,521 | 253.118 |
| español | 22.462.886 | 4.190.512 | 4.568.408 | 0,448 | |
| **Total *MSD-Manual DE-ES home*** | **45.201.759** | **7.628.113** | **8.440.453** | **0,484** | **253.118** |

Tabla 2. Suborpus *MSD-Manual DE-ES home:* datos generales

| lengua | caracteres | palabras | tokens | MSTTRATIO | bisegmentos |
|---|---|---|---|---|---|
| alemán | 31.083.327 | 4.475.198 | 5.083.860 | 0,549 | 330.656 |
| español | 30.054.001 | 5.402.968 | 5.970.625 | 0,467 | |
| **Total *MSD-Manual DE-ES profesional*** | **61.137.328** | **9.878.166** | **11.054.485** | **0,508** | **330.656** |

Tabla 3. Suborpus *MSD-Manual DE-ES profesional*: datos generales

| lengua | caracteres | palabras | tokens | MSTTRATIO | bisegmentos |
|---|---|---|---|---|---|
| alemán | 53.822.200 | 7.912.799 | 8.955.905 | 0,535 | 583.774 |
| español | 52.516.887 | 9.593.480 | 10.539.033 | 0,457 | |

[18] Observamos, por ejemplo, segmentaciones incorrectas tras el punto que sigue a las abreviaturas *z. B.* (*zum Beispiel* "por ejemplo") o *d. h.* (*das heißt* "es decir") o tras el punto que señala en alemán el número ordinal seguido de sustantivos como *Schwangerschaftswoche* "semana de embarazo".

| | | | | | |
|---|---|---|---|---|---|
| **Total *MSD-Manuals DE-ES*** | **106.339.087** | **17.506.279** | **19.494.938** | **0,496** | **583.774** |

Tabla 4. Tabla 6. Corpus *MSD-Manuals DE-ES* (suplemento completo): datos generales

Para calcular la densidad léxica de cada subcorpus y de los suplementos en su conjunto se ha utilizado la medida MSTTR (Mean Segmental Type-Token Ratio), que se considera una medida más fiable que el TTR (Type-Token Ratio) tradicional para evaluar la densidad léxica o riqueza léxica de un corpus (Paquot 2018).

Si comparamos el corpus *MSD-Manuals DE-ES* con el resto de los suplementos disponibles en la actualidad para *PaGeS*[19], vemos que *Merck-DE-ES-professional* y *Merck-DE-ES-home* suman un número de bisegmentos y de palabras similares a las del suplemento TED DE-ES, que cuenta con 321.924 bisegmentos y 9.800.701 palabras. Ambos suplementos se sitúan detrás de Open Subtitles DE-ES y de Europarl DE-ES.

Por subcorpus, podemos observar cómo el registro más divulgativo propio de *MSD-Manual DE-ES-home*, que recoge textos dirigidos al público en general /für Patienten, se relaciona con una densidad léxica general menor (0,484) que la de *MSD-Manual DE-ES-professional* (0,508), que reproduce un registro más especializado (para profesionales/*für Fachkreise*). En su conjunto, la densidad léxica del suplemento *MSD-Manuals DE-ES* (0,496) es menor que la del suplemento *TED-DE-ES* (0,525), con el que se presta a una comparación más estrecha. Probablemente este hecho se deba a que *MSD-Manuals DE-ES* es un corpus más enfocado en un campo de especialización (la medicina) y en un género concreto (el artículo de manual), con unas directrices estilísticas concretas (las dictadas por Merck & Co.), mientras que las charlas divulgativas a las que pertenecen los testimonios de *TED-DE-ES* están ligadas a diferentes ámbitos de conocimiento y reflejan las variaciones estilísticas del discurso de cada orador. Como cabía esperar, tanto *MSD-Manuals DE-ES* como *TED-DE-ES* muestran una densidad léxica menor que el corpus nuclear DE-ES (0,557), compuesto por texto editorial (ficción y no ficción).

*MSD-Manuals DE-ES* está disponible como suplemento en la web del corpus *PaGeS*: https://www.corpuspages.eu/corpus/search/advancedsearch.

Si se desea buscar solo en *MSD-Manual DE-ES home*, ha de introducirse la clave 02301 en el campo ID-Obra en el modo de búsqueda avanzado. Para buscar

[19] Los datos de los suplementos están disponibles en tablas como las aquí presentadas en la dirección https://www.corpuspages.eu/corpus/about/about?lang=es&format=.

en *MSD-Manual DE-ES professional* se usará la clave 02302. Si se desea buscar en ambos corpus, basta con marcar la casilla del suplemento (v. Ilustración 12).


Ilustración 12. Interfaz del corpus bilingüe *PaGeS*
con el suplemento *MSD-Manuals DE-ES* seleccionado[20]

## Conclusiones

En lo que se refiere a la parte técnica de la compilación del corpus *MSD-Manuals DE-ES*, descrita en esta aportación, es necesario subrayar que el trabajo con grandes masas de texto ha sido posible gracias al impulso de proyectos como MTUOC (v. Oliver García 2024f; ASETRAD 2024), que acercan las tecnologías de procesamiento de lenguaje natural a la comunidad investigadora en lingüística del corpus y en traducción. Por otra parte, la compilación de este corpus nos ha permitido apreciar cómo los desarrollos en inteligencia artificial permiten a los investigadores en lingüística y traducción, una vez se han familiarizado mínimamente con el lenguaje Python y al trabajo con sistemas operativos Linux, experimentar con código creado por un bot conversacional, aplicarlo y refinarlo a las necesidades concretas. No obstante, es necesario seguir invirtiendo en el aprendizaje de lenguajes de programación como Python o R para poder adoptar una posición más crítica con los códigos devueltos por los bots conversacionales y acortar los tiempos de prueba y error con los mismos. También así la metodología de compilación aquí presentada, una vez afinada, puede adaptarse para extraer datos paralelos de otros sitios web bi- o multilingües que reproduzcan otros usos especializados de las lenguas de trabajo y otras situaciones de comunicación mediada.

[20] En breve, el nombre del suplemento que aparece en pantalla ("Merck") se actualizará a "MSD Manuals".

**Referencias bibliográficas**

ANETI; ASPROSET, *Situación y perspectivas del mercado de los servicios lingüísticos y la comunicación multilingüe en España. Análisis estratégico*. Madrid: Universidad Complutense de Madrid, 2024.

ASETRAD, "Entrevista a Antoni Oliver, responsable del proyecto MTUOC para el entrenamiento, uso e integración local de la NMT". En: *La linterna del Traductor*, 27, 2024. Disponible en línea en: https://lalinternadeltraductor.org/n27/proyecto-mtuoc.html

BALBUENA TOREZANO, María del Carmen, "La adquisición del léxico veterinario (alemán-español): una propuesta basada en corpus para la traducción de textos". En: *Futhark. Revista de Investigación y Cultura,* 14, 2019, pp. 27-42. Disponible en línea en: https://doi.org/10.12795/futhark.2019.i14.02

BULLERS, Krystal, "Merck Manuals". En: *Journal of the Medical Library Association,* 104, 4, 2017, pp. 369-371. Disponible en línea en: https://doi.org/10.5195/jmla.2016.164

CETTOLO, Mauro; GIRARDI, Christian *et al.,* "Wit3: Web inventory of transcribed and translated talks". En: *Proceedings of the Conference of European Association for Machine Translation* (EAMT), 2012, pp. 261-268. Disponible en línea en: https://cris.fbk.eu/bitstream/11582/104409/1/WIT3-EAMT2012.pdf

DOVAL REIXA, Irene; SÁNCHEZ NIETO, M. Teresa, "Parallel Corpora Spanish (PaCorES): A Collection of Multifunctional Parallel Corpora". En: *RESLA. Revista Española de Lingüística Aplicada,* 39, 2, 2026, en prensa. Disponible en línea en: https://uvadoc.uva.es/handle/10324/76308

ELRC, "ELRC3.0 Multilingual corpus made out of PDF documents from the European Medicines Agency (EMEA)". 2020. Disponible en línea en: https://acortarurl.es/bGljol

GARCÍA JIMÉNEZ, Rocío; VARELA SALINAS, M. José, *Aspectos de la traducción biosanitaria español-alemán / alemán-español*. Berlín: Frank & Timme, 2021.

HOFFMANN, Michael, *Funktionale Varietäten des Deutschen - kurz gefasst*. Potsdam: Universitätsverlag Potsdam, 2007.

LIU, Lei; ZHU, Min, "Bertalign: Improved word embedding-based sentence alignment for Chinese-English parallel corpora of literary texts". En: *Digital Scholarship in the Humanities*, 38, 2, 2023, pp. 621-634. Disponible en línea en: https://doi.org/10.1093/llc/fqac089

MAYOR SERRANO, Blanca; QUIJADA, Carmen *et al.*, "¿Y por qué el alemán, a estas alturas?". En: *Panace@: Revista de Medicina, Lenguaje y Traducción*, *XII*, 34, 2011, pp. 159-160. Disponible en línea en: http://tremedica.org/panacea.html

MERCK&CO, "Conocimiento Médico Global". En: *Manual Merck. Versión para el público general*. 2025. Disponible en línea en: https://www.merckmanuals.com/es-us/hogar/resourcespages/global-medical-knowledge

——, *Manual Merck versión para profesionales*. 2025. Disponible en línea en: https://www.merckmanuals.com/es-us/professional

——, *Manual MSD versión para público general*. 2025. Disponible en línea en: https://www.msdmanuals.com/es/hogar

——, *MSDManual Ausgabe für Patienten*. 2025. Disponible en línea en: https://www.msdmanuals.com/de/heim

——, *MSD Manual Profi-Ausgabe*. 2025. Disponible en línea en: https://www.msdmanuals.com/de/profi

NAVARRO, Fernando A., *Libro Rojo. Diccionario de dudas y dificultades de traducción del inglés médico, Versión 4.07*. Madrid: Cosnautas, 2025. Disponible en línea en: https://www.cosnautas.com/es/libro

——, *Medizin. Gran diccionario médico alemán-español, Versión 1.17*. Madrid: Cosnautas, 2025. Disponible en línea en: https://www-cosnautas-com.ponton.uva.es/es/medizin#

OLIVER GONZÁLEZ, Antoni, "Entrenamiento de motores de traducción automática". En: Sánchez Ramos, María del Mar; Rico Pérez, Celia (eds.), *Traducción automática en contextos especializados*. Berlín: Peter Lang, 2023, pp. 33-70. Disponible en línea en: https://doi.org/10.3726/b20144

——, *MTUOC-download-from-sitemap.py*. 2023. Disponible en línea en: https://github.com/mtuoc/MTUOC-web-downloader/blob/main/MTUOC-download-from-sitemap.py

——, *MTUOC-segmenterDIR.py*. 2023. Disponible en línea en: https://github.com/mtuoc/MTUOC-segmenter/blob/main/MTUOC-segmenterDIR.py

——, *MTUOC-sitemap.py*. 2023. Disponible en línea en: https://github.com/mtuoc/MTUOC-web-downloader/blob/main/MTUOC-sitemap.py

——, *MTUOC-downloadedweb2text.py*. 2023. Disponible en línea en: https://github.com/mtuoc/MTUOC-web-downloader/blob/main/MTUOC-downloadedweb2text.py

——, *MTUOC-create-batchfile.py*. 2023. Disponible en línea en: https://github.com/mtuoc/MTUOC-aligner/blob/main/MTUOC-create-batchfile.py

——, *downloadMSDhtmls.py*. 2024.

——, "GitHub - Mtuoc/MTUOC-Aligner: Scripts and Programs to Automatically Align Text Files Using Hunalign or SBERT". 2024. Disponible en línea en: https://github.com/mtuoc/MTUOC-aligner

——, *Merk2Text_info.py"*. 2024.

——, *MTUOC-Create-Batchfile.py*. 2024.

——, "Semana 3. Creación de corpus paralelos (II). Descarga de sitios web". En: *MTUOC. Seminario Online TAN*. 2024. Disponible en línea en: https://github.com/mtuoc/seminario_online_TAN/wiki/Semana-3.-Creación-de-corpus-paralelos-(II).-Descarga-de-sitios-web

——, *MTUOC project*. 2024. Disponible en línea en: https://mtuoc.github.io/

PAQUOT, Magali, "Corpus research methods for language teaching and learning". En: Aek Phakiti; Peter de Costa *et al.* (eds.), *The Palgrave handbook of applied linguistics research methodology*. Londres: Macmillan, 2018, pp. 359-374.

RAMÍREZ ALMANSA, Isidoro, "La traducción alemán-español de textos médico-jurídicos y su utilidad didáctica: el consentimiento informado". En: *Quaderns de Filologia: Estudis Lingüístics*, XXIV, 2019, pp. 229-245. Disponible en línea en: https://doi.org/10.7203/QF.2

RICO PÉREZ, Celia; GARCÍA ARAGÓN, Álvaro, *Análisis del sector de la traducción en España*. Villaviciosa de Odón: Universidad Europea, 2016. Disponible en línea en: https://abacus.universidadeuropea.com/rest/api/core/bitstreams/18e00a31-18c0-4c6d-911f-b5e78bc42e53/content

TOMES, Nancy, *Remaking the American Patient: How Madison Avenue and modern medicine turned patients into consumers*. Chapel Hill: UNC Press Books, 2016.

———, '"Not just for doctors anymore': How the Merck Manual became a consumer health 'bible"'. En: *Bulletin of the History of Medicine,* 95, 1, 2021. Disponible en línea en: https://doi.org/10.1353/bhm.2021.0000

VARGA, Daniel; HALÁCSY, Peter *et al.,* "Parallel corpora for medium density languages". En: *Recent advances in natural language processing IV: selected papers from RANLP 2005*. Ámsterdam: Benjamins, 2008, pp. 247-258.