Reconocimiento de entidades en corpus de dominios específicos: experimentación con periódicos históricos

Trabajo de Fin de Máster

Eva Sánchez Salido

Trabajo de investigación para el
Máster Universitario en Tecnologías del Lenguaje

Universidad Nacional de Educación a Distancia
Escuela Técnica Superior de Ingeniería Informática

Dirigido por la

Dra. Ana García Serrano

Septiembre de 2022

Agradecimientos

Gracias a Ana por ser mi directora y mentora durante este año y guiarme en mis primeros pasos en el mundo de la investigación.

Gracias al resto de compañeros y compañeras de la UNED, la UAM y el CSIC: a los *seniors* por sus inestimables consejos y ánimos, y a los *juniors* por el gusto que da trabajar con ellos.

Gracias a mis padres por haberme apoyado siempre en mis decisiones.

Gracias a Klein por ser mi compañero durante los últimos años y apoyarme en cada momento.

Gracias a todas mis amigas y amigos que, aunque a veces esporádicamente por cuestiones de tiempo o distancia, me demuestran siempre su cariño y me inspiran para crecer. A los de toda la vida y a los que he encontrado por el camino, gracias.

Resumen

En este trabajo se aborda el problema de la extracción de información en textos históricos a través de un caso de estudio[1] producto de una colaboración entre informáticos y humanistas, cuyo objetivo es facilitar la labor investigadora de los historiadores del arte mediante el reconocimiento de entidades en un dominio específico. Esta resulta una tarea clave para sus investigaciones, que además supone un reto en el ámbito del Procesamiento del Lenguaje Natural y las Humanidades Digitales. Así, este trabajo plantea abordar este problema desde la fase de construcción y etiquetado de un corpus dedicado a tal efecto, hasta la experimentación con los modelos de aprendizaje profundo más actuales, tanto en el nuevo corpus como en otros corpus de naturaleza similar. El trabajo se desarrolla mostrando en primer lugar los pasos para construcción y anotación de un corpus nuevo, junto con una descripción de las herramientas actuales. Después se contextualizan las nuevas tecnologías de aprendizaje automático y se llevan a cabo los experimentos, concluyendo con unas reflexiones sobre los resultados alcanzados y cómo continuar en un futuro próximo.

[1] En el marco del proyecto CLARA-HD de la UNED (PID2020-116001RB-C32).

Abstract

This work addresses the problem of information extraction in historical texts through a case study[2] resulting from a collaboration between computer scientists and humanists. The project aims to facilitate the research work of art historians through the recognition of entities in a specific domain, which happens to be a key task for their research, and also poses a challenge in the field of Natural Language Processing and Digital Humanities. Thus, this work proposes to address this problem from the construction and tagging phase of a corpus, to the experimentation with the cutting edge deep learning models in the new corpus. First, steps for construction and annotation of a new corpus are shown in detail, together with a description of the currently used tools. Then the new machine learning technologies are contextualized and the experiments are carried out, concluding with some considerations on the results achieved and how to continue in the near future.

[2] Within the framework of the CLARA-HD project of the UNED (PID2020-116001RB-C32).

Índice general

Capítulo 1. Introducción

1.1 Motivación

El planteamiento inicial del trabajo que aquí se presenta es el estudio de la extracción de información a partir de documentos antiguos para su posterior análisis en el ámbito histórico, partiendo de una colección de documentos digitalizados, pero no explotados previamente. Uno de los componentes clave para la extracción de información es el reconocimiento de entidades en el dominio específico del corpus. Actualmente, esta es una tarea que alcanza resultados comparables al rendimiento humano en dominios generalistas e idiomas con altos recursos como el inglés, gracias a las recientes tecnologías basadas en aprendizaje profundo, pero que aún mantiene diferentes retos en dominios específicos debido a la escasez y mala calidad de los datos, y el coste que conlleva producirlos, entre otros (Hu et al., 2020; Li et al., 2022). Por ello, hoy en día se considera a menudo un problema que recae más sobre los datos que sobre los algoritmos.

Para los textos anteriores al siglo XX en español, cuando las normas de escritura aún no estaban normalizadas, hay que decidir cómo obtener transcripciones fidedignas y así gestionar tecnológicamente el corpus y recuperar información semántica. En el caso de corpus digitalizados automáticamente desde textos antiguos (donde se trabaja con manuscritos, papiros, papel poco conservado o semidestruido), además suele ser necesaria una etapa de pre-procesamiento para, por ejemplo, eliminar los errores en la transcripción provocados por manchas provenientes de la digitalización de los textos, e identificar correctamente la ortografía.

Esta es la situación en la que muchos investigadores humanistas, como los historiadores, se encuentran ante un material específico de su interés y que solicitan soporte informático adecuado para su trabajo o investigación concreta. Decidir cuál es ese tipo de soporte exige afrontar este problema con una metodología y unas herramientas adecuadas al dominio específico.

Por otra parte, la tecnología lingüística o el procesamiento del lenguaje natural (PLN) se encuentra suficientemente avanzada para resolver algunas tareas de extracción o recuperación de información que podrían adaptarse y aplicarse a

estos corpus no generalistas además de la identificación de entidades, como son la simplificación de partes del texto, la identificación de autoría, las búsquedas semánticas basadas en ontologías de dominio, el enriquecimiento de la información desde fuentes externas, y otras.

El problema general abordado en este trabajo pretende acompañar la tarea de los humanistas y experimentar el proceso de partir de un recurso digitalizado de interés para unos historiadores del arte y un enigma histórico que desean investigar, identificando el tipo de soporte informático necesario. Nos preguntamos, pues, cuáles son los pasos que deben darse para aportar tecnología a su investigación, existente o a desarrollar.

1.2 Propuesta y objetivos

Para abordar un proyecto de investigación interdisciplinar entre técnicos (informáticos especialistas en PLN y minería de datos) y humanistas es necesario conocer a fondo ambos dominios, aunque habitualmente sean a su vez objeto de la investigación. Los objetivos específicos del trabajo que aquí se presenta son los siguientes.

OBJETIVO 1: Estado del arte sobre los corpus digitales

Tras una etapa de estudio tanto de la contextualización de la tecnología existente en gestión de corpus como en las nuevas tecnologías del PLN, se planteará un caso de estudio a partir de los objetivos de investigación de un equipo de humanistas, en este caso historiadores del arte.

Actualmente, las nuevas tecnologías para extracción y recuperación de información se basan en métodos de PLN y exigen tener disponibles corpus o conjuntos de textos "suficientes". Por esto, se estudiarán a nivel teórico-práctico las propiedades de los corpus, se explorarán las herramientas de gestión de corpus y se seleccionarán las más adecuadas para el caso de estudio en un dominio específico.

OBJETIVO 2: Generación y gestión de corpus

En primer lugar, se comprobará si el recurso disponible inicialmente tiene la calidad necesaria, o se debe construir un corpus con calidad suficiente para las posibles tareas de tipo lingüístico o de conocimiento semántico a realizar automática o semiautomáticamente de acuerdo con los objetivos del dominio específico del caso de estudio.

El recurso inicial del caso de estudio de este trabajo[3], es el archivo digitalizado del Diario de Madrid (siglos XVIII y XIX) de la Hemeroteca Digital de la BNE[4]. Se justificará entonces el desarrollo de un nuevo corpus denominado *CLARA-DM* a partir del recurso inicial tras enriquecerlo y aplicarle una serie de procesos.

Dado que el objetivo principal a realizar en este trabajo se basa en el reconocimiento de entidades nombradas en dominios específicos, se plantea este tercer objetivo.

Objetivo 3: Estado del arte en el reconocimiento de entidades nombradas

Una vez establecidos los límites y naturaleza del corpus CLARA-DM, para llevar a cabo los experimentos se realizará un estudio de las tecnologías que lideran hoy en día el campo del reconocimiento de entidades y una contextualización histórica. Se pondrán en evidencia las sinergias entre este campo y el de las Humanidades Digitales, donde se presentan problemas abiertos que guían investigaciones actuales. Con ello, y se plantearán una serie de experimentos.

Objetivo 4: Experimentación con modelos actuales en el corpus CLARA-DM

Finalmente, se llevarán a cabo distintos experimentos en el campo de las HD, acercando los trabajos de un dominio específico, el de periódicos antiguos en diferentes lenguas (francés, alemán e inglés), al dominio específico del corpus CLARA-DM.

1.3 Estructura del documento

En el capítulo 2 de esta memoria se presentan los conceptos relacionados con los corpus y las propiedades que deben verificar para asegurar la validez de los resultados (representatividad y equilibrio). Se aporta una categorización de los

[3] La autora de este trabajo ha sido parcialmente financiada por el proyecto de investigación CLARA-HD (PID2020-116001RB-C32, MÉTODOS DE LA LINGÜISTICA COMPUTACIONAL PARA LA LEGIBILIDAD Y SIMPLIFICACION AUTOMATICA EN HUMANIDADES DIGITALES), financiado por el Ministerio de Ciencia e Innovación. Convocatoria 2020 de «proyectos I+D+i» en el marco del programa estatal de generación de conocimiento y fortalecimiento científico y tecnológico del sistema de I+D +i y del programa estatal de I+D +i orientada a los retos de la sociedad, (2021-2024).

[4] https://hemerotecadigital.bne.es/hd/card?oid=0001510462

corpus y se clasifican algunos de los más relevantes. Además, se revisan herramientas para la gestión de corpus.

A continuación, en el capítulo 3, se presenta la tarea de reconocimiento de entidades y se revisa la evolución de las tecnologías de PLN utilizadas hasta la actualidad. Se introducen las Humanidades Digitales, área de estudio en la intersección entre los estudios humanísticos y las tecnologías informáticas, y se presenta un caso de estudio, la construcción de un corpus histórico en el proyecto CLARA-HD. Después se discuten las dificultades que se encuentran en este caso de estudio, y se expone la propuesta de experimentación de este trabajo en la tarea de reconocimiento de entidades nombradas.

En el capítulo 4 se describen las pruebas realizadas y se discuten los resultados. Finalmente, se realizan unos comentarios sobre el trabajo realizado y los resultados obtenidos y propuestas de trabajo futuro.

Capítulo 2. Sobre los corpus

La *minería de datos* es un campo de la estadística y las ciencias de la computación que engloba un conjunto de técnicas orientadas a la extracción de conocimiento procesable implícito almacenado en grandes volúmenes de datos. Un área particular dentro de este campo es la *minería de textos*, que implica el descubrimiento de información sobre textos desestructurados y escrita en un lenguaje natural, como son los contenidos en ficheros de texto o en Internet (Weiss, Sholom M. et al., 2005). Se trata de un área multidisciplinar basada fundamentalmente en las tareas de recuperación de información, la clasificación automática y el *clustering*, utilizando principalmente técnicas estadísticas, aprendizaje automático y otras tecnologías de la lengua.

Entre las aplicaciones y tareas asociadas a la minería de textos destacan:

- la recuperación de información, que consiste en buscar documentos, información dentro de documentos y metadatos que describan los documentos, e incluye la búsqueda en bases de datos,
- la extracción de información, cuyo objetivo es extraer automáticamente información estructurada, bien definida y clasificada de un cierto dominio a partir de texto plano. Esta tarea se suele descomponer, a su vez, en varias etapas de procesamiento que incluyen la tokenización, el etiquetado gramatical y la identificación de entidades nombradas, entre otros,
- la búsqueda de respuestas, que en realidad es un problema de recuperación de información en el que, dada una colección de documentos, el sistema debe ser capaz de extraer respuestas a consultas en lenguaje natural. Constituye una aplicación directa de la tarea de reconocimiento de entidades, ya que algunos de los sistemas de búsqueda de respuestas clasifican las preguntas de varios tipos y procesan su base de datos para extraer las entidades marcadas con el tipo de respuesta predefinido al que corresponden,
- la categorización de documentos, que consiste en asignar a un documento una o varias categorías (definidas previamente) en función de su contenido,
- el agrupamiento de documentos, conocido como *clustering*, un tipo de organización de documentos en grupos en el que ni la naturaleza de los grupos ni, en ocasiones, el número de ellos están definidos de antemano,

- la generación automática de resúmenes, que consiste en la transformación del texto original en uno resumido a través de la reducción de contenido mediante selección y/o generalización de contenido importante, y
- la simplificación de textos léxica (sustitución de palabras complejas) o basada en sentencias (modificación o reemplazo de frases complejas) con diferentes objetivos para la accesibilidad a la información.

Ahora bien, el primer paso en la minería de textos es la recolección de los documentos "relevantes" para la tarea que alcance el objetivo planteado. En la investigación y el desarrollo de técnicas de minería de datos, se denomina *corpus* a estas colecciones de documentos y, dependiendo del tipo de técnicas que se deseen aplicar, será necesario disponer de mayor o menor cantidad de textos, de manera más o menos ordenada, con metadatos o no, etc. Todas estas decisiones se encuadran en el proceso de diseño de corpus, que se describe en este capítulo.

2.1 Conceptos básicos de la lingüística basada en corpus

La información textual normalmente conforma documentos, que pueden agruparse en colecciones y en corpus, que a su vez a menudo contienen anotaciones y metadatos.

Un **corpus** es una colección estructurada de textos en formato electrónico, representativos de una variedad lingüística, un período histórico u otro fenómeno particular, destinado a servir como base de una investigación. Esta definición presupone varias cosas, tal y como apunta (Rojo, 2016): los textos deben ser naturales, es decir, no artificiales ni creados expresamente para su incorporación al corpus; tienen que estar en formato electrónico para poder recuperar información a partir de ellos, han de ser representativos de su género y deben permitir su estudio científico (no sólo lingüístico), lo que suele implicar la adición de información gramatical, léxica o pragmática. Esto es, un corpus se construye con la intención de reunir determinados requisitos en cuanto a la procedencia de los textos, la época a la que pertenecen, o el tema que tratan, porque el objetivo es obtener el perfil general de ciertas características de la lengua o la sociedad y el modo en que se presentan a lo largo de esta selección de textos.

En el uso cotidiano, se usan la palabra y el concepto de *corpus* (o de *corpus textual*) de manera poco precisa, pero la peculiaridad de los mismos reside en su selección y ordenación. Por ello es preciso distinguirlos propiamente de los

archivos informatizados y las *bibliotecas electrónicas*, ya que las recopilaciones textuales de estos tipos no implican una selección o una ordenación de los textos siguiendo criterios lingüísticos u otros, mientras que la de los corpus textuales sí (Torruella Casañas, 2017).

Una **anotación** es una información adicional asociada a un punto particular de un documento. El proceso de anotación consiste en la aplicación de un esquema para textos, que puede incluir etiquetas estructurales, etiquetado gramatical, análisis sintáctico y otras.

Los **metadatos** son datos sobre otros datos, y suelen utilizarse para facilitar la comprensión, el uso y el manejo de los datos. Por ejemplo, en un corpus de textos literarios, los metadatos básicos que cada texto debería contener son título, autor y fecha de composición o de publicación (Calvo Tello, 2019).

La evolución de la investigación con corpus textuales ha estado determinada por la que han experimentado los ordenadores y los medios de almacenamiento de versiones electrónicas o digitales de los textos. Así, hemos pasado del *Brown Corpus* (Kučera & Francis, 1979) en 1967, formado por un millón de palabras y la posibilidad de consultarlo sólo en el ordenador en que residía, a cualquiera de los corpus que manejamos hoy en día, formados por cientos o miles de millones de palabras, etiquetadas y lematizadas automáticamente y que pueden ser consultados a través de internet desde cualquier lugar. También ha aumentado considerablemente la complejidad de los análisis que es posible realizar: mientras que el Brown Corpus sólo ofrecía resultados totales, hoy realizamos búsquedas selectivas de información que permiten estudiar los corpus de manera mucho más extensiva y detallada. En respuesta a este aumento de posibilidades de acceder y analizar la información, la tipología de los corpus también se ha diversificado enormemente (Rojo, 2016).

Si bien en los ochenta la investigación en lingüística computacional estaba dirigida por los lingüistas gracias a las aportaciones de las gramáticas de unificación y rasgos, a partir de los noventa la disciplina ha estado dominada por los informáticos, y más en concreto por los modelos estadísticos (Moreno Sandoval, 2019). Entre 1980 y 1995 aparecen los primeros corpus en español concebidos con criterios comparables a los que ya se estaban aplicando en otras lenguas. La Real Academia Española (RAE) tomó en 1995 la decisión de construir el *Corpus de Referencia del Español Actual* (CREA) (Real Academia Española: Banco de datos, 2021) que cuenta con unos 160 millones de formas procedentes de textos tanto escritos como orales de todos los países hispánicos, y pocos meses después

el *Corpus Diacrónico del Español* (CORDE) (Real Academia Española: Banco de datos, 2008), con algo más de 250 millones de formas. Mientras que el CREA contiene textos publicados desde 1975 hasta 2004, el CORDE lo complementa con textos hasta 1974 desde los orígenes de la lengua. Después, la necesidad de ampliar y mejorar sus bancos de datos llevó a la RAE a construir el *Corpus del Diccionario Histórico* (CDH) (Real Academia Española: Banco de datos, 2009) pensado para ser el fondo documental del *Nuevo Diccionario Histórico del Español* y del *Corpus del Español del Siglo XXI* (CORPES XXI) (Real Academia Española: Banco de datos, 2021). Así, en español predominan los corpus de referencia como el CREA, el CORDE, el CORPES y el *Corpus del Español* (CdE) (Davies, 2002), que permiten estudiar fenómenos a lo largo del tiempo y del espacio y no sólo de manera general (Rojo, 2016).

En definitiva, si bien la lingüística de corpus en español comenzó de forma relativamente tardía, ha experimentado un desarrollo muy rápido e intenso precisamente por el hecho de haber iniciado su desarrollo en un marco tecnológico más evolucionado.

2.1.1 Parámetros clasificatorios de los corpus

Cuando se lleva a cabo la construcción de un corpus es importante decidir las características deseables que este ha de tener según los objetivos de la investigación. Hemos de responder a las preguntas: ¿qué necesidades debe cubrir el corpus? ¿qué tipo de análisis se espera realizar? ¿qué resultados cabe obtener?

En los corpus escritos la selección de características se puede hacer desde distintos parámetros: el porcentaje y la distribución de los distintos tipos de textos, su especificidad y tamaño, la cantidad de anotaciones y metadatos, etc. En general, la clasificación no responde a un solo parámetro, sino que es la suma de diversos factores. A continuación, se presenta un resumen de los principales parámetros, basada en, pero no limitada a, la clasificación que se propone en (Torruella Casañas, 2017).

1. Modalidad del discurso

Según la modalidad del discurso, los corpus pueden ser **orales** cuando consisten en transcripciones de grabaciones, como el *Corpus oral y sonoro del Español Rural* (COSER) (Fernández-Ordóñez & Pato, 2020), **escritos** o **textuales** cuando consisten en colecciones de textos almacenados en formato electrónico (seleccionados con criterios específicos según su finalidad y con el objetivo de ser

una muestra representativa de la lengua a analizar), como el CDH, o **mixtos** cuando recopila tanto discursos orales como escritos, como el CORPES XXI.

2. Temática y finalidad

Cuando un corpus tiene como finalidad estudiar la lengua en su totalidad, se trata de un corpus **general**. Estos pretenden reflejar la lengua en su ámbito más amplio, recogiendo cuantas más variedades distintas mejor, por lo que debe ser suficientemente amplio. Un ejemplo es el corpus *Tesouro Medieval Informatizado Da Lingua Galega* (TMILG) (Varela et al., 2018). Dentro de este tipo de corpus también se pueden situar los corpus **de referencia**, que se presentarán más adelante.

Si el objetivo es estudiar algunos aspectos concretos de la lengua como un tema, un registro o un dialecto, se trata de un corpus **especializado**. Ejemplos de este tipo son el *Corpus de Documentos Españoles Anteriores a 1800* (CODEA) (Grupo de Investigación Textos para la Historia del Español [GITHE], 2017), o el *Corpus de Documentación Cancilleresca del s. XIII* (CODCAR) (Sánchez González de Herrero et al., 2010), que estudia la lengua de la administración. Dentro de los corpus especializados se pueden establecer distintos tipos, como los **genéricos** condicionados por el género de los textos que contienen (dialectal, periodístico, científico, poético, etc.), como el corpus *Documentación de Lamento en Español desde Orígenes* (DOLEO) (Bravo-Garcia et al., 2013), o los **canónicos** que están formados por todos los textos que configuran la obra completa de un autor, como el *Index Thomisticus* (Alarcón, 2002).

En cuanto a la finalidad de un corpus, se dice que este es un corpus ***ad hoc*** cuando ha sido diseñado solamente para un proyecto específico, como el COSER o el *Corpus Léxico de Inventarios* (CorLexIn) (Rodríguez, 2014), y **universal** cuando ha sido diseñado pensando en la posibilidad de que sea usado para diversas finalidades sin concretar (como el CdE o el CODEA).

3. Época y temporalidad

Según la época que abarca, un corpus puede ser **contemporáneo** si recopila textos actuales, como el CREA, o **histórico** si recoge textos de una o diversas épocas del pasado, como el *Korpus Barokowy* (KorBa) (Gruszczyński et al., 2021).

A su vez, estos dos tipos de corpus pueden ser **sincrónicos** si en su estructura no presentan períodos, o **diacrónicos** si organizan los textos en franjas temporales, como el CODEA, el *Corpus Diacrónico y Diatópico del Español de América* (CORDIAM) (Academia Mexicana de la Lengua, 2015) o el CORDE.

4. Tamaño y evolución

En cuanto a la magnitud, un corpus puede decirse **grande** cuando no se plantea límite en el volumen de textos o palabras que ha de recopilar o, en caso de planteárselo, el número es muy elevado y no se preocupa de las cuestiones de equilibrio y representatividad o de codificación y anotación. En la actualidad, este tipo de corpus suele superar los 100 millones de palabras. Un ejemplo es el CORDE.

En cambio, los corpus **restringidos** establecen un límite de textos con la finalidad de que sean manejables, puedan alcanzar los objetivos de representatividad y equilibrio y sea posible desarrollar procesos de post-edición, como el caso del *Corpus Histórico Judeoespañol* (CORHIJE) (García Moreno & Pueyo Mena, 2013). Aunque no hay un número establecido de textos para diferenciar estos tipos de corpus, se suelen considerar restringidos los corpus con menos de cien millones de palabras. Queriendo afinar más esta clasificación, se pueden considerar dos tipos dentro de los corpus restringidos: **pequeños** cuando cuentan con menos de 20 millones de palabras, y **medianos** cuando tienen entre 20 y 100 millones de palabras.

Según la previsión de que el corpus tenga un final cerrado en cuanto a la inclusión de nuevos documentos, esto es, quede abierto a nuevas incorporaciones, se pueden clasificar en **corpus abierto** cuando no se estipula ningún límite de palabras y se mantiene en constante crecimiento, como el CdE, **corpus cerrado** cuando está compuesto por un número de textos establecido en la fase de definición, como el *Corpus Informatitzat del Català Antic* (CICA) (Torruella Casañas et al., 2009), y **corpus monitor** cuando pretende tener un volumen constante pero en continua renovación, como el CREA o el *Bank of English* (BOE) (University of Birmingham, 1991).

5. Número y tipo de ediciones

A la hora de componer un corpus se suele elegir un testimonio de cada obra, periódico o unidad de agrupación de texto, normalmente el más completo y representativo, aunque a veces para establecer comparaciones se dispone de diversas versiones de una misma obra. Cuando el corpus solo integra un testimonio de cada obra (aunque a veces incorpore distintos tipos de edición —paleográfica, crítica, normalizada, etc.—) se dice que es un corpus **monoedición**, mientras que cuando integra dos o más testimonios de una misma obra, como

transcripciones diferentes de un mismo texto o versiones libres de un mismo hecho histórico, se dice que es un corpus **pluriedición**.

Dentro de este último tipo, existen los corpus **comparables**, como los *Corpus Comparables y Paralelos de Discursos Europeos* (ECPC) (Calzada Pérez, 2006), donde los textos recopilados tienen características parecidas, y los **paralelos**, como el *Digital Corpus of the European Parliament* (DCEP) (Hajlaoui et al., 2014), donde se recopilan distintas ediciones de una misma obra, pero en este caso los diferentes textos tienen que compartir la misma estructura, aunque puede haber partes no coincidentes. Los corpus paralelos se utilizan con frecuencia para textos que incluyen su versión original y su traducción a otro idioma (o varios). Así, los corpus pluriedición también se pueden clasificar según sean monolingües o plurilingües.

Por último, los corpus paralelos se dicen **alineados** cuando permiten la comparación entre versiones más cómodamente al disponer los textos uno al lado del otro. Es el caso del corpus *Biblia Medieval* (Enrique Arias, 2008), que está formado por distintas versiones de la biblia en español, junto con una versión en latín y otra en hebreo.

En cuanto al tipo de edición, las facilidades que ofrece hoy en día la informática propician la confección de corpus **multiedición** (como en el corpus y proyecto CHARTA[5] (Red CHARTA, 2011a, 2011b), donde se ofrece una triple presentación —paleográfica, crítica y facsímil—), que presentan el mismo texto en diversas modalidades de edición: **facsímil** (reproducción fotográfica del original), **paleográfica** (transcripción sin correcciones ni interpretaciones lo más parecida a la original), **normalizada** (transcripción siguiendo la normativa ortográfica, léxica y sintáctica vigente), **crítica** (transcripción que pretende reconstruir el texto original) e **interpretativa** (transcripción que sigue los postulados de la edición paleográfica pero permite corregir ciertos errores para poder explicar el sentido del texto).

Para poder satisfacer las expectativas de todos los usuarios que se aproximan al texto, lo ideal es ofrecer siempre varios niveles de acceso: una transcripción paleográfica, pensada para llevar a cabo estudios que operan al nivel gráfico y fonético; una edición normalizada del texto, que facilite la lectura al usuario no iniciado o favorezca la base para acometer estudios de índole gramatical; y una

[5] https://www.redcharta.es

imagen del facsímil, que permita cotejar el original así como corregir lecturas dudosas o erróneas (Vaamonde, 2015).

6. Número de lenguas

Inicialmente los corpus estaban compuestos por textos en una sola lengua (**monolingües**, como el CODCAR) aunque a veces incluyen también distintos dialectos, pero más tarde se vio la utilidad de recopilar textos más o menos afines en distintas lenguas (**plurilingües**, como en la Biblia Medieval), por ejemplo, para la tarea de aprendizaje automático aplicado a la traducción automática. Normalmente, este último tipo es además comparable o paralelo.

7. Muestras

Los corpus **textuales** recogen el texto completo de cada obra (especificando lo que se considera como "completo"). En cambio, un corpus **de referencia** no recoge obras completas sino fragmentos de ellas. Es el caso del *British National Corpus* (BNC) (BNC Consortium, 2007), el BOE, el *Lancaster-Oslo-Bergen Corpus* (LOB) (Leech et al., 1978) o el Brown Corpus, aunque también suele usarse el término para corpus como el CREA y el CORDE, que están diseñados para proporcionar información exhaustiva acerca de una lengua en un momento determinado. En este caso no interesa tanto el texto sino la lengua que representa, y resultan especialmente importantes los aspectos de equilibrio y representatividad en la selección de los fragmentos. Por ejemplo, el Brown Corpus recoge muestras de aproximadamente 2.000 palabras, y el BNC de 40.000.

Los corpus **léxicos** recogen de cada documento fragmentos de textos pequeños y de longitud constante (usualmente 2.000 palabras), y la finalidad suele ser exclusivamente el estudio lexicográfico (confección de glosarios, diccionarios, etc.).

8. Marcaje

En función de las posibilidades de consulta que se quieran facilitar a los usuarios, los corpus pueden ser **simples** o **etiquetados**. En los corpus etiquetados o marcados se añaden al texto plano una serie de marcas declarativas que describen los elementos formales (cursiva, tamaño de la fuente), estructurales (capítulos, páginas), lingüísticos (entidades, cambios de registro) o semánticos (clase de las entidades y otros aspectos) del texto con etiquetas externas, lo cual facilita el tratamiento informático de los textos.

Los corpus etiquetados se pueden clasificar en un entorno filológico como **codificados** cuando las etiquetas marcan aspectos extralingüísticos referidos a la

estructura interna del texto, como el CODCAR, y **anotados** cuando las etiquetas marcan aspectos lingüísticos. En este último apartado se pueden distinguir los corpus **anotados morfológicamente** donde cada palabra lleva asociada su categoría morfológica (*part of speech*, como el CdE), **lematizados** donde se atribuye a cada palabra su lema (como en el CORPES XXI), lo cual facilita muchas tareas de búsqueda, **parentizados** cuando presentan un primer nivel de análisis sintáctico, **analizados** cuando presentan un análisis sintáctico más profundo y completo, y otros.

Un corpus es **simple** cuando no lleva etiquetas que declaren aspectos formales, estructurales o lingüísticos, y las consultas que pueden realizarse son básicamente de carácter léxico y sin localización de las palabras en la obra. Tan solo los corpus formados por materiales extraídos de Internet se podrían considerar de este tipo. Podría pensarse que este tipo de corpus no tienen ningún interés hoy en día, ya que la presencia de uno o varios tipos de codificación y anotación constituye un aspecto muy importante en las posibilidades de explotación de los corpus (Rojo, 2010). Sin embargo, son necesarios para el pre-entrenamiento de los modelos de aprendizaje actuales que requieren inmensas cantidades de datos. Recientemente, se ha publicado un corpus masivo en español a partir de datos de Internet, esCorpius (Gutiérrez-Fandiño, Pérez-Fernández, et al., 2022), el más grande hasta la fecha en nuestro idioma.

9. Documentación y accesibilidad

Por último, la documentación del corpus hace referencia normalmente al hecho de que se tenga registro de la procedencia de los textos y sea posible utilizar esta información en las búsquedas, ya sea para realizar búsquedas específicas o para conocer la procedencia de los textos.

Actualmente hay diferentes modalidades de acceso, desde previo pago, con licencia para tareas específicas y declaradas, o sin restricciones, independientemente de que los corpus sean públicos o privados.

2.1.2 Equilibrio y representatividad en la composición de corpus

Sobre todo, aunque no únicamente, para que el corpus constituya un modelo adecuado del universo lingüístico que se quiere describir es necesario que se cumplan dos propiedades: la representatividad, y el equilibrio. En realidad, el

primer término engloba el segundo, puesto que para que un corpus sea representativo, tiene que ser equilibrado. A continuación, se presentan estas dos propiedades en el ámbito lingüístico, aunque en algunos casos habría que redefinirlos según el objetivo del corpus construido.

1. Representatividad

Este concepto expresa la necesidad de que la composición del corpus incluya toda la gama de variabilidad de una población (Biber, 1993), es decir, que cubra todas las variedades lingüísticas o todos los fenómenos del tipo que pretenda representar. La representatividad también se define en base al nivel de generalización de los resultados que se pueden obtener, y se dice que es alto cuando los descubrimientos o resultados obtenidos a partir de dicho corpus pueden generalizarse a otros datos que pertenecen al mismo dominio (Leech, 1991). Así, la representatividad determina los tipos de preguntas de investigación que pueden abordarse y el grado de generalización de los resultados que cabe esperar.

Al ser la investigación a partir de corpus un estudio de carácter inductivo (ya que pretende extraer, desde determinadas observaciones, el principio general que en ellas está implícito), los documentos seleccionados deben ser *muestras* representativas del total de una *población*. Por ello es importante prestar especial atención a la cantidad de textos que se recogerán, y la variedad y porcentaje de las muestras. De hecho, la representatividad de un corpus viene determinada por el grado de especificidad del propio corpus.

Es importante reflexionar sobre si un corpus representa la lengua o el uso que la sociedad, sus hablantes y escritores hacen de ella. En términos teóricos es imposible asumir que pueda conseguirse un corpus perfectamente balanceado y representativo: al fin y al cabo, los corpus son ejemplos de un particular dominio del lenguaje o del núcleo general del lenguaje cotidiano. Por otro lado, el habla es una forma de comunicación natural que la escritura no puede llegar a ser, ya que es más formal y suele estar más controlada. La lengua, aunque solo se manifieste en textos, no es la suma de los textos sino algo distinto (Kabatek, 2013). La total representatividad es imposible salvo en los corpus canónicos, que incluyen todas las obras de un autor. Pero, en general, no es posible representar un idioma con un corpus y el principio de representatividad se presupone. Sin embargo, hoy en día contamos con corpus muy grandes y por tanto es más fácil hacer asunciones de representatividad.

2. Equilibrio

El equilibrio o la distribución del corpus establece cómo se organizan los textos que lo van a componer, es decir, la proporción en la que deben aparecer. Es decir, se define en base a la proporción que existe entre la cantidad de categorías o etiquetas asociadas al corpus y la cantidad de datos pertenecientes a cada una de ellas. Cuando se pretende que las cantidades de palabras o de textos de cada apartado (es decir, las muestras) estén en proporción respecto a su distribución en el total de la población, se dice que el corpus es **proporcional** o **equilibrado**. Todos los corpus que tienen como objetivo el estudio de la lengua actual de manera general, usan este tipo de distribución, como es el caso del CORPES XXI, ya que es la manera más adecuada de describir la lengua a partir de cada uno de sus parámetros (épocas, dialectos, tipos textuales, etc.). Cuando, en cambio, se pretende que las cantidades de palabras de cada apartado sean iguales o parecidas, se llama corpus **equivalente**. Este tipo de corpus suele ser más idóneo para los corpus históricos, puesto que difícilmente se puede saber el porcentaje de representatividad que un apartado (temporal, dialectal, tipológico, etc.) tenía respecto al total. Por ello es preferible no prejuzgar ningún valor de proporcionalidad, como en el caso del CICA.

Teniendo en cuenta estas reflexiones, la cuestión de la representatividad y equilibrio de los corpus se puede estructurar de la siguiente manera: por un lado, la representatividad está condicionada por la calidad y diversificación del material que compone el corpus (representatividad cualitativa) y, por otro, está condicionada por la cantidad de obras o palabras que lo componen y sus porcentajes de distribución en su estructura (representatividad cuantitativa). Más concretamente, la representatividad cuantitativa está condicionada por el equilibrio externo, que atiende a la cantidad de obras o palabras que componen el corpus con respecto al total de la población, y el equilibrio interno, respecto a la relación del número de muestras entre apartados, y que puede ser proporcional o equivalente. En esquema, los factores que intervienen en la representatividad se pueden resumir como muestra la Figura 1.

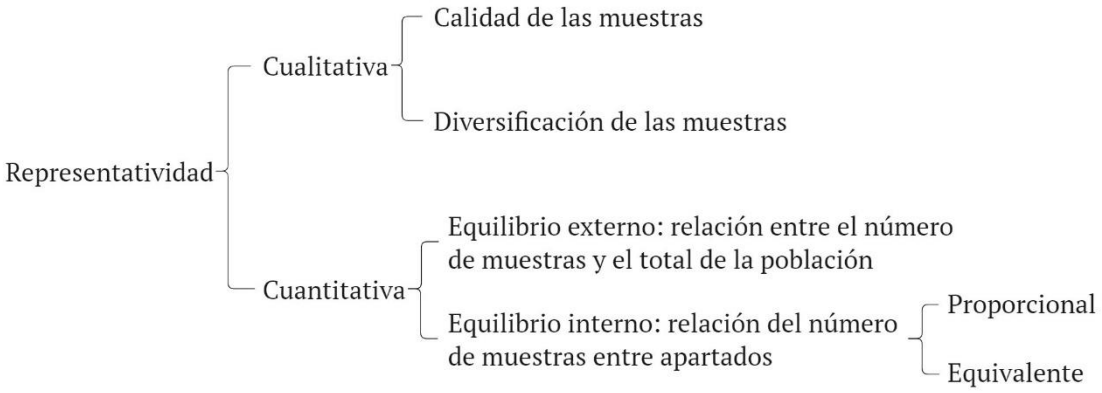


Figura 1. Esquema de los factores que intervienen en el nivel de representatividad. Adaptado a partir de (Torruella Casañas, 2017).

El objetivo principal de los corpus, el equilibrio y la representatividad, se revela mucho más difícil de conseguir en corpus históricos que en corpus de lengua moderna (Gruszczyński et al., 2021). Esto se debe a que, por un lado, a menudo es difícil saber el volumen y la proporción que tenían ciertos apartados con respecto al total de la población (por ejemplo, debido a la subrepresentación de ciertos tipos de textos como los escritos por mujeres, o al desconocimiento de la autoría de los textos), y, por otro, en ocasiones, aunque se sepa o se intuya cuál era el total y la parte proporcional de apartado, no se han conservado suficientes documentos para completar el número de palabras necesario del apartado o del período (limitada conservación de la producción literaria, catástrofes en las que se pierden textos, etc.). De hecho, los textos mejor conservados tienden a ser los literarios porque suelen ser relanzados, a diferencia de otros de carácter mas utilitario, como los periódicos. Esto obstaculiza otro de los objetivos de los corpus, es decir, saber qué tipos de textos se leen más en una sociedad. En cualquier caso, al usar este criterio de proporcionalidad, es importante que incluso los apartados menos representados cuenten con un número mínimo de palabras suficiente que garantice que es representativo de la totalidad de la lengua del apartado.

Para concluir, la Tabla 1 resume los parámetros clasificatorios de los corpus.

Parámetro	**Tipos**
Modalidad	Oral Escrito Mixto
Temática	General Especializado Genérico Canónico
Finalidad	Ad hoc Universal
Época	Contemporáneo Histórico
Temporalidad	Sincrónico Diacrónico
Tamaño	Grande Restringido
Evolución	Abierto Cerrado Monitor
Número de ediciones	Monoedición Pluriedición Comparable Paralelo Alineado
Tipo de edición	Reproducción fotográfica (facsímil) Edición paleográfica Texto normalizado Texto crítico
Número de lenguas	Monolingüe Plurilingüe
Muestras	Textual Referencia Léxico
Marcaje	Simple Etiquetado Codificado Anotado Morfológicamente Lematizado Parentizado Analizado (*full parsing*)
Documentación	Documentado No documentado
Accesibilidad	Público Privado
Equilibrio	Proporcional Equivalente
Representatividad	Si/No

Tabla 1. Resumen de los parámetros clasificatorios de los corpus.

2.1.3 Algunos corpus existentes

Los corpus son un recurso fundamental en el desarrollo de aplicaciones basadas en PLN, ya que permiten evaluar los sistemas desarrollados, proporcionan un marco común en el que comparar técnicas alternativas y permiten entrenar los sistemas de aprendizaje automático supervisados o no, ya que ajustan sus parámetros a partir de una parte del corpus usada para su entrenamiento. Además, proporcionan una muestra amplia y natural del lenguaje y ayudan a ofrecer una mirada útil del uso y el significado de las palabras.

Desde el ámbito lingüístico, los corpus permiten generar listas de frecuencias de palabras, encontrar la frecuencia de palabras, frases y n-gramas (cadenas con una cantidad n de palabras), investigar construcciones sintácticas, o realizar otras tareas y búsquedas que aborden información semántica. Naturalmente, estos tipos de búsquedas se realizan mejor en corpus grandes y bien diseñados (Davies & Parodi, 2022).

La construcción de corpus también propicia la celebración de foros de evaluación competitiva de tareas no resueltas y relacionadas con la gestión de la información (en inglés *shared tasks*), que incentivan la mejora y comparación de los modelos o aplicaciones existentes en nuevas tareas y dominios. El funcionamiento básico de estos foros es que los organizadores de una tarea concreta facilitan a todos los investigadores participantes un mismo corpus, una definición única de tarea a resolver y disponen de un conjunto de resultados esperados, que permiten comparar los resultados de los diferentes sistemas participantes y obtener conclusiones.

Hoy en día contamos con una enorme cantidad de corpus diseñados para diversas finalidades. En el Apéndice I (Corpus, hemerotecas y colecciones) se recoge una selección y descripción de corpus históricos (por ser los de interés en este trabajo), principalmente en español tanto de España como de otros países entre los siglos XV y XIX, aunque no sólo, ya que se incluyen algunos recursos en inglés u otros idiomas y otros períodos, por su interés para el estudio que se presenta.

Como ya se ha indicado, la Real Academia Española ha impulsado la creación de distintos recursos para el español, desde la decisión en 1995 de confeccionar un banco de datos del español. Sin embargo, muchos recursos se desarrollan en el ámbito universitario, como el corpus CODEA, el CODCAR, el DOLEO o el CorLexIn. Otros se desarrollan o bien en instituciones dedicadas al estudio de la lengua, como el CORDIAM o el *Corpus de Textos Antiguos de Galicia*

COTAGAL (Pichel Gotérrez et al., 2016), o en proyectos conjuntos entre universidades e instituciones, como el PRESEEA (Universidad de Alcalá, 2014) o el proyecto Post Scriptum (Universidade de Lisboa, 2014).

Por otro lado, existen recursos avanzados que funcionan como directorio y recolector de recursos digitales relacionados con el ámbito histórico o actual. A nivel europeo existe la biblioteca digital Europeana[6], que facilita encontrar el patrimonio cultural de archivos, museos, bibliotecas y colecciones audiovisuales. Por su parte, HISPANA[7] proporciona acceso al patrimonio cultural y científico español y agrega el contenido nacional a Europeana. Para los corpus de carácter histórico relacionados con las lenguas de la Península Ibérica, existe el Portal de corpus históricos hiberorrománicos (CORHIBER) (Torruella & Kabatek, 2021). La biblioteca digital Memoriademadrid[8] digitaliza y difunde el patrimonio histórico, cultural y documental del Ayuntamiento de Madrid; esta web recoge diversas colecciones de fotografías, libros, publicaciones, pinturas, planos, mapas y otros tipos de documentos.

Por último, cabe destacar iniciativas como CHARTA, que no sólo constituye un corpus que pretende ofrecer una amplia representación del documento archivístico en español de los siglos XII al XIX, con documentos de Europa, América y Asia, sino que se concibe como un proyecto global para la edición y análisis lingüístico de textos cuyos criterios se han conformado como un estándar para la edición de corpus.

Para concluir esta sección, en la Tabla 2 se presenta una caracterización de varios corpus relevantes para este estudio según los parámetros clasificatorios.

[6] https://www.europeana.eu/es

[7] https://hispana.mcu.es/

[8] http://www.memoriademadrid.es

	CdE	CHARTA	CODEA	CORDE	CORPES XXI	COSER	CREA
Modalidad y muestras	mixto, referencia	escrito, textual	escrito, textuales	escrito, textual	mixto, referencia	oral, textual	mixto, referencia o textual
Temática y finalidad	varios géneros (histórico, dialectos...), universal	general con distintos géneros, universal	especializado, universal	varios géneros (líricos, dramáticos, históricos), universal	general con multitud de géneros, universal	genérico (dialectal), ad hoc	general (libros, prensa y otros), universaal
Época y temporalidad	ambos, ambos	histórico, diacrónico	histórico, diacrónico	histórico, diacrónico	contemporáneo, ambos	contemporáneo, diacrónico	contemporáneo, diacrónico
Tamaño y evolución	grande, abierto	restringido, abierto	pequeño, cerrado	grande, cerrado	grande, semiabierto	restringido, semiabierto	grande, monitor
Número y tipo de ediciones	textual y oral cuando procede	multiedición (paleográfica, crítica y facsímil)	multiedición (paleográfica, crítica y facsímil)	textual	textual y oral cuando procede	archivo sonoro y transcripción	textual y oral cuando procede
Número de lenguas	variantes de 21 países hispanohablantes	variantes habladas en la Península	variedades del español entre el siglo XI y el XIX	español de España y América	español de España y América	variantes habladas en la Península	español de España y América
Marcaje	parcialmente lematizado, analizado sintáctica y morfológicamente	sin etiquetar	lematizado	codificado	codificado y lematizado, algunas alineaciones del sonido con la transcripción	lematizado y analizado morfológicamente	lematizado y analizado morfológicamente

Tabla 2. Clasificación de varios corpus.

2.2 Construcción de corpus

En esta sección se describen las fases en la construcción de un corpus, y se presentan algunas de las tecnologías y herramientas más utilizadas en cada paso.

El primer paso es la definición de los límites y la estructura del corpus según su finalidad, determinando la cantidad de muestras que lo constituirán y las dimensiones de estas. Los documentos seleccionados tienen que constituir una muestra representativa del total de la población, ya que son la base que permitirá pasar de los estadios hipotéticos a los estadios empíricos de la investigación.

Es aconsejable que las finalidades del corpus sean múltiples para que pueda ser consultado para distintos fines tanto de carácter lingüístico como histórico-documental. En los trabajos de investigación el tiempo es limitado y por tanto también lo es el tiempo de creación del corpus, por lo que, en general, es mejor diseñar un corpus más pequeño, pero bien estructurado y anotado, que un corpus grande poco estructurado y sin codificar.

Siguiendo a (Torruella Casañas, 2017) y (Nakayama, 2021), la creación de un corpus puede dividirse en diferentes fases, siendo la primera la selección y adquisición de documentos digitalizados, que a su vez puede subdividirse en las etapas de preparación (definir los límites, la estructura y la finalidad del corpus), localización, adquisición y descarga en ficheros.

A continuación, serán necesarias las fases de pre-procesamiento, normalización, marcado o etiquetado (estructural, lingüístico y/o semántico) y un segundo almacenamiento. Después de estas tres fases suele haber una fase importante de revisión, para la que hay que tener muy claros los objetivos del tratamiento posterior del corpus. Finalmente, vendrá la fase de tratamiento del corpus en el que se podrá desde lematizar el corpus hasta categorizar los documentos o extraer diferente información en forma de lexicones, diccionarios u otros.

2.2.1 Magnitud de los corpus

Existe una estrecha relación entre el tamaño del corpus y el rango de fenómenos que se pueden estudiar cuando el estudio es de ámbito lingüístico (Davies & Parodi, 2022). Aunque en la actualidad con la tecnología de aprendizaje profundo (*Deep Learning*) los límites sobre el tamaño de los corpus son mucho mas amplios, en teoría, cuanto más grande sea un corpus, mayor será la probabilidad de encontrar una determinada secuencia o la de afirmar que nunca existió. Lo ideal

desde este punto de vista sería que el corpus contuviese todos los discursos existentes, lo cual no es posible, y menos aun cuando se trabaja con corpus históricos. Por ello los corpus pretenden alcanzar el grado óptimo de representatividad, recogiendo un número de palabras suficientemente elevado.

Para usos generales o para la obtención de datos de carácter léxico, y con las nuevas tecnologías se suelen usar corpus grandes (al menos 100 millones de palabras). Para otros tipos de estudios, en cambio (por ejemplo, lingüísticos de carácter fonético-fonológico o morfosintáctico) es más importante la calidad que la cantidad, por lo que son más útiles los corpus de volumen más reducido que cuiden las ediciones que recogen y sigan criterios de representatividad cualitativa y de equilibrio interno entre apartados.

De acuerdo con (Torruella Casañas, 2017):

1. Los corpus grandes suelen ser poco exigentes en cuanto a la selección de los documentos y la calidad filológica de las ediciones usadas, y pobres en cuanto a la codificación y anotación de los textos. Un corpus pequeño con uno o dos millones de palabras es lo más aconsejable para estudiar fenómenos comunes, como marcadores del discurso, preposiciones o construcciones sintácticas frecuentes.
2. En cuanto al tamaño de las muestras que componen el corpus, si el estudio es de tipo léxico, puede bastar con unas 1.000 palabras, aunque incluso para estudiar fenómenos sintácticos esta puede ser una cantidad estadísticamente significativa. No obstante, en el proyecto se deberá decidir si se necesita recopilar documentos completos (corpus textual) o fragmentos de los documentos (corpus de referencia). Para dar cuenta del comportamiento general de la lengua pueden ser suficientes las muestras de entre 2.000 y 5.000 palabras, como han probado varios estudios (Biber, 1990).
3. El tamaño de los corpus es importante para que se puedan ofrecer datos científicamente válidos considerando algunos principios del método científico a la hora de diseñarlos y utilizarlos como fuente de investigaciones, y así evitar que los resultados resulten irrelevantes, poco fiables o incluso llevar a falsas predicciones.

Para aplicar el método científico en los trabajos de investigación, en primer lugar, hay que apartar las nociones previas que se puedan tener respecto del fenómeno a analizar, a menos que no hayan sido justificadas científicamente. En segundo lugar, se deben analizar los fenómenos desde el ángulo más objetivo posible,

buscando los elementos de clasificación que faciliten un tratamiento de los datos que permita observar lo que entra en la normalidad y lo que presenta algún tipo de especificidad.

En la investigación con corpus se pueden aplicar dos tipos de análisis complementarios: el cualitativo y el cuantitativo. Mientras que el cualitativo pretende hacer descripciones detalladas y completas del fenómeno que se quiere estudiar, el cuantitativo trabaja con las frecuencias y otros parámetros matemáticos de los fenómenos observados en el corpus con el fin de crear modelos estadísticos que los expliquen. Y esta es la gran aportación de los corpus en el avance de este tipo de investigaciones: la posibilidad de realizar un estudio empírico que ofrezca información que pueda ser estadísticamente significativa y resultados que puedan considerarse generalizables. De hecho, la creación de corpus digitales en línea y la disponibilidad de grandes cantidades de datos, ha hecho casi imprescindible el uso de programas para trabajar con todos estos datos (Nieuwenhuijsen, 2016).

En este sentido cabe aclarar, por un lado, que no es lo mismo dar ejemplos de un fenómeno, que demostrarlo. Por ello es importante trabajar con rigor estadístico, y en este proceso metodológico el primer paso es cuantificar los datos del corpus. Sin embargo, la estadística no debe considerarse como la finalidad de la investigación, sino como un instrumento para describir y resumir los datos, y con ello hacer estimaciones de significación y fiabilidad.

2.2.2 Propiedades de los corpus

Las directrices de diseño que exigen las propiedades de representatividad y equilibrio, cruciales en el diseño de corpus, a menudo implican la obtención de corpus cuyos ejes, aunque se distribuyen de manera proporcional, no lo hacen de manera equivalente (debido a las características inherentes a la lengua y sus usos). Esto repercute en el sesgo de los modelos de predicción basados en aprendizaje automático y profundo, ya que aprenden extrayendo patrones de los datos bajo la asunción de una distribución equivalente de los datos (Haibo He & Garcia, 2009). Así, se dice que se tienen datos desbalanceados cuando las diferentes clases del conjunto de entrenamiento contienen un número dispar de ejemplos.

Por ejemplo, supongamos que se recogen las opiniones de los usuarios de un comercio de venta online con el fin de diseñar un clasificador que sea capaz de determinar si los comentarios son positivos o negativos. Si, pongamos, la muestra

contiene un 90% de comentarios positivos, el corpus está desbalanceado y los algoritmos aprenderán a clasificar mejor la categoría más representada respecto a la que es menos frecuente. Para identificar este desequilibrio, basta mirar la distribución de las clases, por ejemplo, en un histograma. En este escenario, antes de poder aplicar técnicas de aprendizaje automático a corpus desbalanceados, hay que reequilibrarlos. En primer lugar, conviene medir cómo afecta al rendimiento del modelo este desajuste, y evaluar los pros y contras de las posibles soluciones.

En la mayoría de los casos se busca optimizar la exactitud (*accuracy*), pero cuando un modelo se entrena con datos desbalanceados es posible obtener una exactitud muy alta de manera trivial, sin que el modelo consiga el objetivo deseado. Por ejemplo, si tan solo el 5% de los pacientes que visitan la consulta del dentista están libres de caries, el modelo podrá predecir que todos los pacientes que llegan padecerán de caries con una exactitud del 95%, pero no obtendremos un modelo que sepa distinguir realmente entre pacientes con y sin caries. Por eso se recomienda trabajar con varias medidas y no sólo con la exactitud, siendo las más comunes la precisión, el *recall* y el valor o medida F (normalmente la $F1$[9], que es la media harmónica entre la precisión y el *recall*).

Además, hay que distinguir entre las medidas *micro* y *macro*: la micro agrega las contribuciones de todas las clases para calcular la medida media, mientras que la macro calcula la medida de manera independiente para cada clase y después computa la media. La micro da la misma importancia a cada clase, lo cual favorece a las clases grandes. En cambio, la macro da la misma importancia a cada clase, y refleja cómo funciona el modelo (queremos que el modelo funcione bien en todas las clases, incluyendo las minorías). Por eso, al trabajar con datos desbalanceados se recomienda dar preferencia a las medidas macro, aunque resulta conveniente analizar los resultados de ambas medidas.

Para abordar este desequilibrio se utilizan distintas estrategias, a nivel de los datos o a nivel de algoritmo. La mayoría de ellas se pueden implementar a través de librerías, como *imbalanced-learn*[10] (basada en *scikit-learn*).

A nivel de los datos, se trata de modificar la distribución de los datos de entrenamiento para reducir el nivel de desequilibrio. De esta manera los gradientes se pueden actualizar y se pueden "ver" un número similar de ejemplos de cada clase. Algunas de las técnicas más comunes son las siguientes:

[9] https://es.wikipedia.org/wiki/Valor-F
[10] https://imbalanced-learn.org/stable/

- Ampliación de los datos (*over-sampling*). Esta técnica consiste en aumentar "artificialmente" el volumen de datos de aquellas categorías que se encuentran menos representadas o que son menos frecuentes. Normalmente, el corpus disponible se divide en dos conjuntos, uno de entrenamiento, del que los modelos de aprendizaje extraen patrones, y otro de evaluación sobre el que se comprueba el acierto de dichos modelos aprendidos). Básicamente, se duplican ejemplos aleatorios de la case subrepresentada, consiguiéndose un conjunto de entrenamiento (o todo un corpus, en el caso de técnicas no supervisadas) donde todas las categorías cuentan con una cantidad de datos similar, lo que facilita que todas ellas sean "igual de visibles" para el modelo. Esta técnica conlleva el riesgo de sobreajuste (*overfitting*), es decir, el modelo puede no ser capaz de generalizar bien y clasificar correctamente datos nuevos.
- Reducir la clase mayoritaria (*under-sampling*). Consiste en eliminar aleatoriamente muestras de la clase mayoritaria para reducirla e intentar equilibrar los datos. Tiene el peligro de que se prescinda de muestras importantes que aportan información valiosa al modelo y por lo tanto puede empeorarlo. Para resolverlo habría que seguir algún criterio para seleccionar qué muestras eliminar.
- Métodos híbridos. Estos o bien combinan *oversampling* y *undersampling*, o bien aplican *undersampling* después del *oversampling* a modo de limpieza de los datos.
- Muestras artificiales. Consiste en crear muestras sintéticas (no idénticas) utilizando diversos algoritmos que intentan seguir la tendencia del grupo minoritario. Según el método, pueden mejorar los resultados, pero lo peligroso de crear muestras sintéticas es que se altera la distribución "natural" de esa clase y puede confundir al modelo en su clasificación.

A nivel de algoritmo, en cambio, lo que se ajusta es el proceso de aprendizaje del modelo. Mediante el ajuste de parámetros se intenta equilibrar la clase minoritaria penalizando a la clase mayoritaria durante el entrenamiento, aumentando así la importancia de las clases más pequeñas. Una manera común de hacerlo es asignando a las clases pesos inversamente proporcionales a sus frecuencias en los datos de entrenamiento. Así, se da más peso a los ejemplos subrepresentados en el cómputo total de la pérdida (*loss*). Por ejemplo, al trabajar con redes neuronales se puede ajustar la métrica *loss* para que penalice a las clases mayoritarias, y en un modelo de regresión logística podríamos utilizar el parámetro

class_weight="balanced", aunque no todos los algoritmos tienen estas posibilidades.

En resumen, en la práctica es raro trabajar con corpus que tengan una proporción de clases equivalente, pero existen distintas soluciones que permiten la aplicación de modelos de aprendizaje automático a cualquier tipo de corpus.

2.2.3 Transparencia

Los datos juegan un papel muy importante en el aprendizaje automático. Sin embargo, todavía no existen procesos estandarizados para la documentación de la creación de los conjuntos de datos y en particular de los corpus. La documentación acerca de la creación y el uso de los conjuntos de datos ha recibido poca atención, a pesar de que sí la haya recibido en concreto la proveniencia de estos.

Desde propuestas como la de (Gebru et al., 2021) se ha puesto énfasis en el hecho de que cada conjunto de datos debe ir acompañado de una ficha técnica (que ellos llaman *datasheet*) que documente la motivación, composición, proceso de colección, usos recomendados, etc., para mejorar la transparencia y la gestión responsable de los datos en la comunidad del aprendizaje automático, así como la reproducibilidad de los resultados y la reducción de sesgos indeseados en los modelos de aprendizaje.

De los creadores se pretende una reflexión minuciosa sobre el proceso de creación, distribución y mantenimiento de los corpus o *datasets*, incluyendo las asunciones subyacentes, potenciales riesgos o perjuicios y las implicaciones que conlleva su uso. De cara a los consumidores de estos recursos se procura asegurar que cuenten con toda la información que necesitan para poder utilizarlos como se debería. Por ello se requiere transparencia por parte de los creadores.

El proceso de creación de una ficha técnica no pretende ser automático, ya que va contra el objetivo de animar a los creadores a reflejar adecuadamente el proceso de creación, distribución y mantenimiento del corpus. En cambio, se propone responder a una lista exhaustiva de preguntas que se agrupan en secciones de acuerdo con las fases de creación y vida del corpus o *dataset*: motivación, composición, proceso de colección, pre-procesamiento, (limpieza, etiquetado), uso, distribución y mantenimiento. Algunos ejemplos son:

- ¿Cuál es la finalidad del *dataset* o del corpus?
- ¿Quién lo ha creado y quién lo ha financiado?

- ¿Qué representan las muestras que lo componen?
- ¿Existen errores o redundancias?

2.2.4 Almacenamiento de corpus

Los corpus se almacenan en distintos formatos estructurados para poder consultarlos y recuperar la información a través de herramientas de búsqueda. Así, distinguimos por un lado los lenguajes de marcado junto con los estándares que siguen, y por otro las herramientas para la consulta según el tipo de almacenamiento de los datos.

Un lenguaje de marcado es una forma de codificar un documento que incorpora etiquetas al texto que contienen información adicional acerca de su estructura o presentación. A continuación, se presentan brevemente uno de los lenguajes fundamentales para el trabajo con corpus, XML, y su predecesor, SGML, junto al estándar TEI que los normaliza, y finalmente se introduce el lenguaje JSON.

El lenguaje de marcado generalizado estándar[11] (SGML por sus siglas en inglés) es un estándar para definir lenguajes de marcado generalizados para documentos, y es una especificación ISO ("ISO 8879: 1986 Tratamiento de la información - Sistemas de texto y de oficina - Lenguaje de marcado generalizado estándar (SGML)"). En la Web, HTML 4, XHTML, y XML son lenguajes populares basados en SGML.

XML es un lenguaje de marcado extensible (eXtensible Markup Languaje). Se trata de un lenguaje abierto derivado de SGML que sigue el estándar del *World Wide Web Consortium* (W3C), está optimizado para su uso en la WWW y permite describir el sentido o la semántica de los datos. Es un formato de datos inherentemente jerárquico, cuya forma más natural de representarlo es mediante un árbol, como veremos a continuación.

A partir de la información en texto plano, esta se estructura y almacena en este formato para poder visualizarla más cómodamente y poder realizar consultas. Todo documento XML comienza con un encabezado que declara que se trata de un documento XML, de la forma:

```
<?xml version=”1.0” encoding=”UTF-8”?>
```

[11] https://es.wikipedia.org/wiki/SGML

El documento se estructura mediante etiquetas anidadas, siempre dispuestas en pares para indicar el comienzo y el final de un elemento, como en el Ejemplo *1*:

```
<biblioteca>
      <libro>
            <autor>Isabel Allende</autor>
            <titulo>La casa de los espíritus</titulo>
            <precio moneda="euros">10</precio>
      </libro>
      <libro>
            <autor>Ana de Miguel</autor>
            <titulo>Ética para Celia</titulo>
            <precio moneda="euros">17</precio>
      </libro>
</biblioteca>
```

Ejemplo 1. Estructura XML.

Las etiquetas (o nodos del árbol) pueden contener texto anidado entre una etiqueta inicial (`<autor>`) y una etiqueta final (`</autor>`) y también pueden contener atributos para almacenar información en metadatos, como en `<precio moneda="euros">` donde se especifica el tipo de moneda en que se muestra el precio.

Una DTD (*Document Type Definition*)[12] es una descripción de la estructura y sintaxis de un documento XML o SGML. Es importante para permitir el procesamiento robusto de los documentos, ya que en base a ella se pueden definir nociones de corrección en documentos XML (válido o bien construido). El documento se relaciona con su DTD en la declaración del tipo de documento en el prólogo.

El *Text Encoding Initiative* (TEI)[13] es un consorcio cuyo objetivo es proponer un esquema de anotación (DTD) para representar los rasgos de un texto que necesiten ser hechos explícitos (estructurales, lingüísticos…). Las ventajas que proporciona el estándar son facilitar el procesamiento de los textos por diferentes aplicaciones y diversos propósitos, y facilitar su intercambio.

En la construcción del corpus *Oralia diacrónica del español* (ODE) (Calderón Campos & Vaamonde, 2020), los textos son transcritos en XML según las directrices TEI, que a pesar de ser un modelo de edición ampliamente generalizado en el mundo de las Humanidades Digitales, todavía es muy poco frecuente en el de la compilación de corpus históricos (Calderón Campos, 2019).

[12] https://es.wikipedia.org/wiki/Definición_de_tipo_de_documento
[13] https://tei-c.org

JSON (*JavaScript Object Notation*) es un formato de intercambio de datos independiente del lenguaje. Partió de *JavaScript*, pero hoy en día numerosos lenguajes de programación incluyen la posibilidad de generar y analizar datos en formato JSON, entre ellos Python. Su uso no excluye el de XML, ya que se pueden complementar en la misma aplicación.

JSON es una alternativa ligera a XML, ya que su procesamiento es más ágil y su formato más sencillo, aunque XML permite describir estructuras más complejas de manera más legible. JSON está limitado a almacenar datos textuales y numéricos, mientras que XML puede almacenar cualquier tipo de datos.

El Ejemplo 1 quedaría como sigue en formato JSON en el Ejemplo 2:

```
{
    "biblioteca": {
        "libro": [
            {
                "autor": "Isabel Allende",
                "titulo": "La casa de los espíritus",
                "precio": {
                    "-moneda": "euros",
                    "#text": "10"
                }
             },
             {
                "autor": "Ana de Miguel",
                "titulo": "Ética para Celia",
                "precio": {
                    "-moneda": "euros",
                    "#text": "17"
                }
            }
        ]
    }
}
```

Ejemplo 2. Estructura del formato JSON.

2.2.5 Anotación de corpus

A pesar de todos los avances en las áreas de la informática de aprendizaje automático e inteligencia artificial en los últimos años, aún no se ha superado la paradoja de la era de la información, y es que, para que los humanos podamos fiarnos de las máquinas, las máquinas necesitan que les enseñemos primero. Así, cuando se aplican técnicas de aprendizaje supervisado (modelos de predicción

entrenados con datos etiquetados) en PLN, la anotación de los datos juega un papel crucial.

La anotación consiste en realizar marcas o anotaciones sobre los textos que describan, analicen o relacionen aspectos concretos. Estas marcas se hacen a nivel de palabra, sintagma, oración, fragmento o documento completo. Esta información adicional se puede usar para para el estudio de la lengua, o para entrenar modelos de aprendizaje o evaluar su rendimiento. La anotación manual de corpus es un proceso costoso, que se lleva a cabo por expertos en el área (lingüistas, historiadores, etc). Un corpus sin buenas anotaciones puede resultar prácticamente inútil, por lo que este paso es de especial importancia.

A la hora de definir el conjunto de etiquetas, es preferible que estas sean directas y descriptivas, y no hacer uso de acrónimos que dificulten aplicar y verificar las etiquetas de manera rápida.

La elaboración de corpus fidedignos es imprescindible para el entrenamiento y la evaluación de los modelos que usan anotaciones. Cuando estos corpus anotados se elaboran manualmente, se suele hacer referencia a ellos como *gold standard*. Estos son los más fiables, aunque su construcción es muy laboriosa en términos de tiempo y esfuerzos (Wissler et al., 2014). En cambio, cuando el corpus anotado se genera automáticamente, por ejemplo, a partir de recursos de la Web, se denomina *silver standard*.

Según el tipo de estudio que se espere realizar, se optará por el tipo de anotación más conveniente. Para los estudios de la lengua, se suelen anotar, como mínimo, las categorías gramaticales y los lemas. Pero también se realizan otras anotaciones de tipo sintáctico, léxico-semántico, fonético y fonológico, de la modalidad del discurso, de la pragmática o de las correlaciones, entre otros. Para el entrenamiento de modelos de aprendizaje automático, se anotan distintas características según la tarea: las entidades nombradas, las palabras complejas, la intención del discurso, etc.

Para fijar los criterios de anotación se suele elaborar una guía detallada con las instrucciones precisas sobre lo que anotar y lo que no anotar. Un ejemplo de este proceso es la guía de anotación del corpus *BioScope* (Szarvas et al., 2008) de textos biomédicos, que contiene anotaciones de negación (*not*, *never*) y especulación (*can*, *possibly*) a nivel de palabras clave y de frase para su estudio lingüístico. El proceso de anotación se llevó a cabo por dos anotadores lingüistas independientes y un anotador principal, responsable de elaborar la guía de anotación y de resolver los casos en que los anotadores no estaban de acuerdo.

Para comprobar la calidad de las anotaciones, el etiquetado es un proceso que suele realizarse por varios anotadores para poder contrastar sus anotaciones, ya que, en general, al anotar existe un grado de subjetividad. Para medir el grado de acuerdo entre los anotadores se utilizan medidas de acuerdo (IAA: *Inter-Annotator Agreement*, también conocido como *Inter-Rater Reliability* o *Inter-Rater Agreement*). Si todos los anotadores realizan las mismas anotaciones de forma independiente, significa que las directrices (la guía de anotación) son claras y que las anotaciones son probablemente correctas. Cuanto mayor sea el acuerdo, mayor será la calidad. Algunas de las medidas más conocidas son el coeficiente kappa de Cohen[14], el Pi de Scott[15], el *kappa de Fleiss*[16], o mediante una matriz de confusión[17].

Cuando no se alcanza el grado mínimo de acuerdo entre anotadores, que generalmente se establece en torno al 70%, puede deberse a distintos motivos. Puede ser señal de que las etiquetas deben ser más claras en cuanto a su significado, o que quizás sea necesario crear nuevas o eliminar algunas. También puede deberse a que las directrices de la guía de anotación no son lo suficientemente descriptivas, o bien que sea necesario formar adecuadamente a los anotadores para que comprendan bien la tarea y la guía.

Distintas organizaciones han puesto sus esfuerzos en la estandarización de la anotación de los corpus, como el *Linguistic Data Consortium*[18] (Bird & Liberman, 1998) o el consorcio *British National Corpus*[19].

Se han propuesto diferentes metodologías para la anotación de corpus. Por ejemplo, la metodología MATTER (Aldama et al., 2022) que siguen en el Instituto de Ingeniería del Conocimiento (IIC) está pensada para la creación de corpus destinados a entrenar algoritmos de PLN. La metodología se divide en cinco pasos: modelo de anotación, anotación, entrenamiento y evaluación de un modelo de aprendizaje, evaluación de los resultados y revisión del modelo de anotación. Además, estas fases se pueden complementar con tres más: idea o hipótesis (previo a concretar el modelo de anotación), provisión de herramientas

[14] https://es.wikipedia.org/wiki/Coeficiente_kappa_de_Cohen
[15] https://en.wikipedia.org/wiki/Scott%27s_Pi
[16] https://en.wikipedia.org/wiki/Fleiss%27_kappa
[17] https://es.wikipedia.org/wiki/Matriz_de_confusión
[18] https://www.ldc.upenn.edu
[19] http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=consortium

y formatos de anotación (previo a la anotación en sí), y distribución (una vez finalizados todos los pasos).

2.3 Herramientas para la gestión de corpus

Hoy en día contamos con diferentes herramientas tanto para la construcción como para el posterior análisis de los corpus, que varían de complejidad según el tipo de usuario destinado a usarlas o la tarea en la que se va a utilizar. Desde aplicaciones web o de escritorio con interfaces gráficas intuitivas diseñadas para los menos familiarizados con la informática, hasta programas más complejos aptos sólo para su uso desde la línea de comandos. Estas herramientas se pueden clasificar en distintos tipos según su finalidad.

2.3.1 Transcripción de documentos

En los últimos veinte años se han realizado multitud de procesos de digitalización para la conservación de colecciones culturales tanto a nivel local como nacional y europeo. Estos proyectos han generado millones de imágenes o ficheros de texto (PDF) que necesitan ser tratados para la transcripción del texto que contienen, ya sea de forma manual o mediante la aplicación de procesos de reconocimiento óptico de caracteres, conocido como OCR (del inglés *Optical Character Recognition*) (Menta et al., 2022). Este proceso de transcripción genera un material de gran riqueza para las investigaciones, ya que permite consultar el texto, analizarlo y manipularlo de manera mucho más eficiente. Hoy en día, este proceso constituye una fase importante en la construcción de corpus históricos.

A continuación, se presenta la herramienta Transkribus (READ Coop), que permite el reconocimiento (transcripción) del texto que compone los documentos tanto manualmente como a través de un OCR. De hecho, cada vez más proyectos hacen uso de esta herramienta (Aranda García, 2022; Ayuso García, 2022; Bazzaco et al., 2022).

Transkribus[20] es una plataforma para la digitalización de documentos, el reconocimiento y transcripción de textos mediante técnicas que provienen del área de la informática de inteligencia artificial, y la búsqueda en documentos históricos.

[20] https://readcoop.eu/transkribus/

Para la transcripción cuenta con una serie de modelos públicos[21] entrenados en distintos idiomas y grafías, lo que facilita encontrar uno que se aproxime al estilo de los documentos que se busca transcribir. De no ser así, la herramienta brinda la posibilidad de entrenar un nuevo modelo para automatizar la transcripción de los documentos del corpus. El flujo de trabajo con la herramienta en general sigue el proceso que se muestra en la Figura 2:

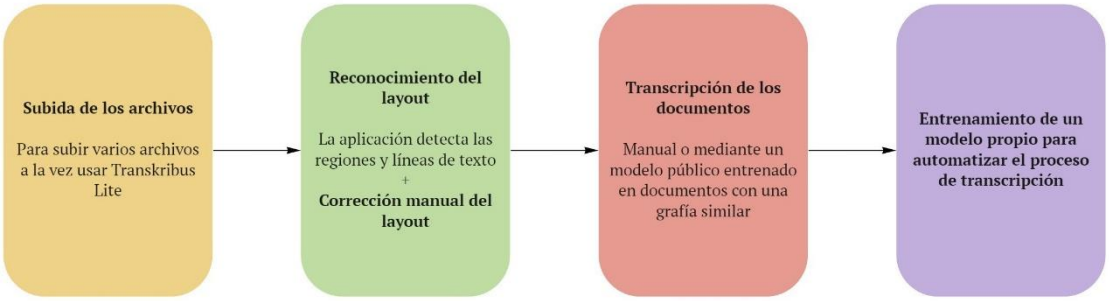


Figura 2. Diagrama del flujo de trabajo en Transkribus

Todas las funcionalidades (reconocimiento de la estructura de los documentos o layout, transcripción manual, entrenamiento de modelos, etc.) son gratuitas en la herramienta, excepto la transcripción de textos con modelos existentes. Para hacer uso de esta funcionalidad se obtienen o compran créditos (con el registro proporcionan 500, lo que equivale a unas 500 páginas).

En la Figura 3 se muestra un ejemplo de trabajo con la plataforma. A la izquierda se ve la lista de los documentos de la colección, a la derecha se visualiza el documento con unas marcas que indican los lugares en los que se encuentran las regiones de texto, y en la parte inferior derecha el texto transcrito del documento.

2.3.2 Acceso y búsquedas en corpus y bases de datos

Existen distintas categorías de buscadores según el tipo de datos sobre los que operan. Para buscar en colecciones de texto sin estructurar o en la Web, se usan indexadores, esto es, motores de búsqueda que almacenan un índice (organización de la información, por ejemplo, mediante el método TF-IDF[22]) para optimizar y acelerar el rendimiento al buscar los documentos relevantes para una consulta.

[21] https://readcoop.eu/transkribus/public-models/
[22] https://en.wikipedia.org/wiki/Tf–idf

Figura 3. Estructura de Transkribus.

Sin estos índices, el motor de búsqueda tendría que escanear cada documento del corpus, lo cual requeriría una cantidad de tiempo y potencia computacional considerable. En cambio, las búsquedas en bases de datos estructuradas, como XML, permiten realizar consultas lógicas y búsquedas por campos. Ejemplos de esto son los lenguajes como SQL (*Structured Query Language*) para bases de datos relacionales, o SPARQL para consultar datos RDF[23]. Estas aproximaciones a menudo se usan de manera conjunta. Por ejemplo, en (Seo et al., 2003) se trabaja sobre documentos XML extendiendo la indexación para crear relaciones y poder realizar consultas avanzadas.

Algunos ejemplos de las tecnologías o herramientas usadas habitualmente para la consulta de corpus son las siguientes.

- **XPath** (*XML Path Language*). Usa expresiones con sintaxis de tipo ruta para identificar y navegar los nodos en un documento XML, e integra más de 200 funciones para ello. Es un elemento principal del estándar XSLT y una recomendación de la W3C. Por ejemplo, en el Ejemplo 1 podríamos seleccionar todos los nombres de los libros de la biblioteca mediante la sintaxis `/biblioteca/libro/titulo`.

  Por su parte, **XQuery** es el lenguaje de consulta para buscar y extraer elementos y atributos en documentos XML, similar a SQL para las bases de datos. Está construido a partir de expresiones XPath, y es una recomendación de la W3C. De hecho, está diseñado para que toda expresión válida en XPath sea también una consulta válida en XQuery.

[23] RDF un estándar de modelado para el intercambio de datos en la Web, y permite mezclar, mostrar y compartir información estructurada y semiestructurada entre diferentes aplicaciones Web (https://es.wikipedia.org/wiki/Resource_Description_Framework)

En el ejemplo, para seleccionar todos los títulos con este lenguaje, la sintaxis sería `doc("biblioteca.xml")/biblioteca/libro/titulo`.

- **Solr**[24] es un proyecto de código libre escrito en Java, basado en Lucene[25], la librería Java de software libre de Apache para la búsqueda e indexación de documentos. Solr es un motor de búsqueda que permite almacenar (indexar) documentos vía JSON, XML, CSV o binarios a través de HTTP de manera muy rápida y eficiente, y que ofrece multitud de funcionalidades como el facetado de los resultados de búsqueda.
  Solr está basada en la tecnología NoSQL. Las Bases de datos SQL son bases de datos estructuradas y de tipo relacional, mientras que las bases de datos NoSQL son de tipo no-relacional y no están estructuradas. Las bases de datos no relacionales, o NoSQL, son más flexibles y no necesitan saber de antemano qué información se va a almacenar y cómo va a ser almacenada. Muchos proyectos usan Solr o Lucene como motor de búsqueda, por ejemplo, el sistema MTAS (Multi Tier Annotation Search) (Brouwer et al., 2016) se propone como solución a la escasa escalabilidad de las búsquedas en textos anotados. Por ello añaden estructura y capas de anotación al enfoque de Lucene y lo implementan como plugin de Solr.
- **SPARQL** (acrónimo recursivo del inglés *SPARQL Protocol and RDF Query Language*) es un lenguaje estandarizado para la consulta de grafos RDF, normalizado por el RDF Data Access Working Group (DAWG) del World Wide Web Consortium (W3C). Es una tecnología clave en el desarrollo de la web semántica que se constituyó como recomendación oficial del W3C el 15 de enero de 2008, siendo actualizado a la versión 1.1 en 2013. En un principio SPARQL únicamente incorpora funciones para la recuperación sentencias RDF. Sin embargo, algunas propuestas también incluyen operaciones para el mantenimiento (creación, modificación y borrado) de datos.

2.3.3 Análisis textual y visualización

En este apartado se presentan algunas de las aplicaciones más conocidas que permiten analizar y explorar textos mediante distintas funcionalidades en entornos visuales.

[24] https://solr.apache.org/
[25] http://lucene.apache.org/

Figura 4. Ejemplo de uso y apariencia de Voyant.

- **Voyant**[26] es un entorno web de lectura y análisis de textos digitales. Con una interfaz muy intuitiva, el mejor modo de conocer la herramienta es usándola. Permite introducir textos a partir de URLs, o abriendo un corpus existente con ficheros en alguno de los (muchos) formatos aceptados: MS Word, PDF, MS Excel, XML, HTML o archivos de texto sencillo, entre otros. La Figura 4 muestra la apariencia de Voyant.

  Al introducir los documentos se realiza un escaneo de estos proporcionando información sobre la frecuencia y tendencia de las palabras que los componen. A nivel visual se muestra una nube de palabras que las representa con una magnitud proporcional a su frecuencia en el texto, y un gráfico que muestra las frecuencias de las palabras en cada segmento del texto. Cambiando estas visualizaciones, podemos ver de manera cuantitativa la frecuencia de cada palabra en la pestaña *Términos*, lo que nos permite hacernos una idea de los conceptos más importantes. Después se muestra un resumen del corpus con datos como el promedio de palabras por oración, y por último algo muy interesante en la lingüística de corpus, un cuadro de búsqueda que permite buscar palabras e identificar los contextos en los que aparecen. Si, por ejemplo, introducimos la página web de Wikipedia sobre María Zambrano, observamos que algunas de las palabras más frecuentes son *razón*, *Madrid*, *realidad* o *pensamiento*. Otra de ellas es *premio*, que podemos poner en contexto para identificar cuándo y por qué aparece. La herramienta es de código libre disponible en GitHub[27] bajo una Licencia Pública General de GNU.

[26] https://voyant-tools.org/
[27] https://github.com/sgsinclair/Voyant

- **AntConc**[28] es una herramienta para el análisis de corpus. Es una aplicación *freeware*, utilizable de forma gratuita, pero no es *open source* y no podemos modificar el código para adaptarlo. En su página web se pueden consultar la licencia y los términos de uso[29]. La interfaz hace muy ameno trabajar con corpus, aunque las funciones que ofrece son limitadas. Solo trabaja con textos planos (.txt), aunque en la misma página de la herramienta se proporciona un software para convertir PDF y DOCX en archivos de texto. Ofrece versiones para los principales sistemas operativos (Windows, Mac y Linux) por lo que la plataforma no es un impedimento en este caso. Tampoco requiere cambios (actualizaciones o versiones específicas de java) ni instalación de ningún software adicional.

  Ofrece siete funciones (se puede consultar el manual de la aplicación[30]):

  - **Concordance Tool** o función de concordancias. Muestra los resultados de búsqueda en formato KWIC (palabras clave en contexto), permitiendo observar cómo se usan las palabras y frases frecuentemente en un corpus de textos, ya que se especifica el texto del que proviene cada frase.

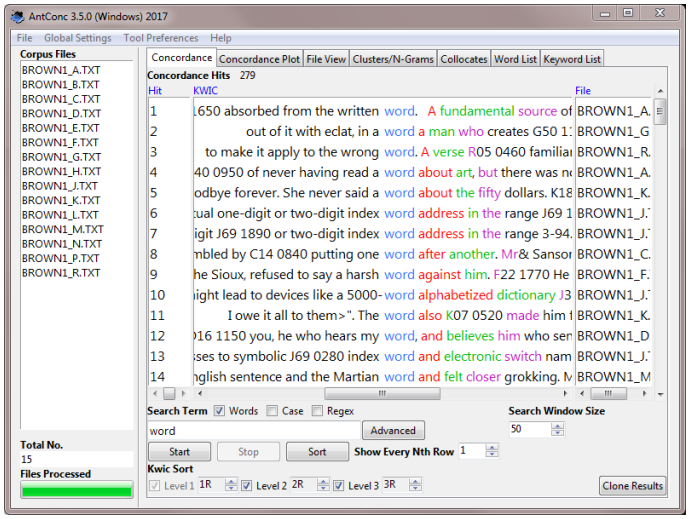


Figura 5. Función de concordancias en Antconc.

  - **Concordance Plot Tool** o Diagrama de concordancias. Muestra los resultados de búsqueda en un código de barras. Su utilidad es la de ver la posición en la que los resultados de búsqueda aparecen en los textos de destino.

[28] https://www.laurenceanthony.net

[29] https://www.laurenceanthony.net/software/antconc/releases/AntConc359/license.pdf

[30] https://www.laurenceanthony.net/software/antconc/releases/AntConc359/help.pdf

- **File View Tool**. Muestra de forma individual cada texto permitiendo examinar con más detalle los resultados generados en otras funciones.
- **Clusters/*N*-grams**. Esta función muestra grupos de palabras (clusters) según los criterios de búsqueda. Básicamente resume los resultados generados en la función *Concordance* (o equivalentemente en *Concordance Plot*). Por otro lado, la función *n-grams* analiza el corpus en su totalidad para clusters de longitud n (por ejemplo, 1 palabra, 2 palabras...) permitiendo encontrar expresiones comunes en un corpus.

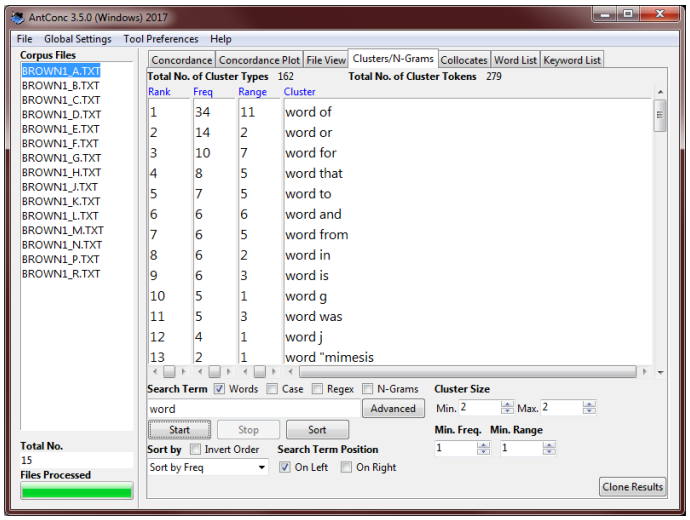


Figura 6. *Clusters* en Antconc.

- **Collocates** o Colocaciones. Se utiliza para generar una lista ordenada de las palabras que aparecen junto al término buscado (antes o después), permitiendo encontrar patrones en el lenguaje.
- **Word List** o Lista de palabras. Cuenta todas las palabras del corpus y las presenta en una lista ordenada junto con su frecuencia, permitiendo encontrar rápidamente las palabras más frecuentes en un corpus.
- **KeyWord List** o Lista de palabras clave. Muestra las palabras que son inusualmente frecuentes (o infrecuentes) en el corpus en comparación con las palabras en un corpus de referencia. Asimismo, permite identificar las palabras características del corpus, por ejemplo, como parte de una temática o estudio de la lengua con fines específicos.

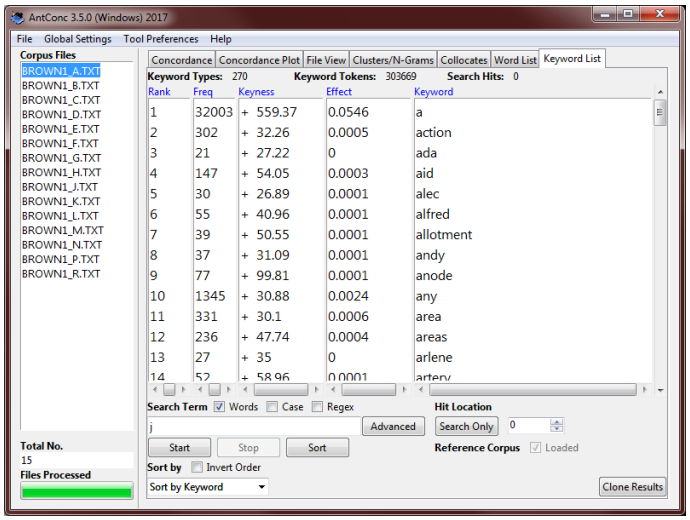


Figura 7. Lista de palabras clave en Antconc.

- **LYNEAL** (Letras y Números en Análisis Lingüísticos)[31] es un desarrollo informático que permite buscar e identificar formas, efectuar análisis cuantitativos, descriptivos y multivariantes y obtener resultados en cifras (absolutas, relativas y normalizadas), en gráficos y en mapas. Diseñado por Hiroto Ueda y Antonio Moreno Sandoval, se trata de una reunión de varios proyectos de corpus digitales, entre los que se encuentran el CODEA, el *Atlas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid* (ALDICAM) (Borja, 2018) y el PRESEEA, entre otros.

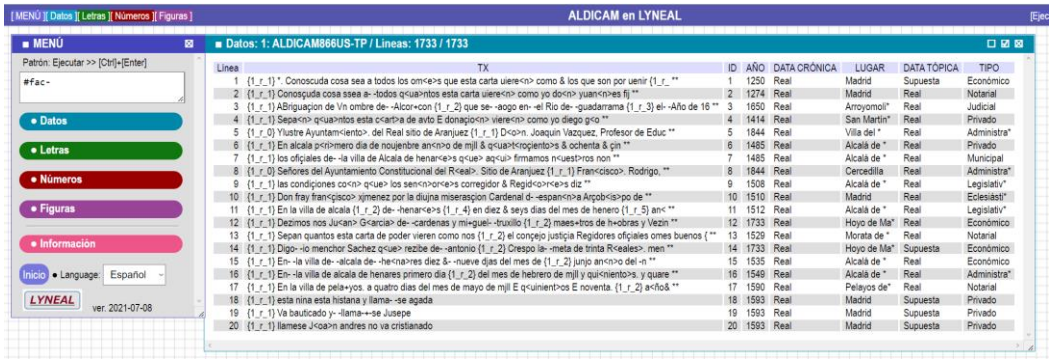


Figura 8. Apariencia de LYNEAL.

2.3.4 Anotación y análisis de corpus

Las herramientas de anotación de corpus suelen incorporar otras funcionalidades relacionadas con el análisis del corpus. Algunas de las mas relevantes son las siguientes.

[31] http://shimoda.lllf.uam.es/ueda/lyneal/

- **Brat**[32] (*Browser-Based Rapid Annotation Tool*) es una herramienta web para la anotación estructurada y colaborativa de textos. Permite asignar etiquetas a partir de un conjunto predefinido y establecer relaciones entre ellas, así como la colaboración en tiempo real entre usuarios y la comparación de las anotaciones.

  La interfaz proporciona un modo de visualización muy completo con una edición intuitiva, e incluye recursos externos para la normalización de varias características con datos externos como Wikipedia y Freebase. Incluye un conjunto de funciones de búsqueda por anotaciones y otros configurables como la búsqueda de palabras clave en contexto. Se puede usar para muchas tareas de PLN como la resolución de correferencias o el etiquetado sintáctico[33].

  No es necesario instalar ningún software local o plugin en los buscadores ya que está construido exclusivamente con tecnologías web. Sin embargo, está pensado para sistemas operativos UNIX y por tanto para usarlo en Windows es necesario usar una máquina virtual y familiarizarse con algunos comandos de dichos sistemas operativos, y preferiblemente también con el manejo de servidores web Apache, aunque no es imprescindible. Solo admite ficheros de texto, y las anotaciones se guardan en formato .ann separadas del fichero de texto las contiene.

- **Doccano**[34] es una herramienta web de código abierto para la anotación. Sus funcionalidades incluyen la clasificación de textos, etiquetado de secuencias, y tareas de conversión secuencia a secuencia, lo que permite crear datos etiquetados para análisis de sentimientos, reconocimiento de entidades, resumen de textos, etc. Tan solo admite ficheros de texto. Cuenta con una interfaz gráfica moderna desde la que se realizan todas las configuraciones. A diferencia de brat, no ofrece la posibilidad de etiquetar relaciones y clasificaciones anidadas.

  Permite la anotación entre varios usuarios, pudiendo escribir y guardar la guía de anotación en la misma aplicación, aunque no incluye otras características para la anotación colaborativa. Soporta varios idiomas y tiene soporte para móviles.

[32] http://brat.nlplab.org
[33] https://brat.nlplab.org/examples.html
[34] https://doccano.herokuapp.com

Para ejecutar doccano hay tres opciones: con pip (Python 3.8+), Docker o Docker Compose. En la documentación[35] proporcionan detalles para cada tipo de instalación. Están disponibles varias demos online en la pestaña superior (*try demo*) de su página web[34].

- **Tagtog**[36] es una herramienta web de anotación. Permite etiquetar tanto a nivel de palabra como de documento y establecer relaciones entre las etiquetas. Al tratarse de una aplicación web no es necesario instalar ningún software, aunque también es posible instalarla en un entorno privado, para lo cual es necesario consultar los requisitos[37] del sistema.

  Permite la colaboración entre usuarios, pudiendo establecer guías de anotación y roles. Además, incluye una herramienta para calcular el acuerdo entre anotadores (*inter-annotator agreement*) y evaluar así la calidad de las anotaciones.

  Admite varios formatos de entrada de textos, como texto plano, .csv, .xml, .html y otros[38].

  La herramienta también ofrece funcionalidades para integrarla con herramientas como SpaCy para automatizar el etiquetado, y entrenar modelos de inteligencia artificial propios a través de Webhooks.

  Para funcionalidades más avanzadas existen diversos planes de suscripción, como la anotación automática para documentos PDF.

- **TEITOK**[39] (Janssen, 2016) es una aplicación web para ver, crear y editar corpus que contienen tanto anotaciones lingüísticas como marcado de texto enriquecido. La interfaz gráfica permite a los usuarios consultar en diversos modos de visualización los documentos anotados, y la edición en tiempo real del documento XML subyacente a los administradores.

  El sistema tiene un diseño modular donde cada módulo incluye interfaces útiles para trabajar con una amplia variedad de corpus. Por ejemplo, en corpus históricos permite alinear las transcripciones de los textos manuscritos con la imagen facsímil, o trabajar con distintas ediciones ortográficas de un texto para combinarlas en un único archivo XML, entre otras opciones. De la misma forma es posible trabajar con audios, y realizar

[35] https://doccano.github.io/doccano/
[36] https://www.tagtog.net/
[37] https://docs.tagtog.net/on_premises_README.html
[38] https://docs.tagtog.net/ioformats.html
[39] http://www.teitok.org

múltiples tareas como visualizar árboles de dependencia y geolocalizar documentos, o ayudar en la creación de nuevos corpus en formato XML de manera más sencilla.

La herramienta es gratuita y el código fuente se mantiene en GitLab[40], y algunas herramientas de conversión también en GitHub[41]. Es necesario descargar la aplicación en un sistema con Linux o MacOS (no está disponible en Windows) y crear una cuenta de usuario. En su página web es posible consultar algunos de estos ejemplos de uso, donde también proporcionan una guía rápida para la instalación en Linux en el apartado *Downloads*.

Algunas de las funcionalidades de TEITOK están disponibles en su web en la sección de herramientas[42]: la creación de árboles sintácticos, la identificación automática del idioma de un texto y la representación gráfica de consultas CQL (Cassandra Query Language).

Numerosos proyectos hacen uso de TEITOK para la construcción de corpus, entre ellos los que se listan en la sección *Projects* de su página web[43]. Un ejemplo es el corpus Oralia diacrónica del español (ODE) (Calderón Campos, 2019), donde se transcriben los documentos en formato XML siguiendo las directrices del consorcio TEI.

- **CATMA**[44] o *Computer Aided Textual Markup and Analysis* (Sullivan, 2013) es una aplicación web para el análisis de textos. Permite etiquetar, analizar, interpretar y visualizar los textos en el modo que mejor se adapte al tipo de búsqueda deseado: cuantitativo o cualitativo, exploratoria, o descriptiva y taxonómica. La interfaz de usuario es muy clara e intuitiva, y trabaja con cualquier tipo de archivo de texto común o directamente introduciendo una URL. La herramienta está orientada a proyectos, por lo que es necesario registrarse para usarla. De esta manera es posible trabajar con un corpus propio y compartirlo fácilmente con otros usuarios para trabajar de manera colaborativa.

  Las funciones que ofrece son muy versátiles. El etiquetado es una parte fundamental de la investigación en humanidades, y por ello los

[40] https://gitlab.com/maartenes/TEITOK
[41] https://github.com/ufal
[42] http://www.teitok.org/index.php?action=tools
[43] http://www.teitok.org/index.php?action=projects
[44] https://catma.de

desarrolladores han puesto énfasis en aportar originalidad en este proceso. La idea es poder anotar los textos como lo haríamos en un libro: permite desarrollar conjuntos de etiquetas propios o crear categorías sobre la marcha, lo cual facilita el desarrollo de interpretaciones de los textos personalizadas y trabajar con modelos teóricos predefinidos. Así, si un párrafo es susceptible de tener más de una interpretación, no hay ningún problema en asignar múltiples etiquetas o incluso que se contradigan entre sí. La flexibilidad que ofrece no compromete la organización de las etiquetas, ya que estas se exportan en formato TEI/XML, lo cual permite reutilizarlas en otros contextos y exportar los datos a otras herramientas, como por ejemplo Voyant, y aprovechar las posibilidades de visualización que esta ofrece.

Para el análisis es posible elegir entre diversas visualizaciones interactivas, desde nubes de palabras hasta grafos de distribución y árboles dobles, además de poder ejecutar consultas (predefinidas o creándolas) para analizar, explorar o evaluar los textos. A continuación, se muestra un ejemplo de la función de visualización *doubletree*, que proporciona una manera más visual de explorar el contexto de palabras clave.

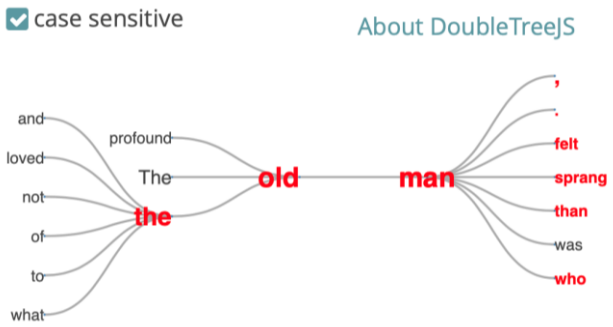


Figura 9. Ejemplo de la función doubletree en CATMA.

En su web cuentan con un manual de uso[45] y tutoriales[46] donde es posible explorar el funcionamiento de la herramienta.

2.3.5 Explotación automática de corpus

En los últimos años se ha popularizado el uso del lenguaje de programación Python para realizar tareas de PLN sofisticadas.

[45] https://catma.de/how-to/compact-manual/
[46] https://catma.de/how-to/tutorials/analyze-and-visualize/

Python[47] es un lenguaje de programación de alto nivel de propósito general, muy usado en ciencia de datos y en la producción de algoritmos de aprendizaje automático y profundo. Tiene muchas características que lo convierten en el lenguaje de programación idóneo para los proyectos de PLN: la simplicidad y consistencia de la sintaxis, la transparencia de la semántica, la flexibilidad, el acceso a librerías y frameworks de IA y aprendizaje automático, e incluso la amplia comunidad que la sostiene. Tan solo son necesarias algunas líneas de código para implementar técnicas de PLN en Python.

Una librería es una colección de módulos relacionados que contienen paquetes de código destinados a usarse repetidamente en diferentes programas. Hace que la programación en Python sea más sencilla y cómoda para el programador, ya que son amplias y proporcionan múltiples funcionalidades.

A continuación, se presentan algunas de las librerías y entornos más populares en el ámbito del PLN, cada una de ellas dedicada a objetivos distintos y, en general, variados.

- **Flair**[48] (Akbik et al., 2019) es un proyecto basado en PyTorch[49] 1.5+ y Python 3.6+. La librería está pensada para trabajar de manera sencilla con modelos de PLN del estado del arte en numerosas tareas: reconocimiento de entidades, etiquetado gramatical, desambiguación y clasificación del sentido de las palabras, entre otros, y soporte en cada vez más idiomas. También incluye distintas implementaciones de *embeddings* a nivel de palabra y de documento, como BERT y ELMo, incluyendo su propia implementación (Akbik et al., 2018). Al estar construida directamente en PyTorch, permite entrenar modelos y experimentar con nuevos enfoques gracias a los *embeddings* y las clases de Flair.
  En su repositorio están disponibles varios tutoriales para aprender a manejarse con la herramienta[50].
- **Gensim**[51] es una librería de código abierto para representar documentos como vectores semánticos. Está diseñada para procesar textos no estructurados (en texto plano) mediante algoritmos de aprendizaje automático no supervisado, es decir, sin ningún tipo de datos etiquetados. Los algoritmos que utiliza Gensim, como Word2vec (Mikolov et al., 2013),

[47] https://www.python.org
[48] https://github.com/flairNLP/flair
[49] https://pytorch.org
[50] https://github.com/flairNLP/flair#tutorials
[51] https://radimrehurek.com/gensim/

FastText (Bojanowski et al., 2017) o *Latent Dirichlet Allocation* (Blei et al., 2003), descubren automáticamente la estructura semántica de los documentos examinando los patrones de coocurrencia estadística dentro de un corpus de entrenamiento. Una vez que se encuentran estos patrones, cualquier documento en texto plano (frases, expresiones o palabras) se puede expresar brevemente en la nueva representación semántica y consultar la similitud con otros documentos.

- **SpaCy**[52] se considera una de las librerías en Python más rápidas para PLN avanzado. Es de código abierto e incluye implementaciones para realizar una gran cantidad de tareas de PLN, como el etiquetado gramatical, análisis de dependencias, reconocimiento y desambiguación de entidades, clasificación de textos y entrenamiento de modelos, incluso los modelos del estado del arte basados en la arquitectura *Transformer* (Vaswani et al., 2017). Tiene soporte para más de 66 idiomas, y 73 *pipelines* entrenados en 22 idiomas, que por defecto cargan el etiquetador gramatical, el analizador de dependencias y el reconocedor de entidades. En su web explican detalladamente cómo usar la herramienta en su guía de uso[53].

  Los creadores de SpaCy, la compañía Explosion[54], tienen su propia herramienta de anotación, Prodigy[55] (Montani & Honnibal, 2018). Está implementada en una librería de Python, es programable mediante *scripts* e incluye una aplicación web para la anotación. Está especialmente pensada para crear datos de entrenamiento y de evaluación para modelos de aprendizaje automático, y en concreto para flujos de trabajo en ciencia de datos. Se puede usar para inspeccionar y limpiar datos, hacer análisis de errores y desarrollar sistemas basados en reglas para usar en combinación con modelos estadísticos, además de etiquetado de entidades, clasificación de textos, etiquetado de imágenes, audios y vídeos, y también etiquetado libre para, por ejemplo, asociar a un texto de entrada su traducción en otro idioma. No es un software como servicio al que se accede a través de internet, sino que es una herramienta que se descarga, instala y ejecuta en un dispositivo en el que los datos permanecen sin viajar a

[52] https://spacy.io
[53] https://spacy.io/usage/spacy-101
[54] https://explosion.ai
[55] https://prodi.gy

otros servidores. La mejor manera de probarla es a través de su demo online[56].

- **Hugging Face**[57] es una startup en el campo del PLN, que ofrece una librería con los modelos más punteros en IA y que ha ido ganando protagonismo desde la aparición de los *Transformers*. Su objetivo es democratizar el PLN para que todo el mundo pueda usarlo, y poder así promover su avance.

  Los *Transformers* permitieron crear modelos de lenguaje que podían realizar tareas de lenguaje natural a niveles equivalentes a los del rendimiento humano. HuggingFace inicialmente comenzó como un proyecto de IA conversacional, y ahora se dedica a desarrollar herramientas de código abierto para el aprendizaje por transferencia en PLN. La librería se divide en tres módulos: Transformers, Datasets, y Tokenizers. La librería Transformers proporciona miles de modelos que permiten realizar numerosas tareas de PLN, como clasificación de textos, extracción de resúmenes, reconocimiento de entidades, traducción, respuesta de preguntas, etc. La ventaja de esta librería es que evita el trabajo inicial de configuración del entorno y la arquitectura, ya que es posible poner los modelos en funcionamiento en pocas líneas de código en Python. Además, son compatiblels con los dos frameworks de aprendizaje profundo más populares en PLN, TensorFlow y PyTorch. La tokenización consiste en asignar valores numéricos únicos a las palabras (o grupos de palabras) de un documento. La librería Tokenizers proporciona algoritmos para tokenizar documentos de manera muy rápida, y en consonancia con el modelo de lenguaje que se quiera utilizar. Por último, la librería Datasets proporciona una plataforma unificada que hace accesibles los *datasets* en diferentes idiomas y tareas para poder ajustar fácilmente los modelos a nuevas tareas.
- **Tensorflow** es una librería útil para construir aplicaciones de aprendizaje profundo. Se basa en grafos computacionales en los que un nodo representa datos persistentes u operaciones matemáticas y las aristas representan el flujo de datos entre los nodos, que es una matriz multidimensional o tensor; de ahí el nombre TensorFlow. Aunque fue diseñado para redes neuronales, funciona bien para otras redes en las que el cálculo puede modelarse como un gráfico de flujo de datos.

[56] https://prodi.gy/demo
[57] https://huggingface.co

Se pueden construir diferentes tipos de redes profundas con TensorFlow, como redes convolucionales, recurrentes, autocodificadores, etc. Sin embargo, no tiene soporte para la configuración de hiperparámetros. Para esta funcionalidad, se puede usar Keras.

- **Keras** también es una librería de aprendizaje profundo que se ejecuta sobre otras librerías de aprendizaje automático de código abierto, como TensorFlow, y que también es de código abierto. Para desarrollar modelos de aprendizaje profundo, Keras adopta una estructura minimalista en Python que hace que sea más fácil de aprender y rápido de escribir.
- **PyTorch** es una librería de código abierto utilizada en *machine learning*, muy relacionada con las dos anteriores. Es imperativa, lo que significa que se ejecuta de inmediato y el usuario puede verificar si está funcionando o no antes de escribir el código completo. Podemos escribir una parte del código y verificarlo en tiempo real, es una implementación integrada basada en Python para brindar compatibilidad como una plataforma de aprendizaje profundo.

  Tiene diversas funciones, aunque se usa principalmente en sistemas de PLN. Respecto a Tensorflow, tiene menos funcionalidades, pero es más fácil de aprender.

Capítulo 3. Reconocimiento de entidades en corpus de dominios específicos

La tarea de reconocimiento de entidades nombradas ha evolucionado de manera importante en los últimos años. Hoy en día destaca la adaptación de sistemas de reconocimiento de entidades a dominios específicos, como el de las Humanidades Digitales (Ehrmann, Romanello, Fluckiger, et al., 2020) o el dominio biomédico (Cho & Lee, 2019; Miranda-Escalada et al., 2022; Moreno Sandoval et al., 2018).

En este capítulo se introduce la tarea de reconocimiento de entidades y se proporciona un contexto desde sus inicios, haciendo un breve repaso del tipo de tecnologías que se han usado hasta las que hoy en día lideran el estado del arte. Asimismo, se describe el modo en que se evalúan estos sistemas con *datasets* creados a tal efecto, y se presentan las herramientas más utilizadas hoy en día. En la segunda parte se introducen las Humanidades Digitales y se presenta el problema del reconocimiento de entidades en textos históricos. En la tercera parte se describe un caso de estudio, referente al trabajo realizado en el proyecto CLARA-HD (PID2020-116001RB-C32) de la UNED. Finalmente, en la cuarta parte se motivan, en este contexto, los experimentos que se reportan en el capítulo 3.5 y se concretiza la propuesta que se llevará a cabo.

3.1 Descripción de la tarea y contexto actual

El reconocimiento de entidades nombradas es la tarea dentro del PLN que implica la localización en los documentos textuales de secuencias de palabras (en este caso, entidades, usualmente nombres propios) y su clasificación en las categorías predefinidas de nombres de personas, lugares, organizaciones u otros tipos de entidades. Es común referirse a esta tarea como *NER* por sus siglas en inglés (*Named Entity Recognition*). Constituye un objetivo clave de prácticamente cualquier tarea de minería de textos, ya que los nombres de personas, los lugares o las organizaciones subyacen en la semántica de los textos y guían su interpretación. Un buen sistema de NER permitirá comprender el tema de un texto y clasificar los documentos según su relevancia, y constituye una tarea clave

en los sistemas de PLN en aplicaciones de respuestaa a preguntas, la extracción y búsqueda de información o su extracción, entre otros (Yadav & Bethard, 2018).

La extracción de entidades se realiza segmentando el texto, siguiendo en general un proceso de dos pasos. En el primero, el sistema detecta la localización de los tokens que forman una entidad y sus límites, para lo que suele utilizarse el método de etiquetado IOB (que se describirá en el siguiente apartado). Después la entidad se clasifica mediante la asignación de una etiqueta que indica la categoría a la que pertenece. A pesar de que existen enfoques basados en reglas para el reconocimiento de entidades, hoy en día se suelen adoptar técnicas de aprendizaje automático y de aprendizaje profundo, ya que la ambigüedad de los datos impide a menudo establecer una regla infalible que determine de qué entidad se trata en ese contexto. Por ejemplo, la palabra "Petra" puede referirse al enclave arqueológico, al nombre propio o a la película, por lo que sería difícil desambiguarla sin el contexto en el que se realiza la mención o un amplio conocimiento del mundo real.

En la mayoría de los idiomas y dominios de conocimiento no existen grandes cantidades de datos etiquetados con entidades para el entrenamiento de sistemas supervisados, ya que estos recursos específicos a la lengua y al dominio son costosos de desarrollar, lo cual complica aún más la tarea de NER. El aprendizaje no supervisado a partir de corpus sin anotar puede ayudar a obtener mejores generalizaciones a partir de corpus supervisados pequeños, pero no puede sustituir completamente a los corpus anotados. En general, en aprendizaje automático, cuanto más relevantes sean los datos de entrenamiento para la tarea, mejor será el rendimiento del modelo.

Como ya se ha indicado, los primeros sistemas de NER hacían uso de algoritmos basados en reglas construidas a mano, lexicones, características ortográficas, ontologías y listados (*gazetteers*), entre otros. Después, a estos sistemas les siguieron los basados en la ingeniería o extracción de características (*feature engineering*) y el aprendizaje automático (Nadeau & Sekine, 2007). Más tarde, empezando con (Collobert et al., 2011), se popularizaron los sistemas basados en redes neuronales que apenas hacían uso de la ingeniería de características. Estos modelos son interesantes porque no requieren recursos de dominio específico como lexicones u ontologías, y por tanto son más independientes del dominio.

En este contexto, se propusieron varias arquitecturas de redes neuronales, la mayoría basadas en alguna forma de red neuronal recurrente (RNN: *Recurrent Neural Networks*) sobre los caracteres, y *embeddings* de palabras o de las

componentes de las palabras. Estos sistemas basados en redes neuronales sin hacer uso de recursos externos superaron los sistemas del estado del arte en 1.59% en español, 2.34% en alemán y 0.36% en inglés (Yadav & Bethard, 2018). La introducción de la arquitectura *Transformer* y los mecanismos de atención mejoraron aún más el rendimiento de los sistemas, y se han convertido en la tecnología dominante en PLN y, en concreto, en el reconocimiento de entidades en los últimos años.

3.1.1 Anotación de entidades

Al ser una tarea de etiquetado de secuencias de palabras, los *datasets* para NER suelen seguir el tipo de formato IOB2, una variante del formato de etiquetado IOB[58].

El formato IOB (*Inside-Outside-Beginning*) es un formato de etiquetado de tokens. En este formato, el prefijo *I-* antes de una etiqueta indica que la etiqueta se encuentra dentro de una entidad, y la etiqueta *O* indica que el token no pertenece a ninguna entidad. El prefijo *B-* antes de una etiqueta se usa solo para indicar el comienzo de una entidad que va inmediatamente a continuación de otra, sin etiquetas *O* entre ellas.

Token	**Tag**
found	O
under	O
the	O
porch	O
of	O
Christ	B-loc
Church	I-loc
,	O
in	O
Boston	B-loc

Figura 10. Ejemplo de etiquetado en formato IOB extraído del *dataset* Hipe2020[59].

En cambio, en el formato IOB2, el prefijo *B-* antes de una etiqueta indica el comienzo de la entidad, el prefijo *I-* indica que el token o palabra se encuentra

[58] https://en.wikipedia.org/wiki/Inside–outside–beginning_(tagging)

[59] Del *dataset* `HIPE-2022-v2.1-hipe2020-dev-en.tsv` disponible en https://github.com/hipe-eval/HIPE-2022-data/tree/main/data/v2.1/hipe2020/en

dentro de una entidad, y la etiqueta *O* indica que el token no pertenece a ninguna entidad. Este es el más usado, y tiene la forma que se muestra en la Figura 10.

El etiquetado de entidades es una tarea compleja. Metodológicamente, en primer lugar, es necesario definir el conjunto de entidades en base al objetivo para el que se realiza la tarea y a los usos previstos del *dataset* y sus anotaciones. Después, se deben definir los límites de cada entidad, esto es, las partes que hay que anotar y las que no. Por ejemplo, al etiquetar nombres de personas, decidir si se incluye o no el tratamiento (Don, Señor, Excmo, etc.), o, al etiquetar nombres de calles, decidir si se incluye la dirección completa (número y piso), y el tipo (calle, plaza, plazuela, etc.). Para que el *dataset* sea consistente, todas estas decisiones o criterios de anotación se deben reflejar en la guía de anotación, con los ejemplos positivos (cómo anotar) y negativos (qué no anotar).

Para realizar una anotación, se pueden usar herramientas como las presentadas previamente (Brat, Doccano, Tagtog, u otras). Algunas de ellas integran la conversión de las anotaciones al formato IOB, y en otras ocasiones cuentan con un formato propio de exportación, lo que podría hacer necesario el desarrollo de un *script* para convertir el *dataset* anotado a un formato estándar como IOB.

3.1.2 Evaluación de los sistemas de NER

Una metodología de evaluación del rendimiento de un sistema que implementa una tarea de NLP consiste en organizar un evento competitivo (*shared task*) en el que diferentes sistemas comparten datos y objetivos. Por ejemplo, en el caso de un sistema de NER, se trata de identificar y clasificar las entidades presentes en un corpus. A continuación, se comparan los resultados de los diferentes sistemas con un *gold standard* desarrollado por los organizadores esto es, un *dataset* cuyas entidades se han anotado previamente manualmente. La comparación de los resultados de los sistemas se realiza sobre la base de diferentes medidas.

Desde la primera tarea compartida de NER (Grishman & Sundheim, 1996), se han creado muchas otras tareas y *datasets* para la evaluación. En las tareas CoNLL 2002 (Tjong Kim Sang, 2002) y CoNLL 2003 (Tjong Kim Sang & De Meulder, 2003) se crearon los *datasets* homónimos a partir de artículos de periódicos en cuatro idiomas distintos (español, inglés, alemán y neerlandés), centrándose en cuatro entidades: persona (PER), lugar (LOC), organización (ORG) y miscelánea, que incluye otros tipos de entidades (MISC). En las distintas

tareas de NER que se han desarrollado a lo largo de los años, los tipos de entidades nombradas suelen variar según el origen y el dominio del *dataset*, y el idioma.

Un *dataset* actual multilingüe es OntoNotes (Hovy et al., 2006) y su versión final OntoNotes 5.0 (Weischedel, Ralph et al., 2013). Este *dataset* contiene textos de varios géneros (noticias, conversaciones telefónicas, blogs, entrevistas y otros) en tres idiomas (inglés, chino y árabe) con anotación estructural (sintaxis y estructura de argumentos predicados) y semántica superficial (sentido de las palabras vinculado a una ontología y correferencias). OntoNotes incluye aproximadamente 1,5 millones de palabras en inglés, 800K en chino, y 300K en árabe.

Otro *dataset* multilingüe para NER, en este caso un *silver standard* (generado automáticamente) construido a partir de los textos y la estructura de Wikipedia es WikiNER (Nothman et al., 2013), que incluye datos en inglés, alemán, español, neerlandés, ruso, francés, italiano, polaco y portugués, con 3,5 millones de tokens en cada idioma y los mismos tipos de entidades que los corpus anteriores. El procedimiento para construir este corpus consistió en asignar etiquetas de entidad a los artículos de Wikipedia, y usar los enlaces entre artículos para asignar las etiquetas de entidad a los tokens, lo cual da lugar a unas anotaciones bastante aceptables. OntoNotes 5.0 y WikiNER, fueron los dos corpus más utilizados para entrenar los modelos estadísticos de la herramienta SpaCy.

Para evaluar el rendimiento de los sistemas de NER, (Grishman & Sundheim, 1996) utilizaron dos medidas: por un lado puntuaron si el tipo de entidad predicho era el correcto, sin tener en cuenta las fronteras (el principio y final) de las entidades, y por otro evaluaron si las fronteras eran las correctas, independientemente del tipo de etiqueta. Para cada categoría, se definió la *precisión* como el número de entidades predichas correctamente por el sistema dividido por el total de predicciones del sistema, el *recall* como el número de predicciones correctas del sistema dividido por el número total de etiquetas que identificaron los anotadores (es decir, las del *gold standard*), y la *medida F1* como la media harmónica entre la *precisión* y el *recall*.

En CoNLL 2002 y 2003 se introdujeron las medidas de acierto *exacto*, que consideraban una predicción correcta cuando se acertaba tanto el tipo de entidad como su principio y final en la frase. Más tarde, se han usado las medidas *F1 relajada* y *estricta* en numerosas *shared tasks* para permitir comparar las anotaciones entre distintos sistemas, que podrían provocar fronteras distintas al hacer uso de distintas técnicas de segmentación (Yadav & Bethard, 2019). En

general, hoy en día el rendimiento de los modelos y los sistemas se evalúa mediante las medidas mencionadas.

3.1.3 Evolución de los sistemas de NER

La primera vez que se definió la tarea de reconocimiento de entidades fue en la sexta edición de MUC (Message Understanding Conferences) (Grishman & Sundheim, 1996), una serie de conferencias que se diseñaron para promover y evaluar la investigación en ciencias de la computación, en concreto en la Extracción de Información. Uno de los objetivos era el de identificar, a partir de las tecnologías que se estaban desarrollando para la extracción de información, funciones que fuesen útiles, independientes del dominio y que pudiesen llegar a ejecutarse de manera automática con alta precisión a corto plazo. Así, se definió la tarea denominada "entidades nombradas", que básicamente comprendía la identificación de todos los nombres de personas, organizaciones y localizaciones geográficas en un texto, y finalmente incluyó también expresiones de tiempo, moneda y porcentajes. Para marcar en el texto estos elementos se usó el marcado SGML, y las etiquetas iniciales fueron ENAMEX (Entity Name Expression) para los nombres de personas y las organizaciones, y NUNEX (Numeric Expression) para las monedas y los porcentajes, como se muestra en el ejemplo de la Figura 11.

```
Mr.  <ENAMEX TYPE="PERSON">Dooner</ENAMEX> met with <ENAMEX TYPE="PERSON">Martin
Puris</ENAMEX>, president and chief executive officer of <ENAMEX
TYPE="ORGANIZATION">Ammirati & Puris</ENAMEX>, about <ENAMEX
TYPE="ORGANIZATION">McCann</ENAMEX>'s acquiring the agency with billings of <NUMEX
TYPE="MONEY">$400 million</NUMEX>, but nothing has materialized.
```

Figura 11. Ejemplo de las primeras anotaciones de entidades extraído de (Grishman & Sundheim, 1996).

El rendimiento de la tarea superó las expectativas los organizadores, ya que los resultados de los sistemas demostraron un rendimiento muy alto e incluso algunos de ellos fueron comercializados o incorporados a sistemas de procesamiento de textos gubernamentales.

A partir del ciclo de evaluaciones MUC, las tareas relacionadas con las entidades nombradas han evolucionado considerablemente hasta ahora, desde el reconocimiento y clasificación (Nadeau & Sekine, 2007), que se denominó NERC por sus siglas en inglés, hasta la desambiguación y la vinculación (Rao et al., 2013). Se suele referir al conjunto de estas tareas como procesamiento o extracción de entidades nombradas. Hoy en día, *NER* suele hacer referencia a la tarea de

reconocimiento y clasificación, aunque no se haga mención explícita a la clasificación.

A continuación, se presenta una clasificación de los sistemas de NER para ilustrar la evolución hasta las complejas arquitecturas de las que hacemos uso en nuestros días.

Los primeros sistemas de NER eran sistemas basados en conocimiento que hacían uso de lexicones y recursos de dominio específicos. Estos funcionan bien cuando el lexicón es exhaustivo, y en tal caso presentan una precisión alta, pero el *recall* suele ser bajo, precisamente debido a la necesidad de conocimiento y reglas específicas de dominio y la incompletitud de los diccionarios. Además, requieren el conocimiento de expertos del dominio, por lo que construir y mantener estos sistemas y los recursos de dominio resulta costoso.

Algunos de los primeros sistemas basados en aprendizaje automático requerían muy pocos datos de entrenamiento, junto con algunas reglas, como la ortografía (por ejemplo, el uso de mayúsculas), el contexto de las entidades, conocimiento sintáctico, o la frecuencia inversa en el documento (IDF, *Inverse Document Frequency*). También se presentaron sistemas no supervisados para generar los lexicones automáticamente y para resolver ambigüedades en las entidades, como en (Nadeau et al., 2006).

Los sistemas supervisados con ingeniería de características fueron la técnica dominante en la tarea de reconocimiento y clasificación de entidades hasta la primera década de los 2000, es decir, hasta la revolución del PLN con la llegada de las redes neuronales.

Los sistemas de aprendizaje automático supervisados más comunes de este tipo que se usaron para NER incluyen los modelos ocultos de Markov (HMM, *Hidden Markov Models*) (Bikel et al., 1997), las máquinas de vectores soporte (SVM, *Support Vector Machines*) (Asahara & Matsumoto, 2003), campos aleatorios condicionales (CRF, *Conditional Random Fields*) (McCallum & Li, 2003), modelos de máxima entropía (ME, *Maximum Entropy models)* (Borthwick et al., 1998), y árboles de decisión (Sekine, 1998). Estos sistemas se basan en asunciones de independencia, esto es, la predicción de la etiqueta de cada palabra depende solo de las palabras contiguas, y no de las etiquetas de las palabras anteriores. En cambio, los métodos posteriores sí que consideraron la secuencia de etiquetas anteriores a la palabra para decidir la etiqueta actual. Ahora bien, se trata de asunciones muy fuertes que no tienen por qué cumplirse siempre.

Estos sistemas se combinaban con reglas ortográficas, listas de palabras clave, y otros tipos de reglas. Un ejemplo de este tipo de modelos es el ganador de CoNLL 2002, AdaBoost (Carreras et al., 2002), que combinaron árboles de decisión con características como el uso de mayúsculas, palabras clave, bolsas de palabras y otros, consiguiendo un 81,39% en la medida F en el *dataset* español de CoNLL 2002.

Aproximadamente desde 2010, el estado del arte para la mayoría de las tareas de predicción de secuencias en PLN lo han liderado los modelos basados en redes neuronales. La mayoría de estos métodos combinan distintas arquitecturas de redes neuronales en un solo modelo. Una de las más importantes, muy presente en los modelos hasta 2018, son las RNNs (*Recurrent Neural Networks*), diseñadas para codificar la información sobre las palabras anteriores en el texto durante períodos más largos. En concreto, la arquitectura específica que se usa en la mayoría de las modelos para el entrenamiento de RNNs es la red LSTM (*Long Short-Term Memory*), que soluciona algunos problemas que presentan las RNNs.

El primer trabajo que intentó usar LSTMs para NER se publicó en 2003 (Hammerton, 2003). Sin embargo, debido a la falta de potencia computacional, estos modelos eran pequeños y no alcanzaban resultados comparables a los que mejor funcionaban en aquel momento. A partir de 2010 se resuelve este problema de rendimiento gracias a la introducción del uso de las GPUs en el aprendizaje profundo.

En el trabajo de (Collobert & Weston, 2008) se propuso uno de los primeros sistemas de NER basado en redes neuronales con vectores de características construidos a partir de reglas ortográficas, diccionarios y lexicones. Los trabajos posteriores reemplazaron estos vectores construidos a mano con *embeddings* de palabras (Collobert et al., 2011), que son representaciones aprendidas a partir de grandes colecciones de datos sin etiquetar mediante procesos no supervisados como el modelo *Skip-gram* (Mikolov et al., 2013). Los estudios han mostrado la importancia de estos *embeddings* para los sistemas de NER basados en redes neuronales (Habibi et al., 2017).

Así, a partir de 2015 aproximadamente se proponen nuevos métodos para las tareas de etiquetado de secuencias basadas en redes neuronales, métodos que se clasifican según hagan uso de representaciones basadas en palabras, caracteres, subpartes de las palabras, o combinaciones de estas.

Cuando la representación se hace a nivel de palabra, las palabras de la frase se dan como input a las RNN y cada palabra se representa con su *embedding*. En

(Huang et al., 2015) se usa por primera vez la arquitectura LSTM-CRF bidireccional, esto es, se usan dos LSTMs, una que lee el texto de principio a final y la otra en sentido contrario desde el final hasta el principio, combinado con una capa CRF, demostrando que el rendimiento mejora. Además, los *embeddings* (de palabras) generados en cada estado de las LSTM se combinan con reglas construidas a mano (mayúsculas, puntuación, patrones en las palabras, contexto de uni/bi/tri-gramas, etc.)

En las arquitecturas con representaciones a nivel de carácter, las frases se toman como secuencias de caracteres. Estas se pasan a través de la RNN, prediciendo etiquetas para cada carácter, que se convierten en etiquetas de palabras en el paso de post procesamiento. El potencial de estos modelos se puso en evidencia por primera vez con (Kim et al., 2015), que usaron *highway networks*[60] en CNNs (*Convolution Neural Networks*) en secuencias de caracteres, usando después otra capa LSTM y *softmax* para las predicciones finales.

Los sistemas que combinan tanto el contexto de las palabras como los caracteres que las componen, han demostrado muy buen rendimiento en la tarea de NER requiriendo muy pocos recursos o conocimiento específico de dominio. En esta categoría se distinguen dos tipos de modelos. Unos representan las palabras como combinación de un *embedding* de palabra y una convolución sobre los caracteres de las palabras, seguido de una capa Bi-LSTM y usando una capa final *softmax* o CRF para generar las etiquetas. Dentro de este tipo, (Chiu & Nichols, 2016) propusieron un modelo híbrido que combina una LSTM bidireccional con una red CNN. Esta última se usa para crear una codificación de cada palabra, que aprende las características tanto a nivel de carácter como de palabra. El modelo hace uso de *embeddings* de palabras, reglas adicionales construidas a mano y características a nivel de carácter extraídas por la CNN, y todo esto se pasa a la LSTM a nivel de palabra. Consiguieron un 91,62% en la medida F1 en el *dataset* CoNLL 2003 para el inglés.

Los modelos del segundo tipo encadenan los *embeddings* de palabras con LSTMs (bidireccionales o no) sobre los caracteres de la palabra, pasando esta representación a otra Bi-LSTM a nivel de frase, y prediciendo finalmente las etiquetas con una capa *softmax* o CRF. Los primeros en introducir esta arquitectura fueron (Lample et al., 2016), consiguiendo una medida F del 90,94% en los *datasets* CoNLL 2002 y 2003 en inglés. No sólo destacaron por introducir esta nueva arquitectura, ya que hasta 2016, los sistemas principalmente hacían

[60] https://en.wikipedia.org/wiki/Highway_network

uso de reglas construidas a mano y conocimiento de dominio específico para poder aprender a partir de corpus pequeños disponibles para el entrenamiento.

A partir del trabajo de (Lample et al., 2016), ganan popularidad las arquitecturas de redes neuronales que apenas hacen uso de extracción de conocimiento (características) y que consiguen resultados que superan el estado del arte. De hecho, fue el primer trabajo en el que se abandonaron completamente las reglas y se usaron solo *embeddings* (es decir, no usan ningún recurso o característica específica de la lengua más allá de una pequeña cantidad de datos de entrenamiento supervisados y corpus sin anotar). La arquitectura que introdujeron, basada en LSTMs bidireccionales y CRFs asombró por su simplicidad.

En el mismo año, (Ma & Hovy, 2016) proponen un sistema muy parecido al anterior, y que se enmarca en el primer tipo de modelo dentro de los sistemas que combinan tanto el contexto de las palabras como los caracteres que las compone: usan una CNN para codificar la información a nivel de carácter de una palabra en su representación a nivel de carácter; esta se combina con una representación a nivel de palabra y se pasa a una LSTM bidireccional para capturar la información contextual de cada palabra. Después la salida de la Bi-LSTM se pasa a la capa CRF para descodificar la mejor secuencia de etiquetas. De hecho, la CNN que usan es similar a la de (Chiu & Nichols, 2016) salvo que estos solo usan *embeddings* de caracteres como input, sin ningún tipo de característica adicional de los caracteres. Como *embeddings* de palabra usan los de GloVe (Pennington et al., 2014).

En resumen, hasta aproximadamente 2018 se mostraba eficaz el uso de LSTMs bidireccionales (una en cada sentido), junto con una capa CRF al final para modelar la transición de las etiquetas.

3.1.4 *Transformers*, la última revolución del PLN

Las redes *Transformer* son un nuevo tipo de arquitectura introducida en 2017 en el artículo *Attention Is All You Need* (Vaswani et al., 2017), y que ha sido la precursora de la reciente y última revolución del PLN. Esta red procesa el texto en dos fases: una para la codificación, que codifica y extrae la información más relevante, y una de decodificación que se encarga de generar una nueva secuencia de texto, cada una de ellas compuesta por múltiples codificadores y decodificadores respectivamente.

En 2018 apareció una nueva arquitectura desarrollada por Google, que marcó el inicio de una nueva era en el PLN: BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019). Se trata de un codificador que obtiene representaciones bidireccionales a partir de redes *Transformer*. El modelo está pre-entrenado con una cantidad inmensa de datos: 800 millones de palabras del BooksCorpus y 2.500 millones de palabras de Wikipedia en inglés. La idea de BERT se basa en entrenar un modelo base que aprenda a interpretar el lenguaje en general, en vez de proponer modelos que resuelven cada tarea individualmente, para después añadir a este modelo unas capas adicionales que permiten que se especialice en una tarea en particular (en lo que se conoce como *fine-tuning*). Por ello, este tipo de modelos se llaman también *modelos de lenguaje*.

BERT es prácticamente el resultado de coger la red *Transformer* y quedarnos con la parte de la codificación, para obtener una representación del lenguaje. La representación de las palabras (de cada token) viene dada por tres codificaciones: un *embedding* de las palabras, una codificación posicional para saber la posición de las palabras, y un *embedding* que indica a qué segmento pertenece la frase. Estos tres vectores se suman, y se da como input al modelo. Durante el preentrenamiento del modelo, se realizan dos tareas. En primer lugar, el modelo aprende a interpretar el lenguaje mediante la técnica de predecir palabras enmascaradas aleatorias en las secuencias (MLM: *Masked Language Modeling*). Gracias a la bidireccionalidad del preentrenamiento en esta fase, es posible codificar de manera precisa el significado de las palabras dependiendo de su contexto, y por tanto sin ambigüedad. En la segunda fase, el modelo aprende a predecir la siguiente frase a una dada, aumentando aún más la robustez del modelo. Una vez completado el preentrenamiento, afinar el modelo para realizar tareas específicas es sencillo: se añaden dos capas, una red neuronal y una *softmax* y se vuelve a entrenar el modelo, requiriendo mucho menos tiempo.

Así, a partir de 2017 se rompe con la tradición de las técnicas de representación mediante vectores de palabras, para dejar paso a los modelos de lenguaje pre-entrenados. Desde entonces se han publicado diferentes modelos de lenguaje tanto monolingües como RoBERTa (Y. Liu et al., 2019) y GPT-3 (Brown et al., 2020) para el inglés, y recientemente modelos en español como BETO (Cañete et al., 2020), los del proyecto MarIA (Gutiérrez-Fandiño, Armengol-Estapé, et al., 2022), los del proyecto BERTin (de la Rosa et al., 2022) y RigoBERTa (Serrano et al., 2022), como modelos multilingües (cuyas arquitecturas y procesos de entrenamiento son similares a los de sus correspectivos monolingües, salvo que en este caso el corpus usado para el preentrenamiento consiste en documentos en

muchos idiomas) como Multilingual BERT (Devlin et al., 2019), XLM (Lample & Conneau, 2019) y XLM-RoBERTa (Conneau et al., 2020).

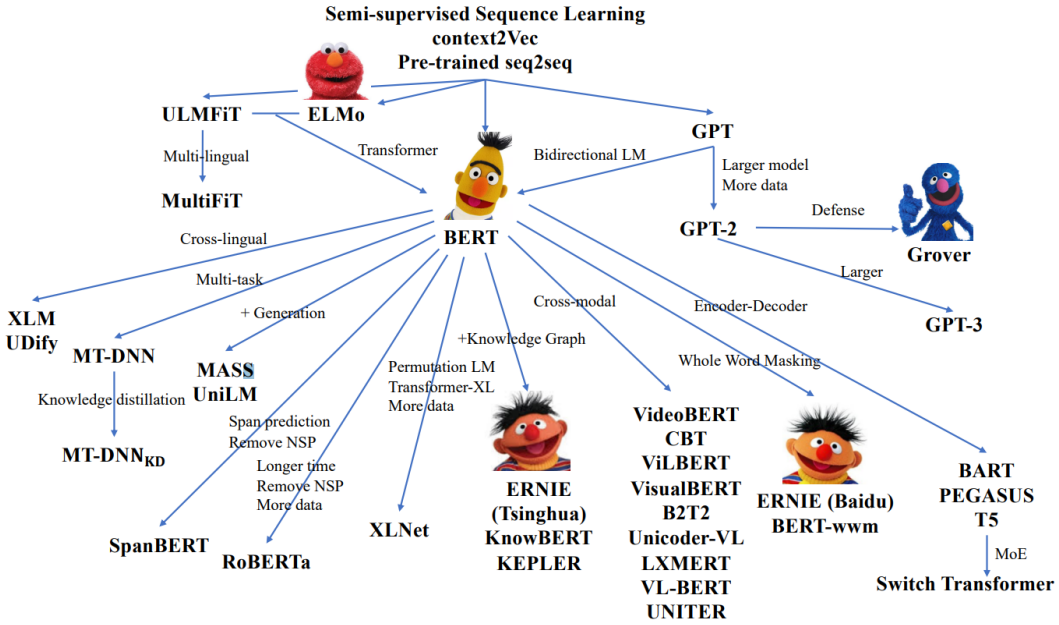


Figura 12. Familia de modelos pre-entrenados. Imagen extraída de *Pre-Trained Models: Past, Present and Future* (Han et al., 2021).

3.1.5 Herramientas para NER

Como se ha visto en los apartados anteriores, las herramientas para el reconocimiento de entidades en textos han variado significativamente a lo largo de los años debido a la mejora de las arquitecturas de los sistemas que se han usado para su implementación. A continuación, se presentan algunas de las herramientas más conocidas, ordenadas cronológicamente para ilustrar la evolución de las tecnologías en las que se basan.

1. Entre los recursos que ofrece **GATE** (Cunningham et al., 2002), un entorno de desarrollo de tecnologías del lenguaje, se encuentra **ANNIE** (*A Nearly-New Information Extraction system*), un sistema de extracción de información, en particular de entidades, que ha sentado las bases de numerosos sistemas de investigación y comerciales. ANNIE permite reconocer diferentes tipos entidades como personas, lugares y organizaciones (estos en particular son los que se encuentran disponibles en su demo online[61]) como también fechas, cantidades de dinero, puestos de trabajo, etc. Inicialmente ANNIE fue desarrollado para procesar documentos en inglés, principalmente

[61] http://services.gate.ac.uk/annie/

noticias y artículos en inglés americano, por lo que sería necesario afinarlo para poder desempeñarse en otros idiomas y poder reconocer las correspondientes entidades.

ANNIE se basa en algoritmos de autómatas finitos y está escrito mediante el lenguaje JAPE. La arquitectura se divide en diferentes funcionalidades o etapas como tokenización, lematización, separación en frases o etiquetado semántico. En GATE *cloud* es posible elegir además el tipo de archivo de entrada (XML, HTML, SGML...) y de salida (XML, JSON).

2. **Stanford NER** (Finkel et al., 2005) es una implementación en Java para el reconocimiento de entidades que fue desarrollada por la Universidad de Stanford en 2005. El sistema etiqueta las secuencias de palabras del texto correspondientes a nombres de cosas como personas u organizaciones, pero también genes o nombres de proteínas. El software proporciona extractores de características para el reconocimiento de entidades y opciones para definir los extractores de características. Además del programa que reconoce las tres clases principales en inglés (persona, organización y lugar), con la descarga se incluyen varios modelos para diferentes idiomas y circunstancias, y es posible usarlo tanto desde la línea de comandos como desde la API de Java.

   Este software se conoce también como *CRFClassifier*, ya que proporciona una implementación general de modelos de secuencias de cadenas lineales, de orden arbitrario, de campos aleatorios condicionales (CRFs). Así, entrenando el modelo con datos etiquetados es posible usar este código para construir modelos de secuencias para NER o cualquier otra tarea. Está disponible online una versión de prueba[62] para explorar la utilidad de la herramienta. Además de las entidades con nombre, en esta demo podemos explorar el etiquetado de otras características: relaciones, sentimiento de las oraciones, análisis gramatical, grafo de dependencias y correferencia entre otros, y podemos elegir entre los idiomas inglés, español, francés, alemán, chino y árabe.

3. **NeuroNER**[63] hace uso del aprendizaje profundo para la detección de las entidades. Permite al usuario crear o modificar anotaciones tanto en corpus nuevos como existentes, es de código libre, gratuito y directo de usar. Puede usarse en Linux, Mac OSX y Windows, y requiere Python 3.5, TensorFlow 1.0 y scikit-learn. El código está disponible en GitHub[64].

[62] http://corenlp.run/

[63] http://neuroner.com

[64] https://github.com/Franck-Dernoncourt/NeuroNER

Los tipos de sistemas para NER basados en redes neuronales, a partir de 2016 empiezan a prescindir del uso de características, ya que aprenden a inferirlas a partir de las representaciones de las palabras y los caracteres, siendo (Lample et al., 2016) los primeros en implementar un modelo de este tipo. Este fue precisamente el modelo que implementaron (Dernoncourt et al., 2017) en NeuroNER. La herramienta también integra el software de anotación Brat para facilitar el desarrollo de modelos de NER en nuevos dominios. NeuroNER alcanzó un 90.50% de medida F en el *dataset* inglés de CoNLL 2003.

4. **SpaCy** y **HuggingFace**. Si bien en la actualidad la librería que lidera la implementación de modelos de PLN es HuggingFace, hace solo algunos años era SpaCy. Estas librerías han aprovechado el auge y la popularidad del aprendizaje automático y el PLN, y han puesto a disposición del usuario implementaciones de los modelos con mejor rendimiento para su reproducción.

   Así, los modelos preentrenados basados en Transformes más actuales, como BERT, RoBERTa o XLM-RoBERTa, se pueden implementar desde estas dos librerías de manera mucho más directa, sin tener que lidiar con el código original de estos modelos.

3.2 Humanidades Digitales y PLN

El término Humanidades Digitales (HD) hace referencia al área de conocimiento y al campo de investigación en la intersección entre los estudios humanísticos y las tecnologías digitales. Describe así tanto las actividades que se llevan a cabo para el desarrollo de herramientas para la investigación y la enseñanza en la era digital, como la producción de trabajo profesional o académico mediante la aplicación de estas nuevas herramientas a las disciplinas de arte, ciencias sociales y humanidades.

Los proyectos en HD responden a preguntas de naturaleza humanística integrando una variedad de formatos multimedia en un entorno dinámico. Se usa la tecnología para el análisis, la estructuración y recogida de datos para permitir a los humanistas encontrar patrones o buscar en grandes colecciones de textos, entre otras tareas.

Los métodos y sistemas relacionados con el PLN son muy utilizados como soporte informático en las HD, ya que son capaces de extraer, procesar y relacionar la información que contienen los documentos para su posterior utilización y que

sirvan de ayuda a los humanistas en sus reflexiones y análisis. Sin embargo, en este momento es necesario reflexionar sobre el modo en que la investigación en humanidades debe evolucionar para hacer frente a la creciente cantidad de recursos digitalizados y herramientas que los gestionan. Además, las soluciones computacionales deben tener en cuenta las necesidades de los humanistas y el grado de familiaridad que estos tienen con la informática.

Por otra parte, la comunidad científica se ha dado cuenta de la dificultad de tratar documentos históricos, y en los últimos años se están realizando esfuerzos por mejorar el acceso y las herramientas disponibles para su consulta (Ehrmann et al., 2020). Las instituciones culturales están llevando a cabo proyectos de digitalización a gran escala donde se obtienen millones de imágenes. Cuando se trata de imágenes que contienen texto (Piotrowski, 2012; Terras, 2011), el contenido se transcribe o bien manualmente con herramientas dedicadas a este fin, como Transkribus, o bien automáticamente mediante la aplicación de procesos de reconocimiento óptico de caracteres (OCR), para poder realizar búsquedas y procesar automáticamente los textos. En este escenario, se revela especialmente interesante la aplicación de tecnologías de PLN de extracción de información, como el reconocimiento de entidades.

3.2.1 NER en las HD

Frente a la necesidad de preservar y acceder al patrimonio cultural, hay un nuevo reto que consiste en adaptar y desarrollar tecnologías del lenguaje capaces de buscar y extraer información de fuentes patrimoniales, y en él juega un papel crucial el reconocimiento y la clasificación de entidades. Gracias a la creciente presencia de grandes cantidades de documentos (históricos o no) digitalizados (Terras, 2011) es posible utilizar las nuevas tecnologías para su análisis. En la actualidad, destacan la adopción de arquitecturas de aprendizaje profundo (Lample et al., 2016) y las representaciones de lenguaje mediante *embeddings*, y en particular los *embeddings* contextualizados (Akbik et al., 2018).

La aplicación de la tarea de NER a textos históricos ha sido el foco de una amplia rama en las investigaciones en PLN y Humanidades Digitales en los últimos años (Ehrmann et al., 2021), ya que reconocer las entidades en textos históricos ayuda a mejorar las búsquedas y la navegación en este tipo de materiales (Neudecker et al., 2014). Sin embargo, aplicar NER a documentos históricos plantea una serie de retos.

Uno de ellos es el margen de error que aún presentan los sistemas de reconocimiento óptico de caracteres (OCR) (Boros, Hamdi, et al., 2020). Los documentos históricos en general son muy heterogéneos y el texto es muy ruidoso (tipografías desconocidas por los sistemas, manchas, etc), lo cual dificulta aún más el reconocimiento de caracteres. Otro problema es la transferencia de conocimiento a nuevos dominios, en este caso de adaptar a documentos antiguos los modelos de NER entrenados con *datasets* en una lengua actual (Baptiste et al., 2021; De Toni et al., 2022). Finalmente, el lenguaje utilizado en siglos pasados suele presentar muchas variaciones respecto al actual, lo cual dificulta la aplicación de recursos como reglas ortográficas o convenciones a la hora de detectar nombres de lugares o personas, al seguir normas distintas (Bollmann, 2019).

Con el fin de avanzar en las investigaciones de este campo se han puesto en marcha distintas iniciativas, como la creación de *datasets* para impulsar el desarrollo y la evaluación de los sistemas de NER en textos históricos (Ehrmann, Romanello, Fluckiger, et al., 2020; Ehrmann, Romanello, Doucet, et al., 2022; Neudecker, 2016). Del mismo modo, se han puesto en marcha eventos como HIPE (*Identifying Historical People, Places and other Entities*) en dos ediciones, 2020[65] y 2022[66], que se llevan a cabo en la conferencia CLEF[67] centrada en la evaluación de sistemas de PLN, extracción y recuperación de información. Los objetivos de HIPE comprenden la mejora de la robustez de los sistemas, permitir la comparación del rendimiento de los sistemas de NER en textos históricos y, a largo plazo, fomentar la indexación semántica eficiente en documentos históricos.

En definitiva, las Humanidades Digitales proporcionan a las tecnologías del PLN un estimulante campo de aplicación, no sólo por la posibilidad de poner a prueba y mejorar los modelos actuales, sino por la utilidad material que estos avances suponen para los humanistas.

3.2.2 Recursos disponibles para dominios específicos

Adaptar las tecnologías existentes a dominios específicos requiere el uso o, en caso de no existir, la construcción de recursos del dominio específico en el que se pretende trabajar. Puede tratarse de recursos léxicos y semánticos como listados

[65] https://impresso.github.io/CLEF-HIPE-2020/
[66] https://hipe-eval.github.io/HIPE-2022/
[67] https://www.clef-initiative.eu

(lexicones o *gazetteers*), tesauros, glosarios, diccionarios, enciclopedias, bases de datos, ontologías, o redes semánticas, que ayudan a detectar o enlazar la información. A continuación, se presentan algunos de estos recursos, que se encuentran descritos con más detalle en el Apéndice II de este trabajo.

La RAE y la Biblioteca Nacional de España (BNE) cuentan con diversos recursos. El Diccionario de autoridades (1726-1739)[68] de la RAE fue su primer diccionario, fundamento de lo que hoy se conoce como el Diccionario de la lengua española. El Nuevo Tesoro Lexicográfico[69] de la lengua española es un "diccionario de diccionarios" que reúne cerca de 70 diccionarios de los siglos XV al XX, en el que se pueden consultar términos descubriendo en qué diccionarios aparecen por primera vez. El Fichero general de la RAE[70], es un fichero que consta de unos diez millones de entradas léxicas y lexicográficas, actualmente digitalizado y disponible para su consulta.

El portal de datos bibliográficos de la BNE[71] propone al usuario un nuevo modo de acercarse a las colecciones y recursos. Es un proyecto de publicación de datos como Linked Open Data, basado en tecnologías y estándares de la Web. El Catálogo de autoridades de la BNE[72] ofrece acceso a más de 300.000 registros de autoridad de los encabezamientos empleados en los registros bibliográficos del catálogo como puntos de acceso, asociados a una persona, entidad corporativa, título o materia.

Existen otros recursos léxicos provenientes de distintos proyectos, como el inventario léxico del corpus CODEA, los Tesauros del Patrimonio Cultural de España[73] o el Inventario léxico del Atlas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid (ALDICAM).

Para establecer relaciones semánticas se pueden usar bases de datos monolingües y multilingües, como WordNet y EuroWordNet (Verdejo Maillo, 1996). WordNet es una base de datos léxica en inglés donde los nombres, verbos, adjetivos y adverbios están agrupados en conjuntos de sinónimos a nivel semántico, o sinónimos cognitivos, llamados *synsets*. Cada *synset* expresa un concepto y entre ellos se relacionan a su vez mediante relaciones semánticas o conceptuales como la hiperonimia, la hiponimia o la meronimia. De hecho, WordNet se usa e

[68] https://www.rae.es/obras-academicas/diccionarios/diccionario-de-autoridades-0
[69] https://www.rae.es/obras-academicas/diccionarios/nuevo-tesoro-lexicografico-0
[70] https://www.rae.es/banco-de-datos/fichero-general
[71] https://datos.bne.es/
[72] https://bnelab.bne.es/dato/catalogo-de-autoridades/
[73] http://tesauros.mecd.es/tesauros/

interpreta como una ontología. EuroWordNet también es una base de datos léxica, en este caso con relaciones semánticas en varios idiomas europeos: alemán, holandés, checo, estonio, italiano, francés y español además del inglés. Cada idioma cuenta con 30.000 conceptos que a su vez se relacionan con el WordNet inglés. Estos recursos son útiles para diversas tareas de PLN porque al representar los términos a nivel conceptual es un recurso idóneo para, por ejemplo, realizar traducciones más precisas, con lo cual es una herramienta útil para la recuperación de información tanto monolingüe (haciendo uso de la sinonimia) como multilingüe (como soporte a la traducción). Del mismo modo, es adecuado para la tarea de *Question Answering* (responder a preguntas) para la expansión de consultas y para desambiguación de las palabras clave.

Otro recurso es BabelNet (Navigli et al., 2021), un diccionario enciclopédico multilingüe y red semántica que conecta los conceptos y las entidades en una gran red de unos 22 millones de entradas. BabelNet sigue el modelo de WordNet basado en synsets, pero lo extiende para contener lexicalizaciones multilingües: cada synset de BabelNet representa un significado dado y contiene todos los sinónimos que expresan ese significado en una gama de idiomas diferentes.

DBPedia (Auer et al., 2007) también es un recurso útil para el enriquecimiento semántico, ya que almacena información semántica a partir de la Wikipedia formalizado en RDF (*Resource Description Framework*), y permite hacer consultas a la base de datos a través del lenguaje SPARQL.

En la actualidad existen muchos problemas para el mantenimiento y acceso a los recursos existentes, por eso se han desarrollado dos infraestructuras europeas, CLARIN[74] y DARIAH[75], que buscan dar soporte a la investigación en humanidades, artes y ciencias sociales, facilitando y unificando el acceso a los diferentes recursos digitales disponibles. Actúan además como pilares para el asesoramiento en la investigación, publicación y colaboración en proyectos, entre los que se involucra la UNED. Uno de los objetivos principales de CLARIN es asegurar que los recursos lingüísticos se almacenan y se ponen a disposición de la comunidad siguiendo los principios FAIR de accesibilidad, interoperabilidad, reusabilidad y localizabilidad (Hinrichs & Krauwer, 2014; Meroño-Peñuela et al., 2020).

[74] https://www.clarin.eu/portal
[75] https://www.dariah.eu

3.3 Caso de estudio: el corpus CLARA-DM

El corpus CLARA-DM se desarrolla en el marco del proyecto CLARA-HD[76] (PID2020-116001RB-C32[77]) de la UNED, que es uno de los tres subproyectos que conforman el proyecto coordinado CLARA-NLP: *Computational Linguistics Approaches to Readability and Automatic Simplification in NLP*. CLARA-HD trabaja en el desarrollo de recursos para el procesamiento del lenguaje en el dominio de las Humanidades Digitales, mientras que los otros dos subproyectos trabajan en los dominios de las narrativas del dominio financiero (CLARA-FINT, PID2020-116001RB-C31) en la Universidad Autónoma de Madrid y el discurso médico (CLARA-MeD, PID2020-116001RA-C33) en el CSIC, respectivamente.

CLARA-HD cuenta con la colaboración del grupo de investigación en Historia del Arte de Alicia Cámara y Álvaro Molina de la UNED, que trabajan en el proyecto CARCEM (*Cartografías de la ciudad en la Edad Moderna: imágenes, relatos, interpretaciones*) (PID2020-113380GB-I00) con el recurso digital de la Biblioteca Nacional de España "Diario de Madrid, siglos XVIII y XIX", sobre la prensa de la vida cotidiana de los siglos XVIII y XIX en Madrid. Durante la primera fase del proyecto se han realizado diversas reuniones entre los grupos CLARA-HD y CARCEM con el objetivo de comprender la dinámica de trabajo de los historiadores del arte e idear soluciones informáticas que puedan ayudar a automatizar algunas de las tareas que realizan. De esta manera, se han ido respondiendo de manera natural a las preguntas que debemos hacernos con relación a la composición del corpus: las necesidades que debe cubrir, el tipo de análisis que se espera realizar, la tecnología que es espera usar, etc.

Descubrimos, sin demasiada sorpresa, que los historiadores del arte se ven obligados a leer y categorizar manualmente cientos de páginas de periódicos, debido a que la mala calidad de las digitalizaciones no permite realizar búsquedas mediante algún tipo buscador automático. Esto se traduce en un trabajo muy costoso en términos de tiempo y esfuerzos. En particular, en el proyecto CARCEM están interesados en lo que se realiza en la ciudad de Madrid, y sobre todo, dónde se realiza, ya que el objetivo es geolocalizar estos eventos en un mapa mediante un sistema GIS (*Geographic Information System*). Para ello, organizan la

[76] http://clara-nlp.uned.es/home/dh/

[77] Proyecto CLARA-HD (PID2020-116001RB-C32). Financiado por MCIN - Agencia Estatal de Investigación (AEI/10.13039/501100011033) en la convocatoria PROYECTOS DE I+D+i (2020) del Programa Estatal de I+D+i Orientada a los Retos de la Sociedad.

información en categorías según los sucesos o los temas tratados en los periódicos. Algunos ejemplos de categorías de su interés son: oficios, establecimientos y servicios; decoración, libros y estampas, animales, ventas y sucesos (incendios, pérdidas, hallazgos…).

Desde el proyecto CLARA-HD se utiliza tecnología de PLN para detectar esos tipos de categorías mediante el reconocimiento de entidades (NER). De aquí se deriva el trabajo que se presenta: desarrollar un modelo de NER que detecte automáticamente los lugares, las personas, los oficios, y otros tipos de entidades útiles para el proyecto CARCEM, que permitan automatizar el proceso de extracción de información de los periódicos del Diario de Madrid.

El proceso para la obtención de este modelo se divide en dos partes principales. En primer lugar, la creación del corpus CLARA-DM, que consta de los periódicos del Diario de Madrid (DM) etiquetados con las entidades acordadas. Y, en segundo lugar, el diseño del entrenamiento de un modelo para el reconocimiento de entidades con el corpus CLARA-DM y la experimentación con diferentes hipótesis y recursos.

En septiembre de 2022 ha tenido lugar en La Coruña el XXXVIII Cogreso de la SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural). En él, varios trabajos (Campillos-Llanos et al., 2022; Martinez-Romo et al., 2022; Ortiz-Zambrano et al., 2022) han señalado que la disponibilidad de corpus de calidad, esto es balanceados y bien anotados, es crítico para el buen rendimiento de los algoritmos basados en modelos del lenguaje. Por eso el esfuerzo de generar el corpus CLARA-DM para contar con un recurso para el reconocimiento de entidades en el ámbito de las Humanidades Digitales.

En los siguientes apartados se detallan los pasos seguidos en la construcción y anotación del corpus que permita el diseño de un modelo de reconocimiento de entidades para este corpus.

3.3.1 Transcripción del Diario de Madrid

El primer paso es la descarga de los periódicos. En este punto cabe preguntarse qué números o años del Diario descargar, en base al interés de los historiadores del arte y a la capacidad de trabajo de la que se dispone. Este es un proceso manual, que puede realizarse desde dos plataformas: la Biblioteca Nacional de

España, que cuenta con los periódicos entre 1788 y 1825[78] y la Biblioteca Digital Memoria de Madrid, con los periódicos entre 1808 y 1814[79].

En la primera fase se descargó el primer día del mes de todos los meses desde 1788 hasta 1825. Después, dado el interés de CARCEM en los primeros números y, en particular del 1791 por haber sido un año de censura, se pasó a descargar el mayor número posible de periódicos entre 1788 y 1791, intentando completar este último.

A 13 de julio de 2022, se tienen unos 750 periódicos descargados, que se distribuyen por años según la Figura 13.

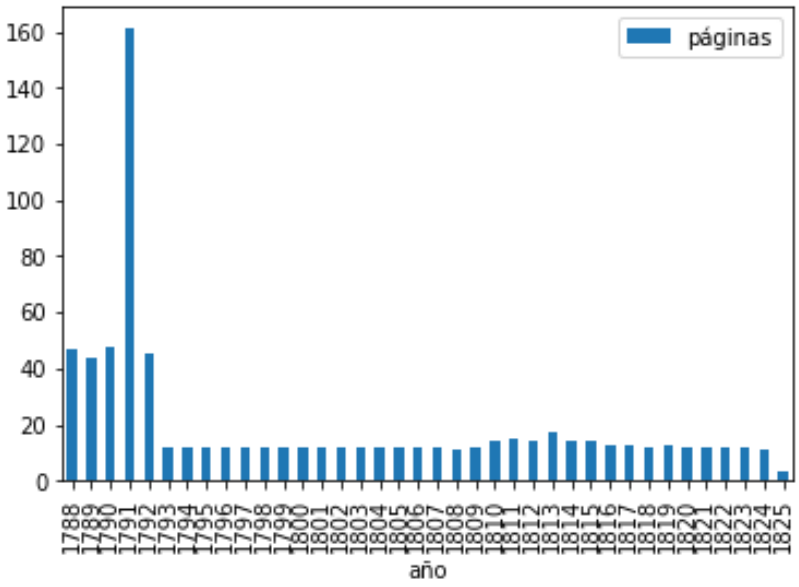


Figura 13. Distribución de los periódicos descargados a 20 de julio de 2020.

Una vez que se tienen los periódicos descargados, el siguiente paso es recuperar el texto que contienen mediante la transcripción. Para ello se elige la herramienta Transkribus, presentada anteriormente.

En primer lugar, se suben los archivos a la plataforma. Para poder transcribir el texto de los documentos es necesario realizar el proceso de reconocimiento de *layout*, que consiste en la aplicación de un modelo de Transkribus que reconoce las regiones y las líneas en las que se encuentra el texto de las páginas. Al aplicar este modelo, se obtienen unas regiones como las que aparecen en la Figura 14. Además de las regiones que contienen texto, también se obtienen muchas regiones que en realidad corresponden a manchas, líneas o logos que no buscamos

[78] http://hemerotecadigital.bne.es/details.vm?q=id:0001510462&lang=es

[79] http://www.memoriademadrid.es/buscador.php?accion=VerFicha&id=4603&num_id=1&num_total=70

transcribir. Por otro lado, el modelo no es perfecto y no siempre es capaz de ordenar las líneas en el orden adecuado de lectura, sobre todo en presencia de tablas o disposición en columnas. Por eso, tras usar este modelo es necesario hacer cambios manuales para eliminar las regiones que no contienen texto, y reordenarlas para que la transcripción resulte coherente.

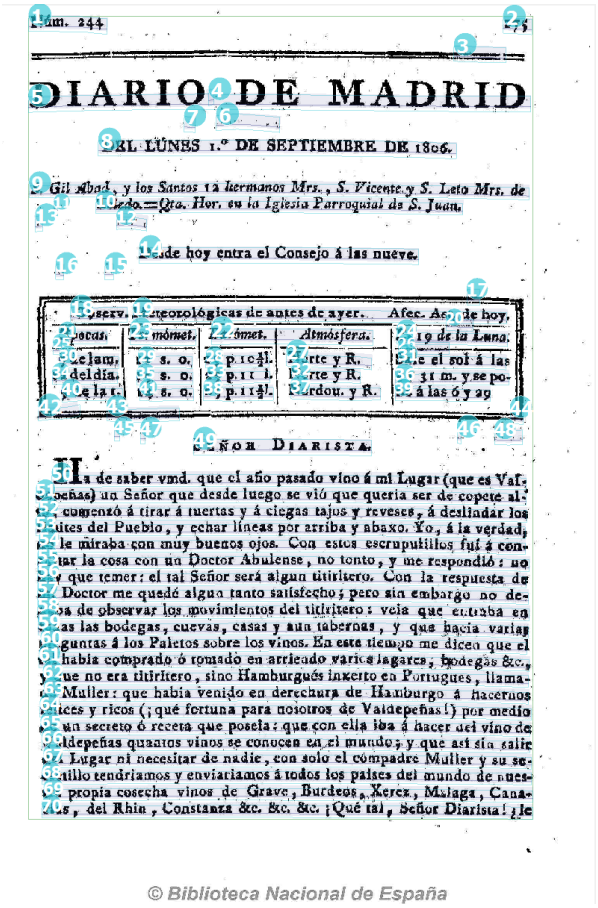


Figura 14. Salida del modelo de reconocimiento de estructura de Transkribus.

Durante los meses de abril, mayo y junio de 2022, se contó con el refuerzo en el proyecto de tres estudiantes de prácticas del ciclo formativo de Grado Superior en Informática para realizar esta tarea. Así, a 20 de julio de 2020 se cuenta con 526 periódicos con el reconocimiento del *layout* realizado, lo que equivale a unas 2291 páginas. Nótese que la Biblioteca Nacional de España cuenta con 13.472 periódicos del Diario de Madrid (1788-1825).

Una vez corregido el reconocimiento de la estructura de los documentos, se puede pasar a transcribirlos. Esto se puede hacer de dos maneras: manualmente, escribiendo línea a línea el contenido del documento, o bien automáticamente (con una posterior corrección manual) aplicando uno de los modelos públicos para la transcripción que ofrece Transkribus. Existen modelos públicos que han sido entrenados con documentos que tienen un tipo de letra parecida a la del Diario

de Madrid por ser de una época cercana, como el modelo "Spanish Golden Age Theatre Prints 1.0". Sin embargo, no consiguen realizar una buena transcripción de los documentos del corpus CLARA-DM. En las pruebas se detecta que no eran capaces de reconocer números, importantes para conservar las fechas, ni las mayúsculas, cruciales para nuestra tarea de detectar entidades. Por eso, hay que entrenar un modelo asociado al corpus para poder transcribir automáticamente el resto del Diario de Madrid.

Para poder entrenar este modelo propio, hay que contar con una cantidad considerable de textos transcritos manualmente. Desde Transkribus recomiendan un mínimo de 75 páginas para entrenar un modelo de reconocimiento de textos manuscritos, lo que equivale a unas 15.000 palabras.

Se realizan distintas pruebas de entrenamiento de modelos de transcripción. El error se mide con el CER[80] (*Character Error Rate*), que indica el porcentaje de error a la hora de predecir los caracteres. Las primeras pruebas se realizan con menos páginas: a febrero de 2022 se tienen 37 páginas transcritas. En estas condiciones se obtiene un error del 4% en el conjunto de validación, con lo cual hay que seguir transcribiendo para poder entrenar con más datos.

En junio de 2022 ya se cuenta con cerca de 200 páginas, con las que se realizan distintas pruebas cambiando los parámetros del modelo (número de épocas, tasa de aprendizaje y documentos seleccionados para el entrenamiento y la validación). Se consigue de esta manera alcanzar un 1% de error. Al comprobar manualmente el rendimiento del modelo se verifica que el rendimiento es muy similar al de la transcripción manual. Este modelo, denominado CLARA-DM, se usará para transcribir automáticamente todas las páginas del Diario de Madrid a las que se aplique previamente el reconocimiento de *layout*.

El último paso antes de pasar a la anotación es la exportación de los documentos. Transkribus ofrece distintas modalidades de exportación, por lo que se han debido tomar algunas decisiones en esta parte.

La exportación a TEI, a pesar de ser popular en las HD, se descarta por quedar una salida demasiado cargada y engorrosa. La exportación a DOCX es útil de por sí para el proyecto CARCEM, ya que permite hacer búsquedas con herramientas de edición. Además, esta exportación es interesante porque cuando una palabra se separa con un guion al final de la frase, esta se une en la transcripción. Esto es necesario para poder mantener las entidades compactas y que los modelos de

[80] https://readcoop.eu/glossary/character-error-rate-cer/

reconocimiento sean capaces de detectarlas. La exportación a texto plano mantiene estos guiones, lo cual complica la aplicación de los modelos de reconocimiento de entidades.

Para el proyecto CLARA-HD y para la experimentación a realizar en este trabajo se necesitan los periódicos en texto plano así que, para solucionar los guiones, se diseña un *script* que convierte los documentos Word a TXT. Así, se cuenta con las exportaciones tanto en formato Word como en texto plano.

Cabe comentar un último aspecto en este proceso. Para poder automatizar completamente todo este ciclo de transcripción, habría sido necesario poder entrenar un modelo de *layout* para evitar el trabajo manual que conlleva. Sin embargo, las pruebas realizadas han evidenciado que esto, de momento, no es posible, concluyendo que el reconocimiento de las regiones de texto es una tarea más complicada que el reconocimiento de caracteres. Para hacer un modelo de transcripción fiable, hubo que revisar manualmente las páginas transcritas y realizar un proceso de normalización. Esto incluye aspectos como la unificación en la forma de transcribir las fracciones, la inclusión u omisión de símbolos como "=" (que se usan antes de la firma de un autor), "&" o "§§", o la corrección de erratas. Después de este proceso, se cuenta con un alto grado de homogeneidad en los datos que el modelo verá en su entrenamiento, y por tanto existen altas probabilidades de un buen aprendizaje.

En el caso de la estructura de los documentos, el grado de heterogeneidad es muy alto: el número de regiones en que se divide el texto, el área que abarcan, los márgenes, el orden en que se leen, etc. Por tanto, los datos de por sí no auspician un buen aprendizaje. Una de las mayores complejidades es la división en columnas cuando el texto se dispone de esta manera, para que puedan leerse en el orden correcto. En la Figura 15 se muestra un ejemplo de ordenación manual de las columnas de un periódico. A pesar de haber entrenado un modelo exclusivamente con páginas que contenían columnas, no se ha conseguido que esta característica se aprenda correctamente. Se realizan tres pruebas, con algunos cambios en los hiperparámetros y en la cantidad y diversidad de los periódicos elegidos tanto para el entrenamiento como para la validación. En la segunda prueba se consigue una ligera mejora respecto al modelo base, aunque no lo suficiente como para evitar el trabajo manual de corrección del *layout*. Para ver si el modelo es capaz de aprender la característica de dividir el texto en columnas, se entrena un tercer modelo con sólo páginas que contienen columnas. Los resultados a priori no son

malos, pero en la práctica, al aplicar el modelo, este sigue sin dividir las columnas. Por tanto, este proceso, por ahora, se mantiene manual.

Figura 15. Ejemplo de página dividida en columnas con un reconocimiento de layout manual.

El libro de estilo para la transcripción de periódicos del Diario de Madrid constituye el Apéndice III de esta memoria.

3.3.2 Anotación de entidades del Diario de Madrid

Con los periódicos transcritos y almacenados en texto plano, se pasa a la fase de anotación. En primer lugar, se decidió la herramienta para el etiquetado. Se compararon cuatro herramientas: Prodigy, Doccano, Brat y Tagtog. La comparativa se muestra en la Tabla 3. Finalmente se eligió la herramienta Tagtog por ser la única que integra una métrica para ver el acuerdo entre anotadores. Además, permite visualizarlo a medida que se anota, lo cual hace el flujo de anotación mucho más dinámico.

En esta parte vuelven a entrar en juego los historiadores del arte del proyecto CARCEM, ya que se debe decidir el conjunto de etiquetas. Por un lado, los historiadores saben la información que necesitan, pero no tienen la perspectiva del informático que sabe qué tipos de categorías es capaz de aprender a generalizar un modelo. Por ello se trata de una tarea delicada que requiere de un esfuerzo de comprensión por ambas partes.

	Prodigy	Doccano	Brat	Tagtog
Funcionalidades generales	etiquetar textos, imágenes y vídeos y entrenar modelos con los datos etiquetados	clasificación de textos, etiquetado de secuencias, tareas secuencia a secuencia	anotación de entidades y relaciones, búsquedas y otras tareas de PLN derivadas	anotar entidades y relaciones, clasificar documentos
Gestión de usuarios	no	sí	sí	sí
Métrica para el acuerdo entre anotadores	no	no	no	sí
Posibilidad de automatización del etiquetado	sí	no	no	sí (de pago)
Formato de los datos de entrada	texto plano, JSONL, JSON, CSV y otros	texto plano, JSONL, CoNLL	texto plano	texto plano, CSV, archivos de código fuente, URLs y otros
Formato de los datos de salida	JSONL	JSONL	.ann	TXT, HTML, XML, CSV, ann.json, EntitiesTSV, y otros
Sistema operativo	Windows, Mac y Linux	Windows, Mac y Linux	Mac o Linux (en Windows se recomienda usar máquina virtual)	Windows, Mac y Linux
Requiere interactuar con la línea de comandos	sí	sí	sí	no (en la versión web)
Necesita Python instalado	sí (3.6+)	sí (3.8+)	sí (2.5+)	no
Se necesitan conocimientos de programación	es deseable tener familiaridad con los entornos en Python	no	se necesitan conocimientos básicos de Linux y servidores Apache	no (en la versión web)
Código abierto	parcialmente	sí	sí	sí
Gratuito	no	sí	sí	sí (distintos planes)

Tabla 3. Comparativa de etiquetadores.

A la hora de decidir el conjunto de etiquetas, estas pueden describir categorías amplias, como persona, lugar, organización, etc., o categorías más finas, como calles y plazas dentro de la categoría lugar, nobles y señores dentro de la categoría persona. Resulta más conveniente comenzar con un conjunto de etiquetas más fino o granulado, ya que la conversión de estas categorías a sus correspondientes versiones generales es una tarea más sencilla que hacer la operación contraria.

A partir del diálogo con CARCEM y de los documentos que nos proporcionan, se elabora una primera propuesta de etiquetas, que se muestra en la Figura 16.

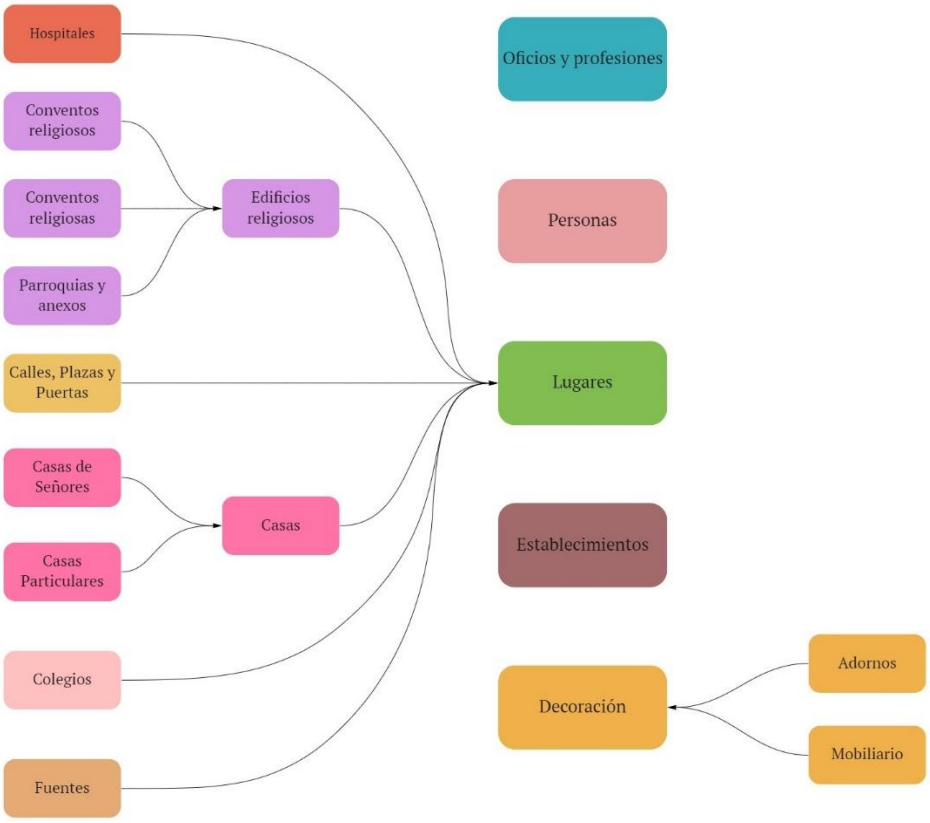


Figura 16. Primera taxonomía de entidades.

Estas entidades se vuelcan en la herramienta de anotación y se realiza el primer ciclo de anotación, como se describe en la sección 2.2.5 (Anotación de corpus). Se eligen cinco periódicos y se anotan por cuatro anotadores mediante un proceso de anotación ciega, es decir, cada persona anota el documento independientemente sin consultar con el resto. Posteriormente la herramienta calcula el acuerdo entre anotadores para cada entidad. Con estas métricas, se vuelve a evaluar la calidad de las etiquetas y se ajusta la taxonomía. En la segunda vuelta la taxonomía de etiquetas es la de la Figura 17.

Junto con la definición de las etiquetas, durante el proceso de etiquetado se elabora una guía de anotación con las directrices sobre lo que anotar o no anotar, y cómo hacerlo, para poder aumentar el número de anotadores en el futuro consiguiendo un acuerdo de anotación alto. Esta parte del proyecto se encuentra en proceso a 22 de julio de 2022, cuando aún contamos con tan solo 5 periódicos etiquetados por al menos 2 personas.

Figura 17. Segunda taxonomía de entidades.

Quedan pendientes algunas reflexiones acerca del etiquetado y el entrenamiento de los modelos de aprendizaje automático. En particular, hay algunos listados de entidades elaborados manualmente en el proyecto CARCEM, que podrían ser de utilidad para pre-etiquetar los documentos con las entidades presentes en estos listados, lo cual podría agilizar el proceso.

Una vez se tengan los periódicos etiquetados, es necesario convertirlos al formato IOB para poder entrenar los modelos de aprendizaje automático. La herramienta elegida, Tagtog, proporciona distintos formatos de exportación, entre los que se encuentra ann.json. Para obtener los documentos en formato IOB, se diseña un nuevo *script* que transforma la salida de Tagtog en formato *EntitiesTsv* al formato IOB.

La primera propuesta de guía de estilo para el etiquetado de entidades del Diario de Madrid (1788-1825) se encuentra en el Apéndice IV de esta memoria.

3.4 Discusión

El reconocimiento de entidades es una tarea que se demuestra especialmente compleja en dominios específicos con pocos recursos. En particular, los documentos históricos presentan un reto particularmente difícil. En esta sección se exponen y ejemplifican algunas de estas dificultades, justificando así la elección de abordar el caso de estudio descrito en la sección anterior, y motivando los experimentos que siguen en el próximo Capítulo 4.

1. Textos antiguos y OCR

Los documentos históricos requieren un proceso de digitalización y transcripción que añade ruido al texto, ya que normalmente la digitalización se ha realizado con escáneres de poca calidad o el documento estaba muy deteriorado. En los *datasets* de la competición HIPE hay numerosos ejemplos[81] de estos errores en los modelos de OCR, como la sustitución de unas letras por otras que se muestra en la Figura 18, o la inserción de caracteres dentro de las palabras debido a manchas o errores en el OCR que muestra la Figura 19.

Token	Tag
from	O
the	O
State	B-loc
of	I-loc
Rboda	I-loc
-	I-loc
Island	I-loc
,	O

Token	Tag
Treaty	O
of	O
Peace	O
with	O
Frtnce	B-loc
.	O

Figura 18. Dos ejemplos de reemplazamiento de caracteres en Hipe2020.

En otras ocasiones, los sistemas de OCR no son capaces de detectar números o letras mayúsculas, que son características fundamentales para la detección de entidades.

Estos errores repercuten negativamente en los sistemas de reconocimiento de entidades (van Strien et al., 2020), ya que inhiben muchos de los mecanismos de activación de los que dependen. Por ejemplo, las entidades con caracteres adicionales o caracteres cambiados no estarán presentes en los listados de

[81] Extraídos del conjunto de validación del dataset Hipe2020 inglés: https://github.com/hipe-eval/HIPE-2022-data/blob/main/data/v2.1/hipe2020/en/HIPE-2022-v2.1-hipe2020-dev-en.tsv

entidades. Además, estas sustituciones, inserciones o eliminaciones, pueden empeorar el rendimiento de los tokenizadores de los que dependen los modelos basados en *Transformers*, y por tanto su rendimiento (Boros, Pontes, et al., 2020).

Token	Tag
as	O
ibe	O
Adjutant	B-pers
of	I-pers
Buonaparte	I-pers
,	I-pers
Ci	I-pers
¬	I-pers
tizen	I-pers
Duroc	I-pers

Token	Tag
of	O
the	O
Island	O
of	O
Barha	B-loc
‘	I-loc
does	I-loc
,	O

Figura 19. Dos ejemplos de adición de caracteres indeseados en los *datasets* Hipe2020.

2. Ortografía no normalizada

Los modelos actuales normalmente están entrenados con textos en español actual, que en general sigue unas normas ortográficas y morfológicas muy distintas a las de hace pocos siglos.

Por ejemplo, las mayúsculas son una característica ortográfica clave que evidencia la presencia de entidades en idiomas como el español, el francés y el alemán; información que los *embeddings* a nivel de carácter pretenden capturar. Estas representaciones se aprenden a partir de grandes corpus. Lo que ocurre con los textos históricos, es que ni se suele disponer de la cantidad suficiente de textos, ni el uso de las mayúsculas sigue las mismas normas que en la actualidad. En el Diario de Madrid, el uso de mayúsculas en muchas ocasiones no responde a la mención de una entidad propia, sino que a veces se utilizan para dar énfasis, como se muestra en la Figura 20.

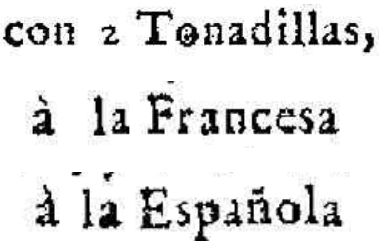


Figura 20. Ejemplos de uso de mayúsculas en el Diario de Madrid.

Otro ejemplo en español de ortografía no normalizada es el uso distinto de la *y* o la *j* y la *x*, como muestra la Figura 21:

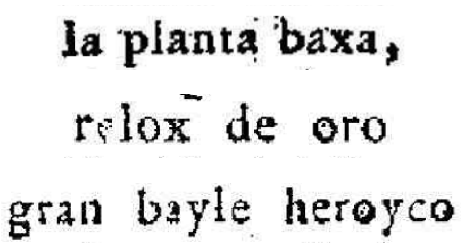


Figura 21. Ejemplos de ortografía no normalizada en el Diario de Madrid.

3. Variación de las entidades en el tiempo

Los textos antiguos contienen entidades a menudo extrañas, que han sufrido cambios significativos a lo largo de los años o que incluso que ya no existen, como localizaciones, para las cuales serían necesarios recursos lingüísticos y bases de datos adecuadas que ayuden a identificarlas (van Hooland et al., 2015).

4. Distintos *datasets*, distintos conjuntos de entidades

La mayoría de *datasets* a gran escala existentes y útiles para la tarea de NER cuentan con las etiquetas genéricas de persona, lugar y organización, y en ocasiones una miscelánea para las entidades que no se encuadran en las anteriores (como el CoNLL 2002 y 2003). Otros *datasets* incluyen un conjunto de etiquetas mucho más amplio, como el OntoNotes 5.0, que incluye 18 tipos de entidades como el tiempo, dinero, cantidades, productos, etc. Cuando se diseña un dataset para un proyecto específico, pueden surgir categorías de entidades que no están presentes en ningún otro dataset.

Por ejemplo, en el corpus CLARA-DM se han identificado entidades de interés relacionadas con el dominio de estudio de los humanistas como la decoración, los oficios o las casas de señores. Así, para aplicar modelos de NER genéricos a *datasets* específicos es necesario eliminar las etiquetas que no estén presentes en el dataset original, y cambiar los nombres de las etiquetas para que se ajusten a las que el modelo conoce. En el caso de CLARA-DM, sería necesario identificar todas las etiquetas de lugar (conventos, hospitales, colegios, etc) con la etiqueta genérica de lugar, con la pérdida de información y el desaprovechamiento de los esfuerzos de anotación que ello conlleva. Por tanto, para que los modelos de NER puedan identificar estas clases nuevas, es necesario contar con suficientes datos etiquetados para el entrenamiento de nuevos modelos.

Como ya se ha expuesto anteriormente, son múltiples los esfuerzos que se están desarrollando en los últimos años para mejorar la extracción de información en textos históricos (García-Serrano et al., 2022), en concreto mediante el reconocimiento de entidades. Para ello se han desarrollado corpus o *datasets* específicos que ayudan a entrenar y comparar los modelos, y se han puesto en marcha tareas competitivas que incentivan la mejora de las técnicas actuales.

Para el desarrollo de un modelo de reconocimiento de entidades sobre el corpus CLARA-DM se tendrán que abordar distintas estrategias, que podrán incluir el uso de los mencionados recursos disponibles. Cuando se cuenta con suficientes datos de entrenamiento, es posible aprovechar los modelos existentes en español mediante el ajuste fino (*fine-tuning*). Esta solución sigue presentando el problema de que los modelos están entrenados con textos actuales, mientras que nuestro corpus contiene textos históricos con errores de reconocimiento óptico de caracteres y una ortografía distinta en muchos casos. Además, para el ajuste de modelos se considera pequeño un corpus con menos de 10.000 ejemplos para el entrenamiento, y estudios demuestran que los modelos basados en BERT pueden ser inestables al entrenar con corpus tan pequeños (T. Zhang et al., 2021).

Al no contar con suficientes datos en el corpus CLARA-DM por encontrarse en construcción, se deben buscar otras aproximaciones. En presencia de escasez de datos, es común recurrir a la estrategia del aprendizaje por transferencia.

El aprendizaje por transferencia (en inglés *transfer learning*) se refiere al conjunto de métodos que permiten transferir los conocimientos adquiridos durante la resolución de un problema para la resolución de otros problemas. Esto es, en vez de abordar una tarea desde cero, se aprovechan las representaciones obtenidas en tareas parecidas para inicializar los modelos. En este sentido, los *Transformers* preentrenados (como BERT y GPT-3) han revolucionado el campo del PLN, permitiendo adaptar estos modelos que encierran una gran cantidad de conocimiento a dominios con pocos datos supervisados.

Este trabajo se propone abordar este problema con el fin de obtener un modelo de reconocimiento de entidades para el corpus CLARA-DM. En la siguiente sección se motiva la necesidad del modelo y se presenta la propuesta de este trabajo, para motivar los experimentos que se describen en el capítulo 4.

3.5 Propuesta

Los obstáculos descritos evidencian la necesidad de desarrollar recursos específicos, o bien mediante la recolección de grandes corpus con los que los modelos puedan aprender, o bien mediante la adición de características específicas a la ortografía de su tiempo, o mediante la creación de recursos externos como listados de entidades de la época. En el caso del corpus CLARA-DM, que se encuentra en desarrollo, estos problemas no son menores. Por eso, en este trabajo se proponen distintas estrategias preliminares para abordar esta primera fase de desarrollo de un modelo de NER.

En la tarea compartida HIPE (*Identifying Historical People, Places and other Entities*) celebrada en 2020 y 2022 (Ehrmann, Romanello, Fluckiger, et al., 2020; Ehrmann, Romanello, Najem-Meyer, et al., 2022), los participantes se enfrentaron al problema de reconocer entidades en textos históricos, en concreto en el corpus Hipe2020 (Ehrmann, Romanello, Clematide, et al., 2020). Se trata de una colección de documentos digitalizados en tres idiomas: inglés, francés y alemán. Los documentos provienen de archivos de distintos periódicos suizos, luxemburgueses y americanos. El dataset fue anotado siguiendo la guía de anotación de HIPE (Ehrmann, Watter, Romanello, et al., 2020), que a su vez se derivó de la guía de anotación Quaero (Rosset et al., 2011). El corpus usa el formato IOB, proporcionando conjuntos de entrenamiento, test y validación para el francés y el alemán, y sin corpus de entrenamiento para el inglés. El objetivo de la tarea consiste en ganar nuevos conocimientos y prospectivas sobre la transferibilidad de los enfoques de reconocimiento de entidades a través de los distintos idiomas, períodos de tiempo, tipos de documentos y conjuntos de etiquetas de anotación.

Los participantes adoptan distintas estrategias, la mayoría de ellas aprovechando la arquitectura *Transformer*, mediante el ajuste de modelos basados en BERT, consiguiendo buenos resultados.

Por tratarse de una tarea muy parecida a la nuestra, que comprende el reconocimiento de entidades en textos históricos, y que además cuenta con *datasets* específicos para ello, nos proponemos abordarla con el objetivo de aprovechar dichos recursos y ganar perspectiva en el enfoque para nuestro dominio, tarea para la cual aún contamos con pocos datos anotados. Se pretende identificar qué tipo de adaptación de procesos y de modelos basados en

Transformers son los que obtendrán mejores resultados para el corpus CLARA-DM, a partir de la experiencia en HIPE.

Así, en primer lugar, se experimentará con los *datasets* de Hipe2020 mediante la aplicación de modelos más actuales, como RoBERTa. De hecho, algunos participantes de HIPE 2020 recomendaron su uso para mejorar sus resultados (Provatorova et al., 2020). En concreto, se usará la versión multilingüe XLM-RoBERTa, que ha demostrado superar el rendimiento de BERT multilingüe.

Después, se experimentará con el aprendizaje por transferencia en el corpus CLARA-DM. Comprobaremos el rendimiento de sistemas tanto en español como en otros idiomas, entrenados en documentos antiguos o actuales. Tal y como demuestran (Vilain et al., 2007), transferir los sistemas de NER de un dominio a otro no es sencillo, especialmente cuando el dominio objetivo cuenta con pocos recursos, y encontramos artículos que abordan el problema en conferencias actuales (Z. Liu et al., 2021; X. Zhang et al., 2022). Además, comprobaremos cómo el rendimiento de sistemas que han sido desarrollados para textos actuales se ve afectado al aplicarlos a textos históricos debido al error añadido de los sistemas de OCR (van Strien et al., 2020), y a otros factores como los descritos en la sección anterior.

Capítulo 4. Modelos y evaluaciones

El objetivo de este capítulo es indagar sobre el funcionamiento de los modelos basados en Transformers que lideran el estado del arte, y experimentar con ellos de cara a obtener un modelo eficiente de reconocimiento de entidades en nuestro corpus CLARA-DM. Para ello se usarán distintos modelos basados en los modelos RoBERTa (Y. Liu et al., 2019), monolingüe, y XLM-RoBERTa (Conneau et al., 2020), multilingüe. El primero es interesante por haber superado a BERT (en las referencias GLUE (General Language Understanding Evaluation) (Wang et al., 2018), RACE (ReAding Comprehension from Examinations) (Lai et al., 2017) y SQuAD (Stanford Question Answering Dataset) (Rajpurkar et al., 2018)), mientras que el segundo es especialmente prometedor, por su buen rendimiento en las tareas en las que se cuenta con pocos recursos.

En primer lugar, se realizarán varios experimentos con el dataset Hipe2020, por tratarse de textos históricos para la tarea de NER, que además presentan problemas de OCR similares a los de nuestro dataset. A septiembre de 2022 se cuenta con tan solo cinco documentos anotados en el corpus CLARA-DM. Por ello, además del enfoque de entrenamiento (fine-tuning) con estos pocos documentos, se adoptarán estrategias de transferencia de conocimiento (transfer learning), que incluyen la aplicación directa de modelos pre-entrenados en otros *datasets* (zero-shot learning), y el entrenamiento con pocos datos de modelos pre-entrenados en otros *datasets* (few-shot learning). Así, tras experimentar con el dataset Hipe2020 se tendrá una idea de cómo funcionan estos modelos, y se delinerá la forma de abordar nuestro dataset para la tarea de NER.

4.1 RoBERTa y XLM-RoBERTa

Para realizar los experimentos de este capítulo se utilizarán una serie de modelos que se basan o en el modelo monolingüe RoBERTa (Y. Liu et al., 2019) o en el modelo multilingüe XLM-RoBERTa (Conneau et al., 2020).

RoBERTa es el acrónimo de *Robustly optimized BERT pretraining approach*, es decir, se trata de un modelo con modificaciones en el proceso de pre-entrenamiento de BERT, en concreto para mejorar el rendimiento de las tareas de aplicación. RoBERTa fue entrenado con un corpus de 160GB de datos (10 veces más grande que el utilizado para BERT) que incluyen Wikipedia, *CommonCrawl News*, *Book*

Corpus y *Webtext Corpus*. El modelo consigue una mejora de entre un 2 y un 20% en las tareas en las que se evaluó BERT.

Por su parte, XLM-RoBERTa[82], siguiendo los trabajos de XLM y de RoBERTa, es una versión multilingüe de RoBERTa, pre-entrenado (mediante MLM) en 2,5TB de datos de *CommonCrawl* en 100 idiomas. Este modelo superó tanto a XLM como a BERT multilingüe por un gran margen, especialmente cuando se trata de idiomas con pocos recursos, e incluso en la transferencia de conocimiento *zero-shot* (Hu et al., 2020).

En base a estos modelos pre-entrenados se han desarrollado diferentes modelos ajustados a idiomas o tareas específicas, como los que se presentan a continuación, disponibles en la plataforma de HuggingFace, y que se usarán en los experimentos de este capítulo:

1. Modelos monolingües para el inglés, francés y alemán:
    a. **DistilRoBERTa**[83] es una versión "destilada" o comprimida del modelo RoBERTa. Mientras que el modelo base de RoBERTa[84] cuenta con 125M de parámetros, el de DistilRoBERTa cuenta con 82M.
    b. **DistilCamemBERT**[85]: CamemBERT es una adaptación de RoBERTa con un corpus francés (Delestre & Amar, 2022), y DistilCamemBERT es la versión destilada del modelo CamemBERT.
    c. **GottBERT**[86] (Scheible et al., 2020) es una adaptación de RoBERTa con un corpus alemán.
2. Modelos multilingües entrenados para la tarea de NER:
    d. **XLM-R-NER-HRL**[87]. Modelo entrenado a partir del modelo base XLM-RoBERTa entrenado para la tarea de NER en 10 idiomas con altos recursos (árabe, alemán, inglés, español, francés, italiano, letón, holandés, portugués y chino), cada uno de ellos con un *dataset* distinto (por ejemplo, el CoNLL 2002 para el español y el holandés y el CoNLL 2003 para el alemán y el inglés) para reconocer las entidades de lugar, persona y organización.

[82] https://huggingface.co/xlm-roberta-base
[83] https://huggingface.co/distilroberta-base
[84] https://huggingface.co/roberta-base
[85] https://huggingface.co/cmarkea/distilcamembert-base
[86] https://huggingface.co/uklfr/gottbert-base
[87] https://huggingface.co/Davlan/xlm-roberta-base-ner-hrl

3. Modelos monolingües para el español:
    e. **RoBERTa-bne**[88]. Modelo basado en RoBERTa, pre-entrenado usando el corpus más grande del español hasta la fecha, con un total de 570GB de datos limpios y sin duplicados procesados para este trabajo, compilados a partir de *crawlings* llevados a cabo por la Biblioteca Nacional de España entre 2009 y 2019. El corpus cuenta con 201.080.084 de documentos y un total de 135.733.450.668 de tokens. Este es uno de los modelos que forman parte del proyecto MarIA[89] (Gutiérrez-Fandiño, Armengol-Estapé, et al., 2022) llevado a cabo en el *Barcelona Supercomputing Center* (BSC). También está disponible la versión *large*.
    f. **BERTin-R**[90]. Modelo basado en RoBERTa desarrollado dentro del proyecto BERTin (de la Rosa et al., 2022), entrenado en la parte española del corpus mC4[91].
4. Modelos entrenados para NER en español:
    g. **RoBERTa-bne-ner-capitel**: se trata del modelo RoBERTa-bne del proyecto MarIA entrenado (en el marco del proyecto MarIA (Gutiérrez-Fandiño, Armengol-Estapé, et al., 2022)) con el *dataset* CAPITEL[92] (Corpus Anotado del Plan de Impulso de las Tecnologías del Lenguaje) para la tarea de NER en español, que cuenta con las entidades de persona, lugar, organización y "otros", en este caso anotadas con el formato BIOES (*Beginning-Inside-Outside-Ending-Single*). También está disponible la versión *large*.
    h. **XLM-RoBERTa-ner-spanish**[93]: Modelo XLM-RoBERTa-large entrenado para NER con la parte española del *dataset* CoNLL 2002.

El entorno de trabajo es un cuaderno de *Google Colaboratory*, que proporciona una GPU NVIDIA Tesla T4 con 16GB de memoria RAM y la versión 11.2 de CUDA. Además, se instalan las librerías Transformers 4.11.3, Datasets 1.16.1, Tokenizers 0.10.3 de HuggingFace y Pytorch 1.12.1+cu113 para la ejecución de los experimentos.

[88] https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne
[89] https://github.com/PlanTL-GOB-ES/lm-spanish
[90] https://huggingface.co/bertin-project/bertin-roberta-base-spanish
[91] https://huggingface.co/datasets/mc4#dataset-summary
[92] https://sites.google.com/view/capitel2020
[93] https://huggingface.co/MMG/xlm-roberta-large-ner-spanish

4.2 Experimentos con el *dataset* Hipe2020

En este primer conjunto de pruebas se ajustarán (*fine-tuning*) modelos basados en RoBERTa (monolingüe) y XLM-RoBERTa (multilingüe) para la tarea de NER en textos históricos.

El *dataset* Hipe2020 cuenta con conjuntos de entrenamiento, validación y test para el francés y el alemán, y de validación y test para el inglés. Está anotado con las etiquetas de persona, lugar, organización, tiempo y producción humana. Contiene 185 documentos en alemán que suman un total de 149.856 tokens, 126 documentos en inglés con un total de 45.695 tokens, y 244 documentos en francés con 245.026 tokens. En total, conforman un corpus de 555 documentos y 440.577 tokens. El *dataset* se somete a un pre-procesamiento para recuperar las frases que componen los documentos y poder pasárselas a los modelos junto con las etiquetas, obteniendo un total de 7.887 frases en francés (de las cuales 5.334 corresponden al entrenamiento —166.217 tokens—, 1.186 a la validación y 1.367 al test), 5.462 frases en alemán (de las cuales 3.185 corresponden al entrenamiento —86.444 tokens—, 1.136 a la validación y 1.141 al test), y 1.437 frases en inglés (938 en la validación y 499 en el test). Nótese que el conjunto de entrenamiento francés es considerablemente más grande que el conjunto de entrenamiento alemán. En cuanto a la distribución de los *datasets*, la Tabla 4 muestra el número de etiquetas de cada tipo presente en cada sección de los *datasets* francés, alemán e inglés respectivamente. Como se puede observar, el *dataset* está desequilibrado, ya que predominan las etiquetas de lugar y persona, seguidas de la de organización, mientras que las clases de tiempo y producción se encuentran notablemente subrepresentadas.

		LOC	**PER**	**ORG**	**PROD**	**TIME**	**Total**
Training	**FR**	3.089	2.525	836	200	276	6.926
	DE	1.740	1.166	358	112	118	3.494
	EN	-	-	-	-	-	-
Validation	**FR**	774	679	159	49	68	1.729
	DE	588	372	164	49	69	1.242
	EN	384	402	118	33	29	966
Test	**FR**	854	502	130	61	53	1.600
	DE	595	311	130	62	49	1.147
	EN	181	156	76	19	17	449

Tabla 4. Distribución de las etiquetas en los *datasets* de Hipe2020.

El *dataset* presenta numerosas inconsistencias debido a la aplicación del OCR y a la etiquetación automática, como las ya mencionadas en el apartado 3.4 (Discusión). En la Figura 22 se muestran algunos ejemplos más: a la izquierda se etiquetan lugares con la etiqueta de organización, mientras que a la derecha se etiquetan como personas.

Token	**Tag**
two	O
kings	O
of	O
France	B-org
king	O
of	O
England	B-org
queen	O
dowager	O
of	O
Poland	B-org

Token	**Tag**
queen	B-pers
of	I-pers
England	I-pers
queen	B-pers
Mary	I-pers
of	I-pers
Hungary	I-pers
queen	B-pers
Eleanor	I-pers

Figura 22. Ejemplos de distintas formas de etiquetar entidades en el *dataset* Hipe2020[94].

En otras ocasiones, la misma entidad se etiqueta con etiquetas distintas, e incluso se encuentran distintas escrituras de la misma entidad, como muestra el ejemplo de la Figura 23.

Token	**Tag**
the	O
Little	B-org
River	I-org
turnpike	I-org

Token	**Tag**
Little	B-pers
Riv	I-pers
«	
r	
Turnpike	I-pers

Figura 23. Ejemplo de distintas escrituras de una misma entidad en Hipe2020[95].

Todas estas inconsistencias dificultan la tarea de predicción que se llevará a cabo a continuación.

[94] Extraídos del conjunto de validación del *dataset* Hipe2020 inglés: https://github.com/hipe-eval/HIPE-2022-data/blob/main/data/v2.1/hipe2020/en/HIPE-2022-v2.1-hipe2020-dev-en.tsv
[95] Idem.

4.2.1 *Fine-tuning* en Hipe2020

Los experimentos aquí descritos se han llevado a cabo contando exclusivamente con los *datasets* de Hipe2020, comparando el rendimiento entre los modelos monolingües y multilingües. Los modelos se configuran para un entrenamiento en 3 épocas y un tamaño del batch de 12 tanto en el conjunto de entrenamiento como en el de validación, y una tasa de aprendizaje de 5e-5. El resto de la configuración de los modelos se establece por defecto mediante las clases AutoConfig, AutoModel y AutoTokenizer de la librería Transformers de Huggingface.

En primer lugar, se realizan pruebas con modelos monolingües. Para el francés se usa el modelo DistilCamemBERT, para el alemán se usa GottBERT y para el inglés se usa DistilRoBERTa. Los resultados de estas tres pruebas, en términos de las medidas micro estrictas de Precisión, *Recall* y F1, al evaluar en los respectivos conjuntos de test los modelos entrenados, se muestran en las tres primeras filas de la Tabla 5.

El *fine-tuning* del modelo DistilCamemBERT se realiza con la parte del entrenamiento del *dataset* francés de Hipe2020. Al evaluar el modelo en el test francés, se obtiene un 0.77 en la medida F1. Al evaluar el modelo en el test alemán se obtiene un 0.19 en la medida F1, lo cual es coherente con el hecho de que el modelo DistilCamemBERT nunca ha visto documentos en alemán. Al evaluar el modelo en el test inglés se obtiene un 0.45 en la medida F1: por un lado, el modelo conoce el idioma inglés al ser una adaptación de RoBERTa, pero, por otro, no ha sido entrenado con documentos históricos en inglés, lo cual explica que el rendimiento supere al del alemán, pero no alcance el del francés.

Después, el *fine-tuning* del modelo GottBERT se realiza con la parte de entrenamiento del dataset alemán de Hipe2020. Al evaluar el modelo en el test alemán se obtiene un 0.72 en la medida F1. Al evaluar el modelo en el test francés, se obtiene un 0.32 en la medida F1, ya que el modelo nunca ha visto este idioma. Al evaluar el modelo en el test inglés se obtiene un 0.41 en la medida F1, de manera similar a lo que ocurría con el modelo DistilCamemBERT.

Para el inglés se usa el modelo DistilRoBERTa. Sin embargo, al no contar con un conjunto de entrenamiento para el inglés, adoptamos una estrategia distinta. En este caso, entrenamos el modelo con los datos de entrenamiento del francés y el alemán. De esta manera, obtendremos un modelo que conoce los tres idiomas (el inglés en la fase de pre-entrenamiento, y el francés y el alemán en la fase de *fine-*

tuning), y que además aprende las características de las entidades en textos históricos durante el *fine-tuning*, luego podremos comprobar si este conocimiento se transfiere al inglés. Al evaluar este modelo en el test francés se obtiene una medida F1 de 0.7, similar a la obtenida por el modelo francés DistilCamemBERT. Al evaluar en el test alemán se obtiene un 0.59 en la medida F1, inferior en 13 puntos respecto a la obtenida con el modelo alemán. Al evaluar en el test inglés se obtienen resultados ligeramente mejores respecto a los anteriores, con una medida F1 de 0.48, solo tres puntos superior a la obtenida con el modelo francés DistilCamemBERT y 7 puntos superior a la obtenida con el modelo alemán GottBERT. Por tanto, el modelo DistilRoBERTa, a pesar de ser monolingüe, ha conseguido transferir algo del conocimiento sobre el reconocimiento de entidades en textos históricos franceses y alemanes, al inglés.

	FR			DE			EN		
	P	R	F1	P	R	F1	P	R	F1
DistilCamemBERT-fr	0.74	**0.8**	0.77	0.13	0.36	0.19	0.38	0.56	0.45
GottBERT-de	0.28	0.38	0.32	0.69	0.75	0.72	0.4	0.52	0.41
DistilRoBERTa-fr+de	0.66	0.75	0.7	0.56	0.63	0.59	0.4	0.6	0.48
XLM-R-fr	**0.76**	**0.8**	**0.78**	0.56	0.72	0.63	0.53	0.61	0.57
XLM-R-de	0.61	0.68	0.65	0.69	0.75	0.72	0.46	0.54	0.5
XLM-R-fr+de	**0.76**	**0.8**	**0.78**	**0.75**	**0.76**	**0.76**	**0.59**	**0.62**	**0.6**

Tabla 5. Experimentos con modelos monolingües y multilingües en los *datasets* francés, alemán e inglés de Hipe2020.

En segundo lugar, se realizan pruebas con el modelo multilingüe. Los resultados se muestran en las tres últimas filas de la Tabla 5. Se entrena el modelo base de XLM-RoBERTa con el *dataset* francés, obteniendo un 0.78 en la medida F1 en el test francés (tan solo un punto más respecto al modelo monolingüe, y ocho puntos más respecto al modelo inglés entrenado en francés y alemán). En el test alemán se obtiene un 0.63, cuatro puntos por encima del modelo inglés entrenado en francés y alemán, y una mejora de más de tres veces respecto al resultado de aplicar el modelo francés al alemán, lo cual demuestra la utilidad del pre-entrenamiento multilingüe. En el test inglés se obtiene un 0.57, superando en 9 puntos al modelo inglés entrenado en francés y alemán, y en 12 puntos al modelo francés. Es decir, el modelo multilingüe entrenado en francés no ha destacado particularmente respecto al modelo francés monolingüe, pero sí ha mejorado los resultados en alemán y en inglés sin ser entrenado para la tarea en esos idiomas. Por tanto, con el modelo multilingüe, además de no perder calidad respecto al monolingüe, se consigue mejorar los resultados en otros idiomas.

Se realiza la misma prueba, pero entrenando el modelo con el dataset alemán. Al evaluarlo en el test alemán se obtiene un 0.72, que al igual que en el caso anterior, no mejora la puntuación del modelo GottBERT monolingüe. En el test francés se obtiene un 0.65 en la medida F1, que es el doble respecto al resultado que se obtuvo al aplicar el modelo alemán GottBERT. En el test inglés se obtiene un 0.5, que mejora en 2 puntos al modelo inglés entrenado en francés y alemán, y en 9 puntos al modelo GottBERT. De nuevo, el modelo multilingüe ha conseguido igualar los resultados respecto a su correspectivo monolingüe cuando se cuenta con datos de entrenamiento para el idioma en cuestión, y además se demuestra útil para transferir el conocimiento a idiomas sin recursos con los que ha sido pre-entrenado (en este caso mejorando los resultados en el francés y el inglés sin haber sido entrenado con textos históricos en estos idiomas).

Al igual que en el caso monolingüe, se realiza la última prueba entrenando el modelo XLM-RoBERTa con los *datasets* francés y alemán. Al evaluar en el test francés, se obtienen prácticamente los mismos resultados que los obtenidos al entrenar sólo con el *dataset* francés, superando por muy poco los resultados del modelo francés monolingüe. Al evaluar en el test alemán se obtiene un 0.76 en F1 y Recall, y un 0.75 en Precisión, los mejores resultados de estas seis pruebas, superando tanto al modelo monolingüe como al modelo multilingüe entrenado solo en alemán por 4 puntos. Al evaluar en el test inglés se obtienen un 0.6 F1, un 0.59 en Precisión y un 0.62 en Recall, los mejores resultados hasta el momento, superando en 12 puntos al modelo inglés entrenado en francés y alemán. Por tanto, la estrategia de entrenar con ambos *datasets* en un modelo multilingüe ha resultado exitosa para mejorar el rendimiento en el dataset inglés.

En conclusión, en caso de contar con datos para el entrenamiento, como es el caso del francés y el alemán, apenas se ha observado diferencia entre el uso de un modelo monolingüe o multilingüe. En cambio, el uso de un modelo multilingüe sí se ha demostrado útil para mejorar los resultados en un idioma sin recursos como es el inglés en este caso, al contar con datos de entrenamiento de la misma naturaleza en otros idiomas en los que el modelo ha sido pre-entrenado (el francés y el alemán). Por otro lado, cabe destacar que los mejores resultados se han obtenido en el *dataset* francés de Hipe2020, lo cual puede deberse al hecho de que es el *dataset* que cuenta con más frases para el entrenamiento.

4.2.2 *Datasets* de NER genéricos multilingües

Visto el éxito del modelo multilingüe en los experimentos anteriores, en esta sección se realizan nuevos experimentos con otro modelo multilingüe. En este caso no se entrenará con los *datasets* de Hipe2020, sino que se evaluará en ellos el rendimiento de un modelo XLM-RoBERTa entrenado para la tarea de NER en 10 idiomas con altos recursos (denominado *XLM-R-ner-hrl*) (*zero-shot*) y después se entrenará con Hipe2020 (*fine-tuning*). Este modelo ha sido entrenado con distintos *datasets* (el CoNLL 2002 para el español y el holandés, y el CoNLL 2003 para el alemán y el inglés, entre otros) para reconocer las entidades de lugar, persona, organización, y tiempo, por lo que en los *datasets* de Hipe2020 tendremos que eliminar las etiquetas que no están presentes, en este caso solo la de producción humana. También será necesario cambiar los nombres de las etiquetas en los *datasets* de Hipe2020 para que se ajusten a los del modelo base: por ejemplo, en CoNLL 2002 y 2003 la etiqueta de persona es 'B-PER', mientras que en Hipe2020 es 'B-pers'.

En este experimento, por un lado, contamos con la ventaja de ser un modelo multilingüe entrenado en varios idiomas para la tarea de NER, y por otro, con la desventaja de no haber sido entrenado en textos históricos.

Los resultados de evaluar este modelo directamente en Hipe2020 se muestran en la primera fila de la Tabla 6, rondando los 50 puntos en la medida F1 en los tres idiomas. A pesar de ser un modelo multilingüe ajustado para la tarea de NER en varios idiomas, no es suficiente para reconocer las entidades en textos históricos al mismo nivel que cuando se entrena con datos específicos para ello, como es de esperar. Además, debido al bajo rendimiento, esta estrategia no compensa el hecho de haber sacrificado una de las etiquetas presentes en el *dataset* Hipe2020.

Veamos qué ocurre si entrenamos este modelo con los *datasets* de Hipe2020: es decir, a los parámetros de NER genéricos del modelo base, les estamos añadiendo el conocimiento del NER en textos históricos. De la misma manera que antes, lo entrenamos primero con el dataset francés, después con el dataset alemán, y por último con ambos *datasets* a la vez. Los resultados se muestran en las tres últimas filas de la Tabla 6.

Al entrenar el modelo con la parte francesa del dataset Hipe2020 y evaluar el modelo en los conjuntos de test, observamos que los resultados mejoran ligeramente respecto a los del mismo experimento realizado con el modelo base XLM-RoBERTa: 1 punto en la medida F1 en el francés, 5 puntos en el alemán y

2 puntos en inglés. Al entrenar con la parte alemana sucede lo mismo, los resultados mejoran respecto al mismo experimento realizado con el modelo base XLM-RoBERTa al evaluar en todos los conjuntos de test, en este caso de manera más notable: 7 puntos en el francés, 3 en el alemán y 10 en el inglés. Por último, al entrenar con los *datasets* francés y alemán, vista la dinámica de los experimentos hechos hasta ahora, cabe esperar obtener los mejores resultados. En efecto, se obtienen las mejores medidas F1 para los tres idiomas, con valores que alcanzan el 80% para el francés y el 78% para el alemán, e incluso la mejor Precisión para el francés y el alemán, y el mejor Recall para el alemán y el inglés. En concreto, se han mejorado las medidas F1 en 2 puntos en el francés y el alemán, y 4 en el inglés.

Los resultados de las pruebas descritas se resumen en la Tabla 6. Respecto a los experimentos realizados utilizando solo los *datasets* de Hipe2020, podemos concluir que el uso de un modelo entrenado con un dataset de NER genérico para el posterior entrenamiento con un dataset específico como es el Hipe2020 ha mejorado ligeramente los resultados. Ahora bien, al no tratarse de una mejora excesiva, se debería sopesar, según las necesidades del proyecto, la rentabilidad o no de haber prescindido de una de las etiquetas de Hipe2020.

	FR			DE			EN		
	P	R	F1	P	R	F1	P	R	F1
XLM-R-ner-hrl	0.54	0.6	0.56	0.53	0.56	0.54	0.46	0.54	0.5
XLM-R-ner-hrl-fr	0.77	**0.82**	0.79	0.67	0.7	0.68	0.56	0.63	0.59
XLM-R-ner-hrl-de	0.71	0.73	0.72	0.73	0.77	0.75	**0.64**	0.57	0.6
XLM-R-ner-hrl-fr+de	**0.78**	0.68	**0.8**	**0.76**	**0.8**	**0.78**	0.6	**0.68**	**0.64**

Tabla 6. Evaluación en Hipe2020 de XLM-RoBERTa entrenado para NER en 10 idiomas con altos recursos.

4.3 Experimentos con el *dataset* CLARA-DM

El *dataset* CLARA-DM cuenta con un conjunto de etiquetas amplio y original. Esto implica que, para poder obtener un modelo de reconocimiento de entidades específico para el *dataset*, será necesario contar con suficientes datos de entrenamiento. Visto que contamos con muy pocos datos etiquetados a septiembre de 2022, adoptaremos dos estrategias para realizar las primeras pruebas, por un lado, haciendo uso de modelos entrenados con *datasets* de NER externos (generalistas o específicos), y por otro, entrenando con los pocos datos con los que se cuenta.

El conjunto de etiquetas del *dataset* es extenso: incluye las etiquetas genéricas de persona, lugar, establecimiento, profesión, adornos, mobiliario, ventas y pérdidas o hallazgos, y además las subetiquetas persona_señores (para nobles, altos cargos, etc), lugar_dirección (para calles, plazas, puertas), lugar_religioso (conventos, parroquias), lugar_hospital, lugar_colegio y lugar_fuente. En total, forman un conjunto de 14 etiquetas, que al duplicarse en el formato IOB y junto con la etiqueta vacía 'O', suman un total de 29 etiquetas. Esto aumenta la complejidad a la hora de que los modelos aprendan, y por tanto la necesidad de contar con suficientes datos de entrenamiento. Por otro lado, para aplicar aprendizaje *zero-shot*, hay que cambiar los nombres de las etiquetas y simplificarlas para que se sean las mismas que las de los *datasets* de entrenamiento de los modelos, con lo cual se perderán las etiquetas más específicas (como lugares religiosos, adornos u objetos en venta) y se reducirá drásticamente el tamaño del conjunto de etiquetas, con la pérdida de información y de esfuerzos realizados durante el proceso de etiquetado que eso conlleva.

Para realizar estas pruebas se cuenta con 5 periódicos etiquetados del Diario de Madrid, que han sido obtenidos a partir de las anotaciones de entre 3 y 4 anotadores, y fusionadas mediante el método del voto mayoritario[96]. Tras la fase de pre-procesamiento, donde se usa el paquete spaCy para poder delimitar las frases que componen los periódicos, se obtiene un *dataset* formado por 928 frases, con un total de 15.145 tokens. La distribución de las etiquetas en estos documentos se puede ver en la Tabla 7. La clase con más ejemplos con diferencia es la de persona, seguida de las de dirección, profesión, establecimiento, lugar, y señores. Las clases de lugares religiosos, pérdidas, colegios, hospitales y adornos están notablemente subrepresentadas, y las de fuentes y mobiliario ni siquiera aparecen en el *dataset* (debido a que no habrán sido anotadas por al menos dos personas y por tanto no aparecen en la versión final).

PER	SEÑOR	LOC	RELIG	DIREC	COLE	HOSP	PROF	ESTABLEC	VENTA	PÉRDIDA	ADOR	**Total**
347	49	78	28	112	1	2	89	81	32	14	4	837

Tabla 7. Distribución de las etiquetas en el dataset CLARA-DM.

[96] https://docs.tagtog.com/collaboration.html#automatic-adjudication-by-majority-vote

4.3.1 *Zero-shot* en CLARA-DM

En esta sección se observa el resultado de aplicar al corpus CLARA-DM modelos pre-entrenados para la tarea de NER en español u otros idiomas, con textos históricos o no, sin ningún tipo de entrenamiento con el corpus CLARA-DM.

La aplicación de cada modelo implica la conversión de las etiquetas de CLARA-DM a las del conjunto de entrenamiento del modelo.

1. Modelos entrenados para NER en español (generalistas)

Para evaluar el modelo *XLM-RoBERTa-ner-spanish* (XLM-RoBERTa entrenado para la tarea de NER en español con la parte española del *dataset* CoNLL 2002), la etiqueta de personas_señores se identifica con la de persona, las de direcciones, hospitales, colegios y lugares religiosos se identifican con la de lugar, las de establecimientos con organizaciones, y el resto (fuentes, profesiones, adornos, mobiliario, ventas, pérdidas y hallazgos) con la miscelánea. Una vez convertidas las etiquetas, se obtienen 221 etiquetas de lugar, 396 de persona, 81 de organización y 139 misceláneas. Los resultados obtenidos rondan el 40% en las tres medidas, lo cual resulta razonable, visto que el modelo no conoce textos históricos, y además se han modificado las etiquetas del *dataset* CLARA-DM.

Después se evalúa el modelo *RoBERTa-bne-ner-capitel* (modelo RoBERTa-bne del proyecto MarIA entrenado con el dataset CAPITEL), y para ello hay que realizar un cambio en el conjunto de etiquetas aún más sofisticado. El modelo cuenta con las etiquetas de persona, lugar, organización y "otros", en este caso anotadas con el formato BIOES (*Beginning-Inside-Outside-Ending-Single*), formando un total de 17 etiquetas. Al convertir las etiquetas de CLARA-DM a este formato y evaluar el modelo, se obtienen, resultados un poco mejores respecto a los anteriores, más cercanos al 50% (Tabla 8). En este caso, el modelo monolingüe vence al multilingüe, lo cual puede indicar que en dominios específicos es importante un conocimiento más profundo del idioma de los textos en cuestión.

	P	R	F1
XLM-RoBERTa-ner-spanish	0.39	0.48	0.43
RoBERTa-bne-ner-capitel	**0.43**	**0.53**	**0.48**

Tabla 8. Evaluación en CLARA-DM de modelos genéricos de NER en español.

2. Modelos entrenados con Hipe2020

Al realizar los experimentos anteriores con el *dataset* Hipe2020 los mejores resultados se obtuvieron al entrenar un modelo multilingüe conjuntamente con los *datasets* francés y alemán. El *dataset* inglés no contaba con parte de entrenamiento, pero, como ahora se trata de evaluar en CLARA-DM, se puede usar la parte de test del inglés para entrenar el modelo. Así, se contará con un modelo entrenado en periódicos históricos en tres idiomas.

La conversión de las etiquetas se mantiene como las anteriores, excepto que en este caso el *dataset* Hipe2020 no cuenta con etiqueta de miscelánea u "otros", luego las categorías de fuentes, adornos, mobiliario, ventas y pérdidas o hallazgos se eliminan (identificándolas con la etiqueta 'O'). Por otro lado, Hipe2020 cuenta con las etiquetas de tiempo y producciones humanas, de las que CLARA-DM carece, por lo que el modelo es susceptible de intentar identificar esta entidad y aumentar el número de errores de predicción. En definitiva, los *datasets* solo "coinciden" (tras realizar las modificaciones convenientes en CLARA-DM) en las etiquetas de persona, lugar y organización.

Al evaluar el modelo XLM-RoBERTa entrenado con la parte francesa de Hipe2020 en CLARA-DM se obtiene una medida F1 de 0.48, la misma que con el modelo *RoBERTa-bne-ner-capitel*, y cinco puntos superior a la obtenida con el modelo *XLM-RoBERTa-ner-spanish*. Es decir, la transferencia de conocimiento del francés al español resulta en un rendimiento similar al de la transferencia de conocimiento desde un *dataset* de NER genérico a un *dataset* de NER específico.

Se evalúa también el modelo XLM-RoBERTa entrenado con las partes francesa y alemana de Hipe2020, para ver si el entrenamiento con más textos históricos ayuda a predecir mejor en el *dataset*, aunque sean en otro idioma. Los resultados resultan peores que los de entrenar solo con la parte francesa, luego descartamos esta hipótesis. Además, es coherente con lo que nos diría la intuición: entrenar con francés y alemán supuso una mejora para evaluar en el inglés, al ser el alemán un idioma germánico como el inglés. En cambio, para evaluar en el español, resulta más beneficioso entrenar solo con otro idioma latino como es el francés. Al entrenar con los tres idiomas (usando la parte de test del inglés para el entrenamiento), se obtienen resultados ligeramente mejores respecto a los anteriores, aunque sin mejorar la medida F1 respecto a entrenar solo con francés. Por ello, en adelante, utilizaremos tan solo el *dataset* francés para los próximos entrenamientos. Las tres últimas pruebas están reflejadas en las tres primeras filas de la Tabla 9.

	P	R	F1
XLM-RoBERTa-hipe-fr	0.43	0.54	0.48
XLM-RoBERTa-hipe-fr-de	0.44	0.49	0.46
XLM-RoBERTa-hipe-fr-de-en	0.46	0.51	0.48
RoBERTa-bne-hipe-fr	**0.47**	0.56	**0.51**
RoBERTa-bne-ner-capitel-hipe-fr	0.46	0.47	0.46
BERTin-hipe-fr	0.42	**0.6**	0.5

Tabla 9. Evaluación en CLARA-DM de modelos entrenados con Hipe2020.

Entrenando un modelo español, el modelo *RoBERTa-bne*, tan solo con el *dataset* francés, obtenemos una medida F1 de 0.51, una Precisión de 0.47 y un Recall de 0.56, que superan los resultados de los tres experimentos anteriores, es decir, que se consiguen mejores resultados con un modelo monolingüe español, que con un modelo multilingüe. Además, también han mejorado los resultados respecto a la aplicación directa del modelo genérico de NER *RoBERTa-bne-ner-capitel*, luego el entrenamiento con textos históricos demuestra ser de utilidad, cuando las lenguas son de la misma familia.

Hasta ahora, se ha evaluado el dataset CLARA-DM con el conjunto de etiquetas de Hipe2020. Al evaluar con el conjunto de etiquetas del dataset CAPITEL (con el que se entrenó el modelo *RoBERTa-bne*), se cuenta con las etiquetas de persona, lugar, organización, y "otros". La diferencia con el dataset Hipe2020 es la presencia de la etiqueta miscelánea, la ausencia de las etiquetas de tiempo y producción humana, y el formato de anotación (IOB2 vs. BIOES). Sorprendentemente, el modelo *RoBERTa-bne-ner-capitel* entrenado con Hipe2020 da peores resultados respecto al modelo sin ajuste fino para NER (*RoBERTa-bne*). Luego, en este caso, la adición del pre-entrenamiento para NER genérico no aporta un beneficio en nuestro dataset y es preferible entrenar solo con los textos históricos de Hipe2020 a pesar de no estar en español.

Por último, utilizamos el modelo español *BERTin-R*, que consigue el mejor *Recall*, de 0.6, y la segunda mejor medida F1, de 0.5. Los resultados se muestran en la Tabla 9.

En definitiva, en estos experimentos de aplicación directa de modelos en el dataset CLARA-DM, ha destacado el uso de modelos monolingües en español respecto al uso de modelos multilingües. Además, los mejores resultados se han obtenido al entrenar con el dataset francés de Hipe2020, sin utilizar pre-entrenamientos con *datasets* de NER genéricos. Ambos indicios son coherentes con la hipótesis planteada en este trabajo: hoy en día, la mejora de los sistemas de PLN es una

cuestión que recae en gran parte sobre los datos, más que sobre los algoritmos, puesto que estos son capaces de aprender correctamente cuando cuentan con los datos necesarios. En términos de tiempo y esfuerzos, los modelos basados en *Transformers* ahorran la tarea de tener que desarrollar algoritmos específicos para cada tarea, pero no evitan que sea necesario ajustarlos en dominios no generalistas. Por ello, aún es necesario desarrollar recursos específicos relativos tanto al idioma como a la tarea en cuestión.

4.3.2 *Few-shot* en CLARA-DM

En esta sección se usarán los documentos anotados de CLARA-DM para ajustar algunos modelos. Como hay tan solo con 5 documentos anotados, se usan 3 para el entrenamiento (698 frases), 1 para la validación (120 frases) y 1 para el test (111 frases).

A pesar de que en este *dataset* los mejores resultados se han obtenido con modelos monolingües en español, se realizan también algunas pruebas con el modelo multilingüe. Las pruebas consistirán en comparar el rendimiento de los modelos entrenados solo con los ejemplos de CLARA-DM, o entrenados, además, con la parte francesa de Hipe2020.

Para estos experimentos se aumenta el número de épocas desde 3 hasta 10, y el resto de los hiperparámetros se mantienen como antes (mencionados en 4.2.1). Al contar con tan pocos datos de entrenamiento, al aumentar el número de épocas se podría caer en sobreajuste. Sin embargo, el análisis de la evolución de la pérdida en los conjuntos de entrenamiento y validación no lo demuestra.

1. Fine-tuning en CLARA-DM

Al entrenar XLM-RoBERTa se obtiene un 0.46 en la medida F1, tres puntos por encima del modelo *XLM-RoBERTa-ner-spanish*, pero dos puntos por debajo de *RoBERTa-bne-ner-capitel*. Respecto a los modelos entrenados con Hipe2020, no consigue superar ningún resultado. Es decir, no se consigue mejora respecto a las pruebas *zero-shot*, tanto las hechas con modelos genéricos como los entrenados con Hipe2020. Al entrenar *RoBERTa-bne* se obtienen resultados similares. Por último, se entrena *BERTin-R*, que mejora en 6 puntos los resultados de los dos modelos anteriores, e incluso consigue mejorar en un punto el mejor resultado obtenido al entrenar con Hipe2020 (conseguido con *RoBERTa-bne*). Los resultados se muestran en la Tabla 10.

	P	R	F1
XLM-RoBERTa-clara	0.41	0.52	0.46
RoBERTa-bne-clara	0.42	0.50	0.46
BERTin-R-clara	**0.48**	**0.58**	**0.52**

Tabla 10. Entrenamiento y evaluación en CLARA-DM.

Los resultados de estas pruebas son bajos, debido a la escasez y baja calidad de datos disponibles. Lo que podemos concluir, es que destaca el uso del modelo español *BERTin-R* respecto a los otros dos, que además consigue superar al mejor modelo entrenado con Hipe2020 evaluado en CLARA-DM. Teniendo en cuenta que Hipe2020 tiene muchos más datos para el entrenamiento, esto es especialmente relevante puesto que estamos evaluando en el conjunto de etiquetas de CLARA-DM, y este resultado anima a seguir mejorando el corpus para no tener que evaluar en un conjunto de etiquetas distinto.

De cara a analizar más en profundidad el rendimiento de los modelos, la Tabla 11 muestra las métricas del modelo *BERTin-R* entrenado en CLARA-DM para cada etiqueta. El *soporte* indica el número de etiquetas presentes en el conjunto de test de cada clase. Además, en las tres últimas filas se muestran también las medidas micro (ya mostrada en la Tabla 10), macro y la media ponderada, calculadas con el informe de clasificación del paquete *seqeval*[97].

	P	R	F1	soporte
establec	0.23	0.31	0.26	29
loc	0.79	0.71	0.75	21
loc_cole	0.00	0.00	0.00	1
loc_direc	0.42	0.78	0.55	23
loc_relig	0.80	0.67	0.73	6
perdida_hallazgo	0.00	0.00	0.00	0
pers	0.53	1.00	0.70	8
pers_señores	0.56	0.64	0.60	14
prof	0.61	0.67	0.64	21
venta	0.50	0.08	0.14	12
Micro avg	0.48	0.58	0.52	135
Macro avg	0.44	0.49	0.44	135
Weighted avg	0.51	0.58	0.51	135

Tabla 11. Métricas de las etiquetas de CLARA-DM con BERTin-R.

[97] https://github.com/chakki-works/seqeval

Las etiquetas de localización, lugares religiosos y personas alcanzan resultados altos, por encima del 70% en la medida F1. Después van las de direcciones, señores y profesiones, con medidas F1 superiores al 50%. El rendimiento baja drásticamente en las etiquetas de establecimientos, colegios, pérdidas y hallazgos, y ventas, aunque en el caso de las pérdidas no había ninguna etiqueta en el conjunto de test, y colegios solo uno. En cambio, sí que hay una cantidad razonable de establecimientos y objetos en venta, que el modelo falla en reconocer. Estos resultados son coherentes con el estado actual del dataset y el grado de acuerdo que existe entre los anotadores. Las etiquetas de personas y lugares tienen un alto grado de acuerdo a la hora de la anotación, y además son entidades generales que el modelo es capaz de reconocer. En cambio, etiquetas como la de direcciones o profesiones, aún necesitan una mejor definición en la guía de anotación para mejorar el acuerdo. Por último, algunas etiquetas como la de ventas son muy específicas o poco generalizables (por tratarse de nombres comunes y no propios), por lo que en aún será necesario acotar sus límites, o incluso prescindir de ellas en el corpus.

2. Hipe2020 y CAPITEL en CLARA-DM

En esta última parte se intentarán mejorar los resultados obtenidos en el *dataset* CLARA-DM, combinando el entrenamiento con los *datasets* Hipe2020 y Capitel. Para cada una de las pruebas, se modificarán los nombres de las etiquetas de CLARA-DM para que coincidan con los del *dataset* del primer entrenamiento, tal y como se ha descrito anteriormente.

En primer lugar, se entrena con CLARA-DM el modelo *XLM-RoBERTa* entrenado anteriormente con la parte francesa de Hipe2020. A continuación, se realiza lo mismo con el modelo *RoBERTa-bne*. Después, el modelo *RoBERTa-bne-ner-capitel* se entrena primero solo con CLARA-DM, y a continuación se entrena primero con Hipe2020 (solo la parte francesa) y después con CLARA-DM. Por último, el modelo *BERTin-R* entrenado con Hipe2020 se entrena con CLARA-DM. Los resultados se muestran en la Tabla 12.

	P	R	F1
XLM-RoBERTa-hipe-fr-clara	**0.59**	0.64	**0.61**
RoBERTa-bne-hipe-fr-clara	0.53	0.61	0.57
RoBERTa-bne-ner-capitel-clara	0.54	0.59	0.57
RoBERTa-bne-ner-capitel-hipe-fr-clara	0.55	0.57	0.56
BERTin-R-hipe-fr-clara	0.54	**0.68**	0.6

Tabla 12. Entrenamiento y evaluación en CLARA-DM de modelos entrenados con Hipe2020 y CAPITEL.

Los mejores resultados se obtienen al entrenar con Hipe2020 y CLARA-DM. Esto es algo que cabe esperar, ya que Hipe2020 contiene textos históricos como CLARA-DM. En concreto, los mejores resultados en la medida F1 y Precisión los consigue el modelo *XLM-RoBERTa*, un 0.61 y 0.59 respectivamente, y en el Recall el modelo BERTin-R, con un 0.68 En general, estos resultados mejoran en torno a 10 puntos los resultados respecto a los obtenidos al entrenar solo con CLARA-DM, luego entrenar también con Hipe2020 resulta en una buena estrategia.

En definitiva, de estas pruebas podemos concluir que la inclusión de *datasets* del mismo dominio específico resulta beneficiosa, a pesar de que el idioma sea distinto. De hecho, resulta más beneficiosa incluso que la inclusión de *datasets* de la misma tarea, aunque estén en el mismo idioma.

4.4 Discusión

En este capítulo se han realizado experimentos para el reconocimiento de entidades con dos corpus de periódicos históricos, Hipe2020 y CLARA-DM.

Los primeros experimentos realizados tan solo con Hipe2020 han servido para obtener una imagen panorámica del funcionamiento de los modelos monolingües y multilingües basados en *Transformers*, y la adición o no de *datasets* genéricos para el reconocimiento de entidades. De estas pruebas concluimos que el uso de modelos multilingües es beneficioso cuando se cuenta con *datasets* en varios idiomas para una tarea específica, ya que permiten transferir este conocimiento a otros idiomas sin recursos. Además, también se ha intuido que la ampliación del entrenamiento con *datasets* genéricos no resulta tan beneficiosa como el entrenamiento con *datasets* de la misma tarea, aunque sean en otros idiomas, ya que además comprometen las etiquetas del *dataset* específico. No obstante, se han conseguido resultados que rondan el 80% en la medida F1 para el francés y el alemán, y el 65% en el inglés. Teniendo en cuenta la calidad del corpus, se trata de resultados prometedores.

Después se ha experimentado con el corpus CLARA-DM, con modelos monolingües y multilingües, y la adición de *datasets* genéricos para NER y específicos para NER en textos históricos. En los experimentos *zero-shot*, es decir, sin entrenamiento en CLARA-DM, han dado mejores resultados los modelos monolingües en español, tanto en los modelos entrenados con *datasets* de NER genéricos, como en los entrenados con Hipe2020. En cambio, en los experimentos *few-shot* donde se ha entrenado con los pocos datos disponibles de CLARA-DM,

tenemos dos escenarios. Por un lado, cuando se ha entrenado y evaluado con CLARA-DM, también ha resultado mejor un modelo monolingüe. Por otro lado, cuando se ha añadido también el entrenamiento con Hipe2020, se han obtenido resultados similares con modelos monolingües y multilingües (la diferencia es solo de un punto en la medida F1). Esta indiferencia en los resultados puede deberse a la similitud que existe entre el español y el francés. Lo que podemos afirmar, es que la transferencia de conocimiento ha funcionado en nuestro *dataset*, ya que el entrenamiento con Hipe2020 y CLARA-DM ha mejorado en 9 puntos la medida F1 respecto a entrenar solo con CLARA-DM (de 0.52 a 0.61).

Los resultados obtenidos son susceptibles de mejorar por varios motivos. El corpus CLARA-DM usado se trata de una versión preliminar con pocos datos, que aún no cuentan con la homogeneidad necesaria. A la hora de fusionar las anotaciones, aún no se contaba con el suficiente acuerdo entre los anotadores. Además, el conjunto de etiquetas es muy amplio y es probable que se reduzca en el futuro, lo cual facilitará la tarea de predicción. Así que, los objetivos del trabajo mostrado en esta memoria a corto plazo consisten en mejorar la calidad de las anotaciones, y posteriormente la cantidad, para contar con un volumen suficiente de datos de entrenamiento.

Conclusiones y trabajo futuro

En este trabajo se ha abordado un caso de estudio de aplicación de las tecnologías de PLN al ámbito de las Humanidades Digitales, en concreto el problema del reconocimiento de entidades en textos históricos. Durante la fase de construcción del corpus ha sido notable el elevado coste en términos de tiempo y esfuerzo manual que suponen la recolección y almacenamiento de los periódicos, el proceso de reconocimiento del *layout* (imposible de automatizar por ahora), y la transcripción. Además, las caractacterísticas de los periódicos relativas a la lengua antigua y la calidad de las digitalizaciones, han puesto en evidencia la necesidad de desarrollar un corpus y un modelo específicos para el reconocimiento de entidades en estos textos. Después, durante la fase de anotación del corpus, se han podido comprobar las dificultades que propone la tarea, siendo necesarias una buena coordinación y metodología para llevarla a cabo y poder diseñar una guía de anotación eficiente. Con los documentos anotados disponibles a septiembre de 2022 se han realizado experimentos con modelos actuales de aprendizaje profundo tanto monolingües como multilingües, haciendo uso además de otros datasets generales y específicos. Los experimentos en el dataset de textos históricos Hipe2020 demuestran la utilidad de usar modelos multilingües para la transferencia de conocimiento entre idiomas, y de la inclusión de *datasets* de NER genéricos para el entrenamiento de los modelos. En cambio, en el *dataset* CLARA-DM destaca el uso de modelos monolingües, revelándose además útil el entrenamiento con *datasets* del mismo dominio específico en una lengua de la misma familia, en este caso el francés.

Los resultados son prometedores dada la escasez y mala calidad de los datos con los que se contaba. El proyecto CLARA-HD comienza su segundo año de un total de tres, y se prevé obtener un dataset más grande y consistente próximamente. Así, como trabajo futuro se plantea el desarrollo de un modelo robusto para el reconocimiento de entidades en el corpus CLARA-DM. Para ello, se mejorará la fase de entrenamiento de los modelos mediante una selección de hiperparámetros más refinada, y un posterior análisis de errores más detallado. Una vez se cuente con un modelo robusto para el reconocimiento de entidades en el corpus CLARA-DM, sería interesante realizar un análisis comparativo con respecto al reconocimiento de entidades en dominios generales, que permita estudiar en profundidad el impacto que tienen los obstáculos descritos en este trabajo (lengua

antigua, calidad de las digitalizaciones y transcripciones automáticas, inconsistencias en los *datasets*, etc).

Es difícil reducir a una sola tarea de PLN la complejidad de las tareas que se realizan en investigación: en el caso de los historiadores del arte, probablemente no sea suficiente detectar las entidades de los textos para analizar la diversidad de la vida cotidiana de hace doscientos años. Siguiendo la línea de este trabajo, se podrían aplicar otras tecnologías de PLN a esta temática que sin duda serían de utilidad para las investigaciones: extracción de resúmenes y modelado de temáticas para captar el tema que tratan los textos de un vistazo, resolución de correferencias y extracción de relaciones, todo ello como engranaje de un sistema de respuesta de preguntas... Los avances de las tecnologías nos permiten imaginar numerosas soluciones. Sin embargo, aún estamos lejos de poder abordar las tareas con el grado de detalle del que somos capaces los humanos, y menos aun cuando se trata con dominios con pocos recursos. Por eso es necesario dedicar esfuerzos a mejorar los recursos existentes, hacerlos más adaptables, y extender su dominio de aplicación para que no se estanquen en el ámbito del desarrollo, sino que encuentren una aplicación que facilite la vida de las personas.

Publicaciones de la autora del trabajo

Menta, A., Sánchez-Salido, E., & García-Serrano, A. (2022). Transcripción de periódicos históricos: Aproximación CLARA-HD. *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations (SEPLN-PD 2022).*

García-Serrano, A., Menta, A., & Sánchez-Salido, E. (2022). Digital Humanities and Text Simplification Tasks: The CLARA-HD Project. *III WS Intele – sesión de pósteres.* En línea: http://ixa2.si.ehu.eus/intele/sites/default/files/posterrak/agarcia_posterIntele_vf.pdf

Referencias

Academia Mexicana de la Lengua. (2015). *Corpus Diacrónico y Diatópico del Español de América (CORDIAM)*. En línea: <www.cordiam.org>

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54-59. https://doi.org/10.18653/v1/N19-4010

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. *Proceedings of the 27th International Conference on Computational Linguistics*, 1638-1649. https://aclanthology.org/C18-1139

Alarcón, E. (2002). *El proyecto Corpus Thomisticum: Descripción y perspectivas*. 11.

Aldama, N., Guerrero, M., Montoro, H., & Samy, D. (2022). *Anotación de corpus lingüísticos: Metodología utilizada en el Instituto de Ingeniería del Conocimiento (IIC)*. 17.

Aranda García, N. (2022). Humanidades Digitales y literatura medieval española: La integración de Transkribus en la base de datos COMEDIC. *Historias Fingidas*, *0*, 127-149. https://doi.org/10.13136/2284-2667/1107

Asahara, M., & Matsumoto, Y. (2003). Japanese Named Entity Extraction with Redundant Morphological Analysis. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 8-15. https://aclanthology.org/N03-1002

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a web of open data. *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, 722-735.

Ayuso García, M. (2022). Las ediciones de Arnao Guillén de Brocar de BECLaR transcritas con ayuda de Transkribus y OCR4all: Creación de un modelo para la red neuronal y posible explotación de los resultados. *Historias Fingidas*, *0*, 151-173. https://doi.org/10.13136/2284-2667/1102

Baptiste, B., Favre, B., Auguste, J., & Henriot, C. (2021, diciembre). Transferring Modern Named Entity Recognition to the Historical Domain: How to Take the Step? *Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*. https://hal.archives-ouvertes.fr/hal-03550384

Bazzaco, S., Ruiz, A. M. J., Ruberte, Á. T., & Molares, M. M. (2022). Sistemas de reconocimiento de textos e impresos hispánicos de la Edad Moderna. La creación de unos modelos de HTR para la transcripción automatizada de documentos en gótica y redonda (s. XV-XVII). *Historias Fingidas*, *0*, 67-125. https://doi.org/10.13136/2284-2667/1190

Biber, D. (1990). Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. *Literary and Linguistic Computing*, *5*(4), 257-269. https://doi.org/10.1093/llc/5.4.257

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*(4), 243-257. https://doi.org/10.1093/llc/8.4.243

Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1997). Nymble: A High-Performance Learning Name-finder. *Fifth Conference on Applied Natural Language Processing*, 194-201. https://doi.org/10.3115/974557.974586

Bird, S., & Liberman, M. (1998). *TOWARDS A FORMAL FRAMEWORK FOR LINGUISTIC ANNOTATIONS*. 12.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993-1022.

BNC Consortium. (2007). *The British National Corpus*. XML Edition, Oxford Text Archive. http://www.natcorp.ox.ac.uk

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching Word Vectors with Subword Information* (arXiv:1607.04606). arXiv. http://arxiv.org/abs/1607.04606

Bollmann, M. (2019). A Large-Scale Comparison of Historical Text Normalization Systems. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3885-3898. https://doi.org/10.18653/v1/N19-1389

Borja, P. S.-P. (2018). El corpus ALDICAM-CM Geografía lingüística diacrónica de la Comunidad de Madrid. *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos*, *5*(1), 69-75. https://doi.org/10.15366/chimera2018.5.1.004

Boros, E., Hamdi, A., Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Sidere, N., & Doucet, A. (2020). Alleviating Digitization Errors in Named Entity Recognition for Historical Documents. *Proceedings of the 24th Conference on Computational Natural Language Learning*, 431-441. https://doi.org/10.18653/v1/2020.conll-1.35

Boros, E., Pontes, E. L., Cabrera-Diego, L. A., Hamdi, A., Moreno, J. G., Sidère, N., & Doucet, A. (2020). *Robust Named Entity Recognition and Linking on Historical Multilingual Documents*. 17.

Borthwick, A., Sterling, J., Agichtein, E., & Grishman, R. (1998). NYU: Description of the MENE Named Entity System as Used in MUC-7. *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. MUC 1998. https://aclanthology.org/M98-1018

Bravo-Garcia, E., Pons Rodriguez, L., Garrido Martín, B., & Octavio, A. (2013, enero 1). *Preliminares para la construcción de un corpus discursivo diacrónico del español: Las quejas en su historia*.

Brouwer, M., Brugman, H., & Kemps-Snijders, M. (2016). *MTAS: A Solr/Lucene based Multi Tier Annotation Search solution*. 19.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Calderón Campos, M. (2019). La edición de corpus históricos en la plataforma TEITOK. El caso de «Oralia diacrónica del español». *Chimera: Romance Corpora and Linguistic Studies*, *6*, 21-36.

Calderón Campos, M., & Vaamonde, G. (2020). «Oralia diacrónica del español»: Un nuevo corpus de la Edad Moderna. *Scriptum digital. Revista de corpus diacrònics i edició digital en Llengües iberoromàniques*, *9*, 23.

Calvo Tello, J. (2019). Diseño de corpus literario para análisis cuantitativos. *Revista de Humanidades Digitales*, *4*, 115-135. https://doi.org/10.5944/rhd.vol.4.2019.25187

Calzada Pérez, M. (2006). Corpus Comparables y Paralelos de Discursos Europeos (ECPC). *VII Congreso de Lingüística General*.

Campillos-Llanos, L., Terroba Reinares, A. R., Zakhir Puig, S., Valverde-Mateos, A., & Capllonch-Carrí´on, A. (2022). Building a comparable corpus and a benchmark for Spanish medical text simplification. *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations (SEPLN-PD 2022)*.

Campos, M. C., & Godoy, M. T. G. (2009). El Corpus diacrónico del español del Reino de Granada (CORDEREGRA). En *Diacronía de las lenguas iberorrománicas* (pp. 229-250). Vervuert Verlagsgesellschaft. https://doi.org/10.31819/9783865278685-014

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Perez, J. (2020). Spanish pre-trained BERT model and evaluation data. *PML4DC, ICLR*.

Carreras, X., Màrquez, L., & Padró, L. (2002). Named Entity Extraction using AdaBoost. *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. https://aclanthology.org/W02-2004

Chiu, J. P. C., & Nichols, E. (2016). Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, *4*, 357-370. https://doi.org/10.1162/tacl_a_00104

Cho, H., & Lee, H. (2019). Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics*, *20*(1), 735. https://doi.org/10.1186/s12859-019-3321-4

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th international conference on Machine learning*, 160-167. https://doi.org/10.1145/1390156.1390177

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *NATURAL LANGUAGE PROCESSING*, 45.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale* (arXiv:1911.02116). arXiv. http://arxiv.org/abs/1911.02116

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: An Architecture for Development of Robust HLT applications. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 168-175. https://doi.org/10.3115/1073083.1073112

Davies, M. (2002). *Un corpus anotado de 100.000.000 palabras del español histórico y moderno*. http://www.corpusdelespanol.org

Davies, M., & Parodi, G. (2022). Constitución de corpus crecientes del español. En G. Parodi, P. Cantos-Gómez, C. Howe, M. Lacorte, J. Muñoz-Basol, & J. Muñoz-Basol, *Lingüística de corpus en español* (1.ª ed., pp. 13-32). Routledge. https://doi.org/10.4324/9780429329296-3

de la Rosa, J., Ponferrada, E. G., Villegas, P., Salas, P. G. de P., Romero, M., & Grandury, M. (2022). *BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling* (arXiv:2207.06814). arXiv. http://arxiv.org/abs/2207.06814

De Toni, F., Akiki, C., De La Rosa, J., Fourrier, C., Manjavacas, E., Schweter, S., & Van Strien, D. (2022). Entities, Dates, and Languages: Zero-Shot on Historical Texts with T0. *Proceedings of BigScience Episode #5 -- Workshop on Challenges & Perspectives in Creating Large Language Models*, 75-83. https://doi.org/10.18653/v1/2022.bigscience-1.7

Delestre, C., & Amar, A. (2022). *DistilCamemBERT: A distillation of the French model CamemBERT* (arXiv:2205.11111). arXiv. https://doi.org/10.48550/arXiv.2205.11111

Dernoncourt, F., Lee, J. Y., & Szolovits, P. (2017). NeuroNER: An easy-to-use program for named-entity recognition based on neural networks. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 97-102. https://doi.org/10.18653/v1/D17-2017

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2021). *Named Entity Recognition and Classification on Historical Documents: A Survey* (arXiv:2109.11406). arXiv. http://arxiv.org/abs/2109.11406

Ehrmann, M., Romanello, M., Clematide, S., Ströbel, P. B., & Barman, R. (2020). Language Resources for Historical Newspapers: The Impresso Collection. *Proceedings of the 12th Language Resources and Evaluation Conference*, 958-968. https://aclanthology.org/2020.lrec-1.121

Ehrmann, M., Romanello, M., Doucet, A., & Clematide, S. (2022). *HIPE-2022 Shared Task Named Entity Datasets* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6089968

Ehrmann, M., Romanello, M., Fluckiger, A., & Clematide, S. (2020). *Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers*. 38.

Ehrmann, M., Romanello, M., Najem-Meyer, S., Doucet, A., & Clematide, S. (2022). *Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents*. 26.

Ehrmann, M., Watter, C., Romanello, M., & Clematide, S. (2020). *Impresso Named Entity Annotation Guidelines*. https://doi.org/10.5281/zenodo.3604227

Enrique Arias, A. (2008). *Biblia Medieval*. En línea: <http://www.bibliamedieval.es>

Fernández-Ordóñez, I., & Pato, E. (2020). El COSER (Corpus oral y sonoro del Español Rural) y su contribución al estudio de la variación gramatical. *Ángel Gallego y Francesc Roca*.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 363-370. https://doi.org/10.3115/1219840.1219885

García Moreno, A., & Pueyo Mena, Fco. J. (2013). Corpus Histórico Judeoespañol (CORHIJE). *CSIC*. En línea: http://esefardic.es/corhije

García-Serrano, A., Menta, A., & Sánchez-Salido, E. (2022). Digital Humanities and Text Simplification Tasks: The CLARA-HD Project. *III WS Intele – sesión de pósteres*. En línea: http://ixa2.si.ehu.eus/intele/sites/default/files/posterrak/agarcia_posterIntele_vf.pdf

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92. https://doi.org/10.1145/3458723

Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6: A brief history. *Proceedings of the 16th conference on Computational linguistics - Volume 1*, 466-471. https://doi.org/10.3115/992628.992709

Grupo de Investigación Textos para la Historia del Español [GITHE]. (2017). *Corpus de documentos españoles anteriores a 1800 (CODEA+ 2015)*. https://doi.org/10.37536/CODEA.2015

Gruszczyński, W., Adamiec, D., Bronikowska, R., Kieraś, W., Modrzejewski, E., Wieczorek, A., & Woliński, M. (2021). The Electronic Corpus of 17th- and 18th-century Polish Texts. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-021-09549-1

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Armentano-Oller, C., Rodriguez-Penagos, C., Gonzalez-Agirre, A., & Villegas, M. (2022). *MarIA: Spanish Language Models*. 22.

Gutiérrez-Fandiño, A., Pérez-Fernández, D., Armengol-Estapé, J., Griol, D., & Callejas, Z. (2022). *esCorpius: A Massive Spanish Crawling Corpus* (arXiv:2206.15147). arXiv. https://doi.org/10.48550/arXiv.2206.15147

Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., & Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, *33*(14), i37-i48. https://doi.org/10.1093/bioinformatics/btx228

Haibo He, & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., & Varga, D. (2014). *DCEP -Digital Corpus of the European Parliament*. 8.

Hammerton, J. (2003). Named Entity Recognition with Long Short-Term Memory. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 172-175. https://aclanthology.org/W03-0426

Hinrichs, E., & Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1525-1531. http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). OntoNotes: The 90% solution. *Proceedings of the Human Language*

Technology Conference of the NAACL, Companion Volume: Short Papers, 57-60.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). *XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization* (arXiv:2003.11080). arXiv. http://arxiv.org/abs/2003.11080

Huang, Z., Xu, W., & Yu, K. (2015). *Bidirectional LSTM-CRF Models for Sequence Tagging* (arXiv:1508.01991; Versión 1). arXiv. http://arxiv.org/abs/1508.01991

Janssen, M. (2016, mayo 1). *TEITOK: Text-Faithful Annotated Corpora*.

Kabatek, J. (2013). ¿Es posible una lingüística histórica basada en un corpus representativo? *Iberoromania*, *77*(1). https://doi.org/10.1515/ibero-2013-0045

Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2015). *Character-Aware Neural Language Models* (arXiv:1508.06615). arXiv. https://doi.org/10.48550/arXiv.1508.06615

Kučera, H., & Francis, W. N. (1979). *Brown Corpus.* Department of Linguistics, Brown University, Providence, Rhode Island, US.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). *Neural Architectures for Named Entity Recognition* (arXiv:1603.01360). arXiv. https://doi.org/10.48550/arXiv.1603.01360

Lample, G., & Conneau, A. (2019). *Cross-lingual Language Model Pretraining* (arXiv:1901.07291). arXiv. https://doi.org/10.48550/arXiv.1901.07291

Leech, G. (1991). *The state of the art in corpus linguistics* (Routledge). https://www.taylorfrancis.com/chapters/edit/10.4324/9781315845890-11/state-art-corpus-linguistics-geoffrey-leech

Leech, G., Johansson, S., & Hofland, K. (1978). The LOB Corpus. *Lancaster University, University of Oslo, University of Bergen*.

Li, J., Sun, A., Han, J., & Li, C. (2022). A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50-70. https://doi.org/10.1109/TKDE.2020.2981314

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Liu, Z., Xu, Y., Yu, T., Dai, W., Ji, Z., Cahyawijaya, S., Madotto, A., & Fung, P. (2021). CrossNER: Evaluating Cross-Domain Named Entity Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(15), 13452-13460.

Ma, X., & Hovy, E. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1064-1074. https://doi.org/10.18653/v1/P16-1101

Martinez-Romo, J., Araujo, L., Reneses, B., Seara-Aguilar, G., & Martínez-Capella, I. (2022). Detección de Indicios de Autolesiones No Suicidas en Informes Médicos de Psiquiatría Mediante el Análisis del Lenguaje. *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations (SEPLN-PD 2022).*

McCallum, A., & Li, W. (2003). Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 188-191. https://aclanthology.org/W03-0430

Menta, A., Sánchez-Salido, E., & García-Serrano, A. (2022). Transcripción de periódicos históricos: Aproximación CLARA-HD. *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations (SEPLN-PD 2022).*

Meroño-Peñuela, A., Boer, V., Erp, M., Melder, W., Mourits, R., Rijpma, A., Schalk, R., & Zijdeman, R. (2020). *Ontologies in CLARIAH: Towards Interoperability in History, Language and Media.*

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. http://arxiv.org/abs/1301.3781

Miranda-Escalada, A., Farré-Maduell, E., González Gacio, G., & Krallinger, M. (2022). *LivingNER corpus: Named entity recognition, normalization & classification of species, pathogens and food.* https://doi.org/10.5281/zenodo.6572503

Montani, I., & Honnibal, M. (2018). Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence.*

Moreno Sandoval, A. (2019). *Lenguas y computación*. Síntesis.

Moreno Sandoval, A., Díaz García, J., Campillos Llanos, L., & Redondo, T. (2018). *Biomedical Term Extraction: NLP Techniques in Computational Medicine*. https://doi.org/10.9781/ijimai.2018.04.001

Nadeau, D., & Sekine, S. (2007). A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, *30*. https://doi.org/10.1075/li.30.1.03nad

Nadeau, D., Turney, P. D., & Matwin, S. (2006). Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity. En L. Lamontagne & M. Marchand (Eds.), *Advances in Artificial Intelligence* (pp. 266-277). Springer. https://doi.org/10.1007/11766247_23

Nakayama, E. (2021). *Implementación de un corpus comparable de español y japonés de acceso abierto para la traducción especializada*. 29.

Navigli, R., Bevilacqua, M., Conia, S., Montagnini, D., & Cecconi, F. (2021). Ten Years of BabelNet: A Survey. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 4559-4567. https://doi.org/10.24963/ijcai.2021/620

Neudecker, C. (2016). An Open Corpus for Named Entity Recognition in Historic Newspapers. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4348-4352. https://aclanthology.org/L16-1689

Neudecker, C., Wilms, L., Faber, W. J., & van Veen, T. (2014). *Large-scale refinement of digital historic newspapers with named entity recognition*. 15.

Nieuwenhuijsen, D. (2016). Notas sobre la aportación del análisis estadístico a la lingüística de corpus. En *Notas sobre la aportación del análisis estadístico a la lingüística de corpus* (pp. 215-237). De Gruyter. https://doi.org/10.1515/9783110462357-011

Nothman, J., Ringland, N., Radford, W., Murphy, T., & Curran, J. R. (2013). Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, *194*, 151-175. https://doi.org/10.1016/j.artint.2012.03.006

Ortiz-Zambrano, J., Espin-Riofrio, C., & Montejo-Ráez, A. (2022). *Transformers for Lexical Complexity Prediction in Spanish Language*.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), 1532-1543. https://doi.org/10.3115/v1/D14-1162

Pichel Gotérrez, R., Alonso Parada, R., Cabana Outeiro, A., Couceiro Pérez, X. L., Dono López, P., Dourado Fernández, R., Martínez Lema, P., Mariño Paz, R., & Varela Barreiro, F. X. (2016). *Corpus de Textos Antiguos de Galicia (COTAGAL).* Instituto da Lingua Galega. https://ilg.usc.es/es/proxectos/corpus-de-textos-antiguos-de-galicia-cotagal

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. Graeme Hirst, University of Toronto.

Provatorova, V., Vakulenko, S., Kanoulas, E., & van Hulst, J. M. (2020). *Named Entity Recognition and Linking on Historical Newspapers: UvA.ILPS & REL at.* 8.

Rao, D., McNamee, P., & Dredze, M. (2013). Entity Linking: Finding Extracted Entities in a Knowledge Base. En T. Poibeau, H. Saggion, J. Piskorski, & R. Yangarber (Eds.), *Multi-source, Multilingual Information Extraction and Summarization* (pp. 93-115). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28569-1_5

Real Academia Española: Banco de datos. (2008). *Corpus diacrónico del español (CORDE).* En línea: <https://www.rae.es/banco-de-datos/corde>

Real Academia Española: Banco de datos. (2009). *Corpus del Diccionario Histórico de la Lengua Española (CDH).* En línea: <https://www.rae.es/banco-de-datos/cdh>

Real Academia Española: Banco de datos. (2021). *Corpus del Español del Siglo XXI (CORPES XXI).* En línea: <https://www.rae.es/banco-de-datos/corpes-xxi>

Red CHARTA. (2011a). *CHARTA (Corpus Hispánico y Americano en la Red: Textos Antiguos).* En línea: <www.corpuscharta.es>

Red CHARTA. (2011b). *Criterios de edición de documentos hispánicos (Orígenes – siglo XIX).* 39.

Rodríguez, J. R. M. (2014). *El CorLexIn, un corpus para el estudio del léxico histórico y dialectal del Siglo de Oro. 3*, 24.

Rojo, G. (2010). Sobre codificación y explotación de corpus textuales: Otra comparación del Corpus del español con el CORDE y el CREA. *Lingüística, 24*, 11-50.

Rojo, G. (2016). *Los corpus textuales del español*.

Rosset, S., Grouin, C., & Zweigenbaum, P. (2011). *Entités nommées structurées: Guide d'annotation Quaero*. http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf

Sánchez González de Herrero, M. N., Martín Aizpuru, L., Sánchez Romo, R., & Marcet Rodríguez, V. J. (2010). *Corpus de documentación de cancillería real castellana, siglo XIII y primera década del s. XIV (CODCAR) en el Corpus CHARTA*. En línea: <http://www.corpuscharta.es/>

Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V., & Boeker, M. (2020). *GottBERT: A pure German Language Model* (arXiv:2012.02110). arXiv. https://doi.org/10.48550/arXiv.2012.02110

Sekine, S. (1998). Description of the Japanese NE System Used for MET-2. *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. MUC 1998. https://aclanthology.org/M98-1019

Seo, C., Lee, S.-W., & Kim, H.-J. (2003). An efficient inverted index technique for XML documents using RDBMS. *Information and Software Technology*, *45*(1), 11-22. https://doi.org/10.1016/S0950-5849(02)00157-X

Serrano, A. V., Subies, G. G., Zamorano, H. M., Garcia, N. A., Samy, D., Sanchez, D. B., Sandoval, A. M., Nieto, M. G., & Jimenez, A. B. (2022). *RigoBERTa: A State-of-the-Art Language Model For Spanish* (arXiv:2205.10233; Versión 2). arXiv. http://arxiv.org/abs/2205.10233

Sullivan, D. (2013). Tech Services on the Web: CATMA: Computer Aided Textual Markup & Analysis; http://www.catma.de/. *Technical Services Quarterly*, *30*(3), 337-338. https://doi.org/10.1080/07317131.2013.788370

Szarvas, G., Vincze, V., Farkas, R., & Csirik, J. (2008). The BioScope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing - BioNLP '08*, 38. https://doi.org/10.3115/1572306.1572314

Terras, M. M. (2011). The Rise of Digitization. En R. Rikowski (Ed.), *Digitisation Perspectives* (pp. 3-20). SensePublishers. https://doi.org/10.1007/978-94-6091-299-3_1

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002).* https://aclanthology.org/W02-2024

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142-147. https://aclanthology.org/W03-0419

Torrens Álvarez, M. J. (2016). *Corpus Histórico del Español Norteño (CORHEN).* En línea: <corhen.es>

Torruella Casañas, J. (2017). *Lingüística de corpus: Génesis y bases metodológicas de los corpus (históricos) para la investigación en lingüística*. Peter Lang.

Torruella Casañas, J., Pérez Saldanya, M., & Martines, J. (2009). *Corpus informatitzat del català antic (CICA).* En línea: http://www.cica.cat

Torruella, J., & Kabatek, J. (2021). *Portal de corpus históricos hiberorrománicos (CORHIBER). 10*, 11.

Universidad de Alcalá. (2014). PRESEEA: Corpus del Proyecto para el estudio sociolingüístico del español de España y de América. *Alcalá de Henares*. [http://preseea.uah.es]

Universidade de Lisboa. (2014). *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna.* http://ps.clul.ul.pt

University of Birmingham. (1991). *Bank of English*.

Vaamonde, G. (2015). *Limitaciones en el uso de corpus diacrónicos del español. Nuevas aportaciones desde el proyecto de investigación Post Scriptum.* 10.

van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., & Van de Walle, R. (2015). Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities*, *30*(2), 262-279. https://doi.org/10.1093/llc/fqt067

van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B., & Colavizza, G. (2020). Assessing the Impact of OCR Quality on Downstream NLP Tasks: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, 484-496. https://doi.org/10.5220/0009169004840496

Varela, X., Doval Iglesias, L., & Osorio Peláez, C. (2018). *Tesouro Medieval Informatizado da Lingua Galega (TMILG).* En línea: http://ilg.usc.es/tmilg

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Verdejo Maillo, M. F. (1996). *EuroWordNet: Building a Multilingual WordNet Database With Semantic Relations between Words.*

Vilain, M., Su, J., & Lubar, S. (2007). Entity extraction is a boring solved problem: Or is it? *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, 181-184.

Weischedel, Ralph, Palmer, Martha, Marcus, Mitchell, Hovy, Eduard, Pradhan, Sameer, Ramshaw, Lance, Xue, Nianwen, Taylor, Ann, Kaufman, Jeff, Franchini, Michelle, El-Bachouti, Mohammed, Belvin, Robert, & Houston, Ann. (2013). *OntoNotes Release 5.0* (p. 2806280 KB) [Data set]. Linguistic Data Consortium. https://doi.org/10.35111/XMHB-2B84

Weiss, Sholom M., Indurkhya, Nitin, Zhang, Tong, & Damerau, Fred J. (2005). *Text Mining: Predictive Methods for Analyzing Unstructured Information.* Springer. https://link.springer.com/book/10.1007/978-0-387-34555-0

Wissler, L., Almashraee, M., Monett, D., & Paschke, A. (2014, junio 26). *The Gold Standard in Corpus Annotation.* https://doi.org/10.13140/2.1.4316.3523

Yadav, V., & Bethard, S. (2018). A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. *Proceedings of the 27th International Conference on Computational Linguistics*, 2145-2158. https://aclanthology.org/C18-1182

Yadav, V., & Bethard, S. (2019). *A Survey on Recent Advances in Named Entity Recognition from Deep Learning models* (arXiv:1910.11470). arXiv. http://arxiv.org/abs/1910.11470

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., & Artzi, Y. (2021). Revisiting Few-Sample BERT Fine-Tuning. *ICLR 2021*, 22.

Zhang, X., Yu, B., Wang, Y., Liu, T., Su, T., & Xu, H. (2022). Exploring Modular Task Decomposition in Cross-domain Named Entity Recognition. *Proceedings of the 45th International ACM SIGIR Conference on*

Research and Development in Information Retrieval, 301-311. https://doi.org/10.1145/3477495.3531976

Apéndice I. Corpus, hemerotecas y colecciones

Biblia Medieval

Recurso de libre acceso en la red destinado a facilitar el estudio y la difusión de un aspecto singular de la lengua y cultura medievales hispánicas: las traducciones de la Biblia al castellano llevadas a cabo durante la Edad Media. El corpus permite consultar en paralelo transcripciones paleográficas de los manuscritos que han transmitido los romanceamientos medievales existentes junto a sus fuentes latinas o hebreas, con posibilidad de consulta de imágenes digitales de los códices originales.

Brown Corpus

El *Brown University Standard Corpus of Present-Day American English* (o simplemente *Brown Corpus*) es una colección electrónica de muestras de texto del inglés estadounidense, el primer corpus estructurado importante de géneros variados. Este corpus estableció por primera vez el listón para el estudio científico de la frecuencia y distribución de las categorías de palabras en el uso cotidiano del lenguaje. Compilado por Henry Kučera y W. Nelson Francis en la Universidad de Brown, en Rhode Island, es un corpus de lenguaje general que contiene 500 muestras de de unas 2.000 palabras cada una, con un total de aproximadamente un millón de palabras, compilado a partir de trabajos publicados en los Estados Unidos en 1961.

BNC: British National Corpus

El BNC es una colección de 100 millones de palabras con muestras de la lengua escrita y hablada recogidas de una variedad de fuentes, diseñada para representar un amplio espectro del inglés británico de finales del siglo XX. La última edición es la *BNC XML Edition*, lanzada en 2007.

BOE: Bank of English

El *Bank of English* es un subconjunto representativo del corpus COBUILD, una colección de textos en inglés con un total de 4.500 millones de palabras. Los textos son principalmente de origen británico, pero también se incluye contenido de América del Norte, Australia, Nueva Zelanda, Sudáfrica y otros países de la Commonwealth. La mayoría de los textos son del inglés escrito, recopilados de

sitios web, periódicos, revistas y libros. También hay un gran componente de datos hablados utilizando material de la radio, la televisión y conversaciones informales. El Bank of English cuenta en total con 650 millones de palabras y, lanzado en 1991, fue diseñado originalmente como un corpus monitor (en continua actualización).

CdE: Corpus del Español

El Corpus del Español de Mark Davies es un corpus de referencia dividido en bloques. El original es el *Género/Histórico* (CdEGH), lanzado en 2001, que contiene textos orales, ficción, prensa y académicos desde el siglo XIII hasta el siglo XX y cuenta con más de 100 millones de palabras y más de 20.000 textos, siendo uno de los corpus históricos más grandes. En concreto, respecto a los siglos XVI-XVIII contiene:

- 323 textos del 1500 con un total de 17,774,762 de palabras,
- 498 textos del 1600 con un total de 13,355,483 de palabras y
- 176 textos del 1700 con un total de 10,324,328 palabras.

Para realizar consultas en el CdEGH es necesario registrarse. En el cuadro de búsqueda se muestran los tipos de consulta posibles (Lista, Gráfico, Colocados, Comparar y Palabras Clave en Contexto, además de poder filtrar por categoría gramatical), y a la derecha su explicación.

Una versión del corpus más actual y mucho más grande es el *Web/Dialectos*, que contiene casi 2.000 millones de palabras e incluye textos extraídos de páginas web de 21 países hispanohablantes en los últimos años. Permite estudiar las variaciones dialectales del español moderno y es descargable. Estos dos subcorpus del Corpus del Español contienen los datos de los que hace uso la herramienta *WordAndRephrase* (que a su vez forma parte del CdE), que permite buscar entre las 40.000 palabras con mayor frecuencia en estos corpus mostrando además las palabras cercanas más comunes, sinónimos y otras características.

El *NOW* es la adición más reciente y contiene más de 7,3 mil millones de palabras de 21 diferentes países de habla hispana, desde 2012 hasta 2019. El *Google Books n-grams* es su interfaz para los datos de *n*-gramas de Google Books y se basa en 45 mil millones de palabras en decenas de millones de libros del 1800 al 2000.

CDH: Corpus del Diccionario Histórico de la Lengua Española

Este es el corpus del actual *Diccionario Histórico de la Lengua Española* de la RAE. Cuenta con 355.740.238 registros que se distribuyen en tres capas de consulta:

- CDH *nuclear* con más de 53 millones de ocurrencias, de las cuales 32 pertenecen a textos españoles y más de 20 millones a obras americanas. Los textos que conforman el corpus (en buena medida, comunes al CORDE y al CREA) se han sometido a un proceso semiautomático de anotación lingüística.
- Desde el siglo XII hasta 1975, formado por una selección de obras procedentes del CORDE con un total de 199.387.676 formas. Estas obras poseen una preanotación morfosintáctica, realizada con herramientas de software libre (Freeling) en el marco del proyecto del Diccionario histórico de la lengua española.
- Desde 1975 hasta el 2000, con títulos procedentes del CREA, anotados lingüísticamente por el Departamento de Tecnología de la Real Academia Española. Cuenta con 103.173.014 registros.

El corpus dispone de una interfaz de consulta que permite filtrar estas capas y realizar distintos tipos de consulta por lema, forma, categoría gramatical y otras características.

CHARTA: Corpus Hispánico y Americano en la Red: Textos Antiguos

La red CHARTA[98] se concibe como un proyecto global para la edición, análisis lingüístico y publicación de documentos antiguos de los siglos XII al XIX de España e Hispanoamérica con un sistema de presentación triple riguroso (paleográfica, crítica y facsimilar) con el fin de satisfacer distintas necesidades de investigadores y usuarios en general.

El corpus CHARTA está compuesto por varios subcorpus, entre los que se encuentran el CODEA, el DOLEO o el *Corpus Histórico del Español Norteño* (CORHEN) (Torrens Álvarez, 2016).

La consulta del corpus se realiza sobre una primera fase que consta de 2076 documentos seleccionados de acuerdo con tres parámetros: geográfico, cronológico y tipológico. Es posible filtrar por siglo y provincia además de otros parámetros. Por ejemplo, si filtramos el siglo XVI en la provincia de Madrid obtenemos 12 textos del CODEA, en el XVII obtenemos 2 de CODEA y 2 de DOLEO, y ningún

[98] https://www.redcharta.es/

texto del XVIII. Se trata de un corpus desequilibrado; en https://www.corpuscharta.es/estadisticas.php se pueden consultar las estadísticas de la distribución de los documentos según el año, el lugar, la tipología y los subgrupos que lo conforman.

CICA: Corpus Informatitzat del Català Antic

El Corpus Informatizado del Catalán Antiguo reúne una colección de textos desde el siglo XI (primeros documentos de la lengua catalana) hasta el siglo XVII. Estos textos están dispuestos de tal modo que, gestionados con el programa de consulta *Estación de Análisis Documentales* (EAD), desarrollado en el Seminario de Filología e Informática de la Universidad Autónoma de Barcelona, puede facilitar a sus usuarios toda una serie de datos y de información útil para estudios tanto de carácter lingüístico como documental.

CODCAR: Corpus de Documentación Cancilleresca del s. XIII

Este proyecto ha pasado por tres grandes fases: Hispanic Seminary (1995-2002), red CHARTA (2008-actualidad) y, por último, una tercera en la que se explora el aprovechamiento de las herramientas digitales para la creación de una edición digital en XML-TEI de dichos documentos y una extracción de datos con información estadística por medio del sistema LYNEAL. CODCAR ofrece en línea 756 documentos originales de cancillería real de Fernando III a Fernando IV, con el reparto siguiente: 45 documentos de Fernando III, 386 de Alfonso X, 224 de Sancho IV y 101 de Fernando IV.

CODEA: Corpus de Documentos Españoles Anteriores a 1800

Este es un corpus general que recopila documentos españoles desde los orígenes de la lengua hasta 1800 con la finalidad de establecer, difundir y facilitar el acceso a unos textos valiosos por sí mismos. Creado por el Grupo de Investigación de Textos para la Historia del Español (GITHE[99]) de la Universidad de Alcalá, este subcorpus de CHARTA contiene 2500 documentos en español de toda la geografía peninsular del español y de diferentes registros (desde la Cancillería a las notas de manos inhábiles), aunque el CHARTA contiene sólo 800 de estos documentos. CODEA+ 2015 es la versión avanzada de CODEA 2011. Además, proporciona una forma avanzada de visualización, convirtiéndose así en un Atlas Lingüístico Diacrónico y Dinámico del Español (ALDIDI). Naturalmente, este corpus sigue

[99] http://textoshispanicos.es/index.php?title=P%C3%A1gina_principal

los criterios de la red CHARTA y también permite filtrar[100] por siglo, lugar, fecha y otros. El corpus también está disponible para su consulta en LYNEAL[101].

Actualmente, el proyecto sigue en curso y se pretenden añadir otros 1500 documentos y ampliar el arco temporal para incluir el siglo XIX, dando lugar así al CODEA+ 2020: Corpus de documentos españoles anteriores a 1900.

CORDE: Corpus Diacrónico del Español

Se trata de un corpus textual de todas las épocas y lugares en que se habló español, desde los inicios del idioma hasta el año 1974, en que limita con el Corpus de Referencia del Español Actual (CREA). El CORDE está diseñado para extraer información con la cual estudiar las palabras y sus significados, así como la gramática y su uso a través del tiempo. Hoy es fuente obligada para cualquier estudio diacrónico relacionado con la lengua española. La Academia utiliza sistemáticamente el CORDE para documentar palabras, para calificarlas de anticuadas o en desuso, para saber el origen de algunos términos, su tradición en la lengua, las primeras apariciones de las palabras, etc. Sirvió, además, de material básico para la confección del Nuevo diccionario histórico del español.

Cuenta en la actualidad con 250 millones de registros correspondientes a textos escritos de muy diferente género. Se distribuyen en prosa y verso y, dentro de cada modalidad, en textos narrativos, líricos, dramáticos, científico-técnicos, históricos, jurídicos, religiosos, periodísticos, etc. Se pretende recoger todas las variedades geográficas, históricas y genéricas para que el conjunto sea suficientemente representativo.

Las consultas[102] permiten filtrar por autor, obra, período, lugar, medio de publicación y tema.

CORDEREGRA: Corpus diacrónico del español del reino de Granada (1492-1833)

Se ci,en la lengua no literaria del periodo 1492-1833. Principalmente, este corpus se compone de tres tipos textuales: protocolos notariales, cartas y pleitos criminales. Estos últimos suelen incluir declaraciones de testigo, en las que se reproducen literalmente las palabras propias o ajenas (discurso directo).

[100] http://corpuscodea.es/corpus/consultas.php
[101] http://shimoda.lllf.uam.es/ueda/lyneal/codea.htm
[102] https://corpus.rae.es/cordenet.html

CORDIAM: Corpus Diacrónico y Diatópico del Español de América

Contiene textos escritos en América y comprende cuatro siglos de profundidad histórica. El primer documento corresponde al año 1494 y el último al año 1905. *Cordiam-Prensa* contiene únicamente textos de los siglos XVIII y XIX, ya que solo hubo prensa en América a partir del siglo XVIII.

CORHIBER: Portal de Corpus Históricos Hiberorrománicos

Portal que reúne corpus de carácter histórico relacionados con las lenguas de la Península Ibérica que pueden ser útiles para el estudio de las lenguas iberorrománicas.

CORHIJE: Corpus Histórico Judeoespañol

El CORHIJE es un corpus lingüístico accesible en línea, representativo de la evolución de la lengua sefardí, y está concebido tanto para el investigador como para el lector curioso en general por su carácter añadido de colección documental. Desde una interfaz web se pueden efectuar búsquedas lingüísticas complejas sobre un número creciente de ediciones críticas de textos sefardíes de distintos lugares, géneros y épocas, y acceder a los documentos originales. Estos han sido metadescritos según el estándar Dublin-Core, lo que facilita, tanto su filtrado (por autor, título, lugar, tipo de texto, palabras clave, etc.), como el establecimiento de corpus paralelos alineados en ciertos casos. Los textos, a su vez, están siendo lematizados y etiquetados lingüísticamente mediante una versión adaptada de Freeling, en un nuevo prototipo.

CORHEN: Corpus Histórico del Español Norteño

Subcorpus de CHARTA formado por documentación privada medieval de las variedades castellanas norteñas. De libre acceso, ofrece a los usuarios los documentos en una doble presentación, paleográfica y crítica (y la imagen cuando se cuenta con el permiso de los archivos), así como la posibilidad de hacer búsquedas en los textos y filtrar por diferentes parámetros. Cuenta con 253 documentos del fondo del monasterio de San Salvador de Oña, con fechas extremas que van del año 822 (en copia del s. XIII) a 1280.

CorLexIn: Corpus Léxico de Inventarios

Corpus de finalidad muy específica, destinado al estudio del léxico de la vida cotidiana del Siglo de Oro. Constituido por documentos notariales del Siglo de Oro y realizado en colaboración con el equipo del Diccionario histórico de la lengua

española (DHLE) Está disponible para su consulta pública en los enlaces externos de la web de la Real Academia Española[103].

CORPES XXI: Corpus del Español del Siglo XXI

El CORPES XXI es, al igual que CREA, un corpus de referencia. Pretende ser un corpus que se mueva en los parámetros utilizados actualmente en esta línea de trabajo: 25 millones de formas por año y una distribución general del 70 % para textos americanos y el 30 % para textos españoles. Se concibe como un corpus semiabierto, un corpus que se irá incrementando en los próximos años con las cantidades previstas. El 90% de los textos corresponde a la lengua escrita y el 10% a la lengua oral.

Los materiales escritos proceden en un 40% de libros, publicaciones periódicas (40%), material de Internet (7,5%) y miscelánea (2,5%).

COSER: Corpus Oral y Sonoro del Español Rural

Formado por grabaciones de la lengua hablada en enclaves rurales de la Península Ibérica. Las entrevistas se obtuvieron con el propósito de ofrecer una muestra representativa de la variedad dialectal, pero también permiten conocer los modos de vida en el campo en la época previa a la mecanización agraria y a la despoblación rural.

COTAGAL: Corpus de Textos Antiguos de Galicia

Subcorpus de CHARTA, recoge obras literarias y prosa documental producidos fundamentalmente en la Galicia medieval (siglos IX-XVI) y moderna (XVI-XVIII).

CREA: Corpus de Referencia del Español Actual

Conjunto de textos de diversa procedencia para la extracción y estudio de palabras, sus significados y contextos. Está disponible para consulta en https://corpus.rae.es/creanet.html. En marzo de 2021 se publica una nueva versión del corpus anotado, donde es posible hacer la búsqueda por formas, lemas y categorías gramaticales.

DCEP: Digital Corpus of the European Parliament

El Corpus Digital del Parlamento Europeo contiene la mayoría de los documentos publicados en el sitio web oficial del Parlamento Europeo. Comprende una variedad de tipos de documentos, desde comunicados de prensa hasta documentos

[103] https://apps2.rae.es/CORLEXIN.html

legislativos y de sesión relacionados con las actividades y los órganos del Parlamento Europeo. La versión actual del corpus contiene documentos que se produjeron entre 2001 y 2012. El corpus está disponible como documentos de texto completo y como datos alineados por oraciones. Incluye información de alineación para los documentos completos, así como para las oraciones, producidas por separado para cada par de idiomas.

DOLEO: Documentación de Lamento en Español desde Orígenes

Subcorpus de CHARTA, recoge textos desde la edad media hasta el siglo XIX, en concreto textos de quejas. CHARTA incluye solo una muestra parcial de DOLEO: un texto del siglo XVI, dos del XVII y dos del XVIII.

ECPC: European Comparable and Parallel Corpora

El Equipo de Corpus Comparables y Paralelos de Discursos Parlamentarios Europeos (ECPC) es un grupo de investigación multidisciplinar y plurinacional que emplea la tecnología en el estudio de una de las actividades potencialmente más cooperativas de la sociedad: la traducción. ECPC se encarga de analizar discursos pronunciados en las sesiones plenarias de varios parlamentos europeos: el Parlamento Europeo (EP), el Congreso de los Diputados español (CD) y la Cámara de los Comunes británica (HC). Sus objetivos son alinear discursos, estudiarlos y desarrollar herramientas online para ponerlas a disposición de la comunidad científica http://www.ecpc.uji.es.

Europeana

Europeana[104] es la biblioteca digital europea, de acceso libre, cuyo prototipo comenzó a funcionar el 20 de noviembre de 2008, que reúne contribuciones ya digitalizadas de reconocidas instituciones culturales de los 27 estados miembros de la Unión Europea. Sus fondos incluyen libros, películas, pinturas, periódicos, archivos sonoros, mapas, manuscritos y otros archivos.

Desde el punto de vista técnico, Europeana es el portal del patrimonio cultural europeo que comenzó con dos millones de objetos digitales y cuya colección alcanzó los 29 millones de documentos en 2013, aportados por unas 2300 instituciones formadas por bibliotecas, archivos, galerías y museos. La colección está formada por una gran variedad de documentos de 45 idiomas: libros, periódicos, revistas, cartas, diarios, documentos de archivo, cuadros, pinturas,

[104] https://www.europeana.eu/es

mapas, dibujos, fotografías, música, tradición oral grabada, emisiones de radio, películas y otros programas televisivos.

Hemeroteca Digital (Biblioteca Digital Hispánica)

La Hemeroteca Digital[105] forma parte del proyecto Biblioteca Digital Hispánica, que tiene como objetivo la consulta y difusión pública a través de Internet del Patrimonio Bibliográfico Español conservado en la Biblioteca Nacional. Esta Hemeroteca nace en marzo de 2007 para proporcionar acceso público a la colección digital de prensa histórica española que alberga la Biblioteca, con una colección inicial compuesta por 143 títulos de prensa y revistas. Actualmente su colección alberga más de 1000 títulos, con intención de seguir ampliándose hasta cubrir la evolución histórica de la prensa española, desde sus inicios hasta principios del siglo XX.

HISPANA

HISPANA es un directorio y recolector de recursos digitales. Se trata de un recurso avanzado de acceso a la información digital producida por todo tipo de instituciones españolas que se constituye en la red mediante la interconexión de sus bases de datos. Este directorio incluye algunos de los más importantes recursos, como la Biblioteca Virtual de Prensa Histórica del Ministerio de Cultura, ofreciendo un sistema de búsqueda centralizado que permite lanzar una consulta sobre todos ellos desde el formulario de Hispana.

Index Thomisticus

El proyecto *Corpus Thomisticum* pretende poner a disposición de los investigadores un conjunto de instrumentos para el estudio de Tomás de Aquino, accesible gratuitamente a través de Internet. Consta de cinco partes: una edición íntegra de las obras completas de Sto. Tomás conforme, en lo posible, a los mejores textos críticos; el catálogo bibliográfico de todos los trabajos aparecidos desde el siglo XIII hasta nuestros días sobre Sto. Tomás y su doctrina; el índice de los principales instrumentos de investigación tomista existentes, y la edición de los más relevantes; un sistema de gestión de bases de datos capaz de encontrar, reunir y ordenar palabras, frases, citas, semejanzas, correlaciones y datos estadísticos, y la edición digital de los manuscritos principales de las obras de Sto. Tomás.

El latín es la lengua principal del Corpus Thomisticum, pues cualquier estudioso de Santo Tomás conoce la lengua en que están escritas sus obras.

---

[105] https://www.bne.es/es/catalogos/hemeroteca-digital

KorBa

El KorBa es un corpus electrónico de textos polacos de los siglos XVII y XVIII. Consiste en muestras extraídas de más de 700 textos escritos y publicados entre 1601 y 1772 tanto en su forma transliteral como transcrita, conteniendo un total de 13,5 millones de formas, lo cual lo hace uno de los corpus históricos más grandes en una lengua eslava. Es un corpus grande, balanceado, lematizado y anotado estructural y morfológicamente, y está disponible para su consulta avanzada[106] mediante el motor de búsqueda MTAS (Multi Tier Annotation Search).

LOB: Lancaster-Oslo-Bergen Corpus

Es una colección de un millón de palabras de textos en inglés británico que se compiló en la década de 1970 en colaboración entre la Universidad de Lancaster, la Universidad de Oslo y el Centro Noruego de Computación para las Humanidades de Bergen, para proporcionar una contraparte británica al Brown Corpus compilado por Henry Kučera y W. Nelson Francis para el inglés americano en la década de 1960. Su composición fue diseñada para coincidir con el corpus original de Brown en términos de tamaño y géneros lo más cerca posible utilizando documentos publicados en el Reino Unido por autores británicos. Ambos corpus constan de 500 muestras cada uno compuestas por unas 2.000 palabras.

Memoriademadrid (Bibioteca Digital)

La Biblioteca Digital *Memoriademadrid* es un proyecto de difusión de la memoria histórica de la ciudad a través de la digitalización del patrimonio histórico del Ayuntamiento de Madrid, especialmente el custodiado en sus archivos, museos y bibliotecas. Algunos de los documentos que se pueden consultar son: fotografías, periódicos y revistas, tarjetas postales, libros de los siglos XVI al XX, objetos de museos, etc. que nos dan una idea de la evolución de la ciudad y de la vida cultural madrileña. Actualmente cuenta con un total de 202.383 documentos en sus fondos.

ODE: Oralia diacrónica del español

Este corpus es una continuación del corpus CORDEREGRA (Campos & Godoy, 2009), formado principalmente por declaraciones de testigos, inventarios de bienes y certificaciones de barberos y cirujanos. ODE presenta dos novedades respecto del antiguo CORDEREGRA: la ampliación geográfica, y la transcripción de los manuscritos en XML siguiendo el modelo del proyecto Post Scriptum.

[106] https://www.korba.edu.pl

PRESEEA

Corpus de lengua española hablada, representativo del mundo hispánico en su variedad geográfica y social. Esos materiales se reúnen atendiendo a la diversidad sociolingüística de las comunidades de habla hispanohablantes. Está disponible para consulta en LYNEAL y también en su web.

Proyecto Post Scriptum

Desarrollado con la herramienta TEITOK, tiene por objeto la investigación sistemática, edición y estudio histórico-lingüístico de cartas privadas escritas en España y Portugal durante la Edad Moderna. Contiene anotación morfosintáctica y sintáctica y cuenta con un sistema de búsqueda avanzada.

TMILG: Tesoro Medieval Informatizado de la Lengua Gallega

El corpus TMILG abarca unas 16.000 unidades textuales distribuidas en un total de 82 obras, representativas de las tres grandes categorías reconocibles en la producción textual de la Galicia medieval: prosa notarial, prosa no notarial y poesía (verso).

Apéndice II. Diccionarios, lexicones y otros recursos léxicos y semánticos

ALDICAM-CM (Inventario léxico del Atlas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid)

El proyecto ALDICAM-CM[107] tiene como objeto elaborar un Atlas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid. Es un proyecto innovador, pues nunca hasta ahora se había llevado a cabo un desarrollo similar para ninguna otra lengua. El punto de partida de la propuesta es la transcripción de un número importante de documentos elaborados entre el siglo XIII y el XIX en diferentes localidades de la actual Comunidad de Madrid, procedentes del Archivo Regional de la CM, del de la Villa y de los diferentes archivos municipales, de instituciones eclesiásticas e incluso de particulares. Los documentos se ofrecerán en una triple edición, de acuerdo con una metodología largamente ensayada, y con los criterios de la red internacional CHARTA. Es un proyecto interdisciplinar que nace de la conjunción del trabajo de paleógrafos y archiveros, historiadores de la lengua y de la cultura general, de la historia de la escritura, y tecnólogos especializados en humanidades digitales.

Hasta ahora se habían elaborado mapas lingüísticos (en papel o en formato digital) con imágenes fijas de la distribución espacial de determinadas variantes lingüísticas, mientras que en el ALDICAM los resultados de cualquier búsqueda de variantes gráfico-fonéticas (cosa/cossa, ffijo/fijo/hijo/ijo/ixo, yugo/yubo), morfosintácticas (leísmo, laísmo), léxicas (prínsoles, sobrefalso, azadón polaino, espadador) y sintácticas (neutro de materia, lana blanco), pueden proyectarse directa e inmediatamente en un mapa digital de la CM, de manera que se obtenga la distribución espacial de las formas buscadas. El mapa no es solo dinámico, sino, además, interactivo, pues el investigador, o usuario general, puede establecer los parámetros de cada consulta al corpus (límites temporales, localidad o localidades de la CM, tipo documental, elaboración femenina del documento, tipo de letra, escribano, etc.), combinados según los interese del usuario. De este modo puede generarse un número ilimitado de mapas, que nos proporcionan información fehaciente, respaldada documentalmente, de la trayectoria histórica del habla de Madrid. Así el proyecto pone a la CM en la vanguardia de la investigación

[107] http://aldicam.blogspot.com/p/que-es-aldicam-cm.html

geolingüística, permitiendo estudiar la distribución geográfica de las variantes lingüísticas, pero, además, la interacción entre los factores diacrónico, diatópico, diastrático, diafásico... contribuyendo a un mejor conocimiento de la lengua y la sociedad de la CM a lo largo del tiempo.

El corpus ALDICAM cuenta con un inventario léxico[108] que ha servido de base para elaborar el atlas. Por un lado, una Excel con una tabla de que contiene ID del documento, fecha con intervalo de 50 años, lema, categoría gramatical, forma modernizada, forma crítica, forma paleográfica simplificada y detallada, y contexto amplio. Por otro, ficheros en PDF que contienen los lemas, formas críticas y paleográficas junto con sus frecuencias normalizadas por mil palabras (con la frecuencia absoluta no se pueden apreciar los cambios históricos comparativamente).

BabelNet

BabelNet[109] es un diccionario enciclopédico multilingüe, con una gran cobertura tanto lexicográfica como enciclopédica de los términos, y una red semántica u ontología que conecta los conceptos y las entidades en una gran red de unos 22 millones de entradas. Ideado en el grupo Sapienza NLP de la universidad Sapienza de Roma y desarrollado y mantenido por Babelscape, BabelNet sigue el modelo de WordNet basado en *synsets*, pero lo extiende para contener lexicalizaciones multilingües: cada synset de BabelNet representa un significado dado y contiene todos los sinónimos que expresan ese significado en una gama de idiomas diferentes.

Catálogo de autoridades de la BNE

El catálogo de autoridades ofrece acceso a más de 300.000 registros de autoridad de los encabezamientos empleados en los registros bibliográficos del catálogo. Los registros de autoridad establecen de forma normalizada el encabezamiento que se utiliza en los registros bibliográficos como puntos de acceso, asociados a una persona, entidad corporativa, título o materia y conforme a la normativa de los Grupos de Trabajo específicos de IFLA.

Además de los formatos convencionales, la BNE pone a disposición de la ciudadanía su catálogo estructurado semánticamente, aplicando tecnologías de la Web Semántica y enriqueciendo los registros con recursos de fuentes externas como es la ontología VIAF. Esta iniciativa se ha puesto en marcha con la

[108] http://shimoda.lllf.uam.es/ueda/lyneal/il/aldicam/
[109] https://babelnet.org

publicación, conforme a los principios de Linked Data, de información procedente de los catálogos, haciéndolos disponibles como bases de conocimiento RDF (Resource Description Framework). Por un lado, se presentan los datos en formato original (MARC y MARC-XML), y por otro lado los datos mapeados en formatos CSV, JSON, ODS, TXT y XML.

DBpedia

DBpedia es un proyecto para la extracción de datos de Wikipedia para proponer una versión Web semántica. Este proyecto es realizado por la Universidad de Leipzig, Universidad Libre de Berlín y la compañía OpenLink Software.

La información se almacena con RDF, y se pueden hacer consultas a la base de datos a través de SPARQL.

El proyecto DBpedia ha generado durante mucho tiempo información semántica a partir de la Wikipedia inglesa. Desde junio de 2011 el proceso de generación de información extrae infomación de Wikipedia en 15 de sus versiones (o idiomas, uno de ellos el español). El comité de internacionalización de DBpedia ha asignado un sitio web y un SPARQL Endpoint para cada uno de estos idiomas. En el caso de **es**.dbpedia.org (sitio en español), el proceso de extracción produce 100 millones de triples RDF a partir de la versión para el español de la wikipedia. En el SPARQL endpoint están disponible todos estos triples.

Este trabajo depende de investigadores, pertenecientes a la Red Temática Española de *Linked Data*, así como de particulares que dedican su tiempo y su esfuerzo a esta iniciativa.

Diccionario de autoridades 1726-1739 de la RAE

Primer repertorio lexicográfico del español confeccionado por la Real Academia Española, fundamento de lo que hoy se conoce como el Diccionario de la lengua española. Disponible en el catálogo de la Biblioteca Nacional de España[110], cuenta con un total de 69.410 entradas, y también es posible realizar consultas desde la web de la RAE[111].

Diccionario español medieval - español[112]

[110] http://catalogo.bne.es/uhtbin/authoritybrowse.cgi
[111] https://apps2.rae.es/DA.html
[112] https://es.glosbe.com/osp/es

Un diccionario colaborativo que permite traducir términos del español al español medieval y viceversa, y muestra ejemplos de uso en sus respectivos contextos.

Diccionario Histórico de la Lengua Española (DHLE) de la RAE[113]

Este diccionario nativo digital persigue describir en su integridad (en el eje diatópico, diastrático y cronológico) la historia del léxico de la lengua española. Una característica definitoria de este repertorio radica en su voluntad de analizar la historia del léxico en una perspectiva relacional, atendiendo a los vínculos etimológicos, morfológicos y semánticos que se establecen entre las palabras. Concebido desde sus orígenes como una base de datos léxica electrónica (y diacrónica), lo que permite elaborar sus artículos de acuerdo con un criterio de organización del trabajo por campos semánticos (o voces relacionadas por su significado) y familias léxicas.

EuroWordNet

En 2014 se celebraron las V jornadas de la red temática en Tratamiento de Información Multilingüe y Multimodal (TIMM), una red de investigación financiada por el Ministerio de Economía y Competitividad cuyo objetivo es promover la investigación en Tecnologías del Lenguaje. Como resultado, se elaboró una colección de recursos lingüísticos entre los que se engloba EuroWordNet, una base de datos multilingüe que contiene relaciones semánticas entre palabras. Originariamente se creó WordNet, una base de datos léxica en inglés en la que los nombres, verbos, adjetivos y adverbios están agrupados en conjuntos de sinónimos a nivel semántico, o sinónimos cognitivos, llamados *synsets*. Cada synset expresa un concepto y entre ellos se relacionan a su vez mediante relaciones semánticas o conceptuales como la hiperonimia, la hiponimia o la meronimia. La principal motivación de su creador, George A. Miller, psicólogo cognitivo de Princeton, fue desarrollar un modelo que fuese consistente con el modo en que los humanos procesan el lenguaje. Así, WordNet es usado e interpretado como una ontología.

El objetivo de EuroWordNet es desarrollar una base de datos léxica con relaciones semánticas en varios idiomas europeos: alemán, holandés, checo, estonio, italiano, francés y español además del inglés. Cada idioma cuenta con 30.000 conceptos que a su vez se relacionan con el WordNet inglés, de manera que cada concepto en inglés cuente también con los equivalentes en el resto de idiomas. Este recurso es útil para diversas tareas de procesamiento del lenguaje natural: al representar

[113] https://www.rae.es/dhle/

los términos a nivel conceptual es idóneo para realizar traducciones más precisas, con lo cual es una herramienta útil para la recuperación de información tanto monolingüe (haciendo uso de la sinonimia) como multilingüe (como soporte a la traducción). Del mismo modo, es adecuado para la tarea de Question Answering para la expansión de consultas y desambiguación de las palabras clave.

Fichero general de la RAE

A lo largo de su historia, la Academia ha ido construyendo, al tiempo que se utilizaba para la elaboración de sus obras, un fichero que consta de unos diez millones de papeletas, léxicas y lexicográficas, actualmente digitalizado y disponible para su consulta[114].

Inventario léxico del corpus CODEA

El corpus CODEA cuenta con un inventario léxico dividido en secciones alfabéticamente[115]. Actualmente este inventario incluye términos de Castilla la Vieja[116] entre los siglos XII y XVIII. Cada sección consiste en una hoja de Excel que contiene el lema de cada palabra, su inicial, categoría gramatical, forma moderna, crítica y paleográfica, un contexto mínimo, año, franja cronológica con intervalo de 50 años, provincia, identificador CODEA, tipo de documento, hoja y línea.

Nuevo tesoro lexicográfico de la lengua española (NTLLE)

Se trata de un "diccionario de diccionarios" que reúne una amplia selección de las obras que durante los últimos quinientos años han recogido, definido y consolidado el patrimonio léxico de nuestro idioma. Contiene, dentro de un entorno informático de consulta, los facsímiles digitales de las obras lexicográficas de Antonio de Nebrija, fray Pedro de Alcalá, Sebastián de Covarrubias, Francisco del Rosal, César Oudin, Esteban Terreros, Ramón Joaquín Domínguez, Vicente Salvá, Elías Zerolo, Aniceto de Pagés, etc., además de toda la lexicografía académica, desde el Diccionario de autoridades hasta la 21.ª edición del Diccionario de la Real Academia Española, pasando por las diversas ediciones del Diccionario manual e ilustrado y lo publicado del Diccionario histórico de 1933-1936.

Ofrece la posibilidad de tener juntos y reunidos cerca de 70 diccionarios que ninguna biblioteca en el mundo está en condiciones de custodiar de forma

[114] https://apps2.rae.es/fichero.html
[115] http://shimoda.lllf.uam.es/ueda/lyneal/ilc-cv.htm
[116] https://es.wikipedia.org/wiki/Castilla_la_Vieja

conjunta, al tiempo que permite buscar de una sola vez, a través de una única operación de consulta, una o varias palabras de forma simultánea en la totalidad de los diccionarios que lo integran.

Portal de datos bibliográficos de la BNE

En fase beta, propone al usuario un nuevo modo de acercarse a las colecciones y recursos de la Biblioteca Nacional de España. Es un proyecto de publicación de datos como Linked Open Data, basado en tecnologías y estándares de la Web. Permite también hacer consultas mediante SPARQL[117].

Tesauros del Patrimonio Cultural de España

El Ministerio de Cultura y Deporte difunde y hace accesible a todos los ciudadanos un recurso para el conocimiento de nuestro patrimonio cultural. El objetivo es dar a conocer la riqueza de nuestros bienes culturales mediante el vocabulario utilizado para su identificación, clasificación, descripción y catalogación. El acceso a estos vocabularios a través de un formato abierto y enriquecido semánticamente los convierte en un punto de unión con otros recursos similares en Internet (Linked Open Data).

[117] https://datos.bne.es/sparql

Apéndice III. CLARA-DM: Guía de estilo para la transcripción de periódicos del Diario de Madrid

En esta sección se presenta el caso de uso de Transkribus para la creación del Corpus CLARA-DM, los pasos seguidos y las dificultades encontradas en el proceso.

1. Descarga de documentos del Diario de Madrid (1788-1825)

Para descargarlos desde la hemeroteca digital de la BNE hay que acceder a la dirección http://hemerotecadigital.bne.es/index.vm?q=id:0001510462&lang=es, escribir en el título "Diario de Madrid" y seleccionar el año y el día concretos[118] (Figura 24).

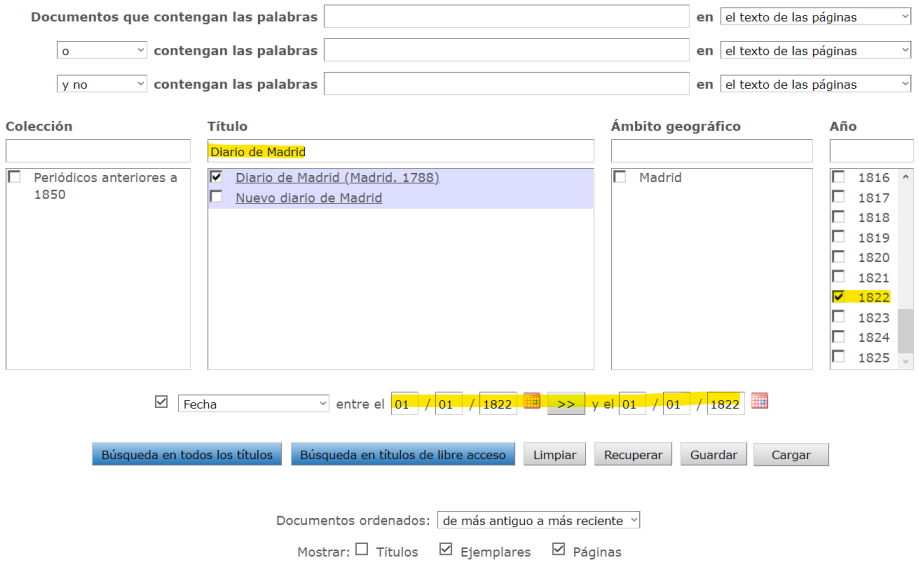


Figura 24. Descargar ejemplares del Diario de Madrid desde la Hemeroteca Digital.

Después, para descargar el documento completo (suelen ser 4 páginas, 8, o 12 en algunos casos) hay que pinchar arriba a la izquierda en la página (Figura 25).

En la primera tanda se acuerda descargar el primer día del mes de todos los años.

[118] Los números entre 1808 y 1814 se pueden bajar desde la biblioteca digital Memoriademadrid (http://www.memoriademadrid.es/buscador.php?accion=VerFicha&id=4603&num_id=1&num_total=70) ya que algunos tienen mejor calidad.

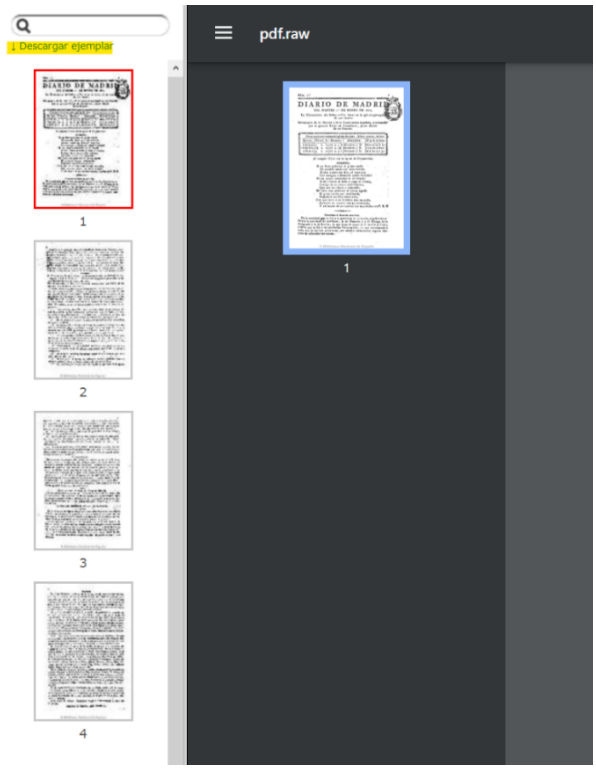


Figura 25. Descarga del archivo completo.

El primer ejemplar de cada trimestre tiene un índice de los temas de los periódicos de ese trimestre, por lo que esos ejemplares no tienen 4 sino 12 páginas. Si, por un casual, un periódico sólo contiene números de sorteos de acciones (no suele ser el caso), se decide descargar el día 2 del mes.

2. Normalización de los nombres de los documentos

Como los documentos proceden de dos páginas web distintas (la hemeroteca digital de la BNE y Memoriademadrid), se conviene normalizar el nombre de todos los documentos con la estructura *Diario de Madrid día-mes-año.*

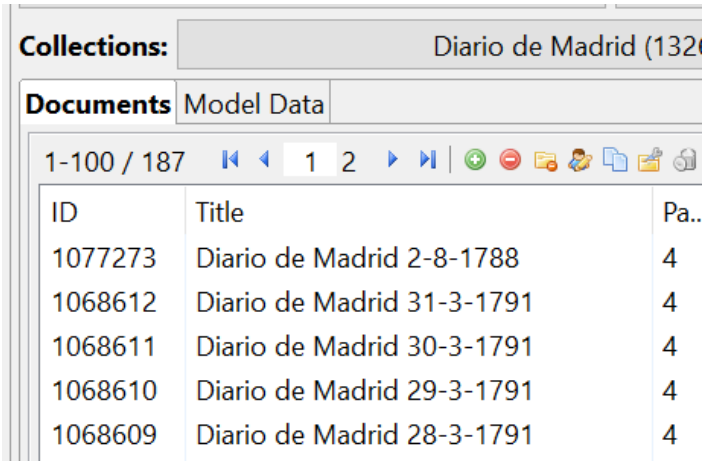


Figura 26. Normalización de los nombres de los documentos.

3. Colecciones

Para llevar un seguimiento de la cantidad de documentos subidos a la plataforma, transcritos y con el reconocimiento del layout realizado, se crean tres colecciones distintas. Cuando se concluye una tarea en un documento, este se enlaza a la colección correspondiente a esa tarea y se desliga de la colección anterior.

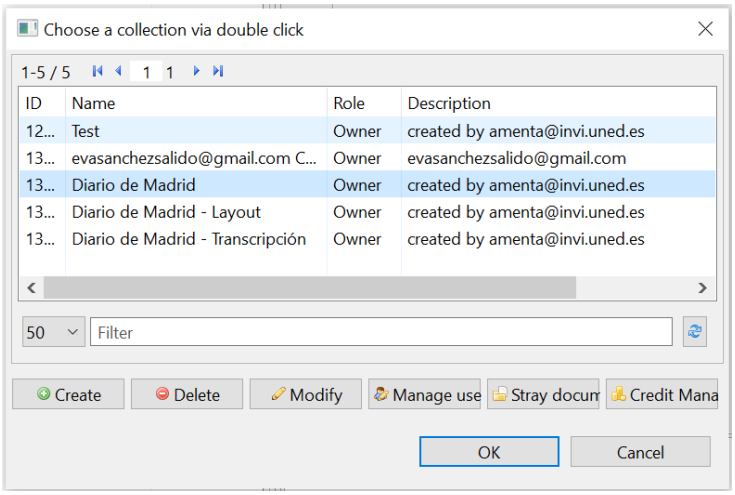


Figura 27. Gestión de documentos mediante colecciones.

4. Reconocimiento del layout

A la hora de hacer el reconocimiento de la estructura de los documentos, ha sido necesario discutir algunos aspectos y llegar a un acuerdo en la forma de realizarlo.

La Tabla 13 recoge las decisiones tomadas respecto a los asuntos que se identificaron en el proceso.

5. Transcripción

Al igual que con el reconocimiento de la estructura, a la hora de transcribir los documentos ha sido necesario tomar algunas decisiones para obtener unas transcripciones lo más homogéneas posible entre todos los colaboradores. Estas decisiones se recogen en la Tabla 14.

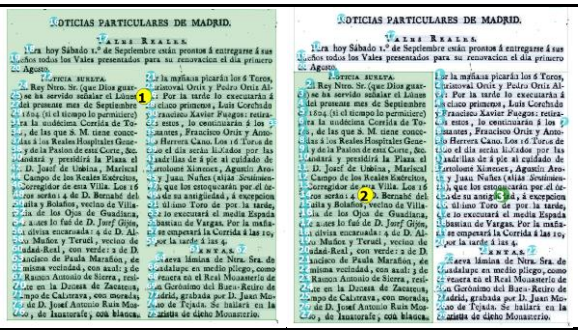
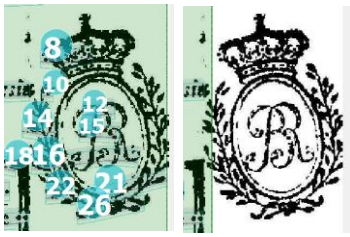
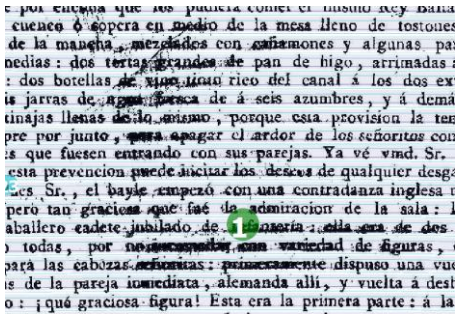
Tablas	En presencia de tablas se divide el texto en regiones para que la transcripción se haga en un orden coherente	
Columnas	Las regiones de texto en columnas se dividen para obtener el orden de lectura adecuado	
Manchas fuera del texto y sellos	Se ignoran o se eliminan cuando el modelo los haya incluido	
Manchas dentro del texto	Se realiza el layout normalmente. Si el texto es legible o se puede adivinar, se transcribe y se ignora la mancha. Si no es legible, se marcará en la transcripción con la etiqueta **[ilegible]**.	
Líneas de puntos	Los puntos son caracteres que el modelo intentará predecir con precisión, y por ello es importante que no haya puntos de más o de menos en la transcripción. En caso de que incluirlos suponga mayor complicación, es mejor eliminarlos del layout.	

Tabla 13. Guía de estilo para el reconocimiento del layout de los documentos del corpus CLARA-DM.

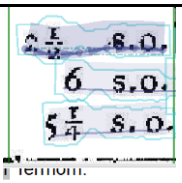
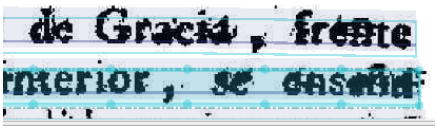
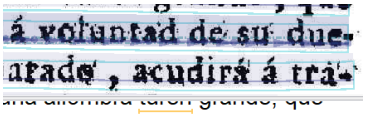
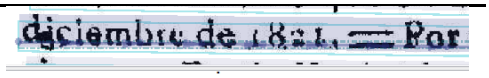
Ortografía	Se conserva la ortografía original. También se conserva el uso de mayúsculas y cursiva.	 el bayle de la maroma, y el bayle bo- de Laureno, el bayle de la maroma, y el bayle bo-
Número de página y del Diario	Se escriben en arábigo (1, 2,…) a pesar de que en la primera página parezca romano.	 Pág. 1 1-1 Núm. 1 1-2 Pág. 1
Fracciones	Las fracciones como la del ejemplo se escriben 7½, ¼, ⅓ o, en su defecto, la fracción después de un espacio (7 1/2).	 2 2½ s.o. 3 6 s.o. 4 5¼ s. o.
Palabras en duda	Se intenta adivinar lo que pone ya que el modelo también debe ser capaz de hacerlo. Es más importante pensar en lo que el escritor quería poner que en lo que el modelo va a interpretar. Cuando no sea posible interpretar el texto, escribir **[ilegible]** en la transcripción de esa parte.	 qto principal interior, se [ilegible]
Palabras divididas por guion al final de una línea	Se conservan imitando el documento original.	 do se vende á voluntad de su due- junto ó separado, acudirá á tra-
Espacios antes de las comas	Se conservan cuando sea necesario.	
Símbolos	Se intentan reproducir en la transcripción.	 diciembre de 1821. = Por el Ayuntamiento  3. reparo. Que omito algunos §§. sin da

Tabla 14. Guía de estilo para la transcripción de los documentos del corpus Diario de Madrid.

6. Exportación de los documentos

Transkibus permite la exportación de los documentos en distintos formatos. El formato TXT es útil para poder etiquetar los textos posteriormente, y poder pasarlos a un modelo de aprendizaje automático. El formato Word también es útil de por sí, ya que permite hacer búsquedas en el documento. La exportación a TEI queda demasiado cargada y engorrosa, por lo que la omitimos.

Además, la exportación a DOCX tiene una cosa interesante, y es que cuando una palabra se separa con un guión al final de la frase, es posible unirla en la transcripción. La exportación a texto mantiene los guiones, lo cual complica la aplicación de modelos de reconocimiento de entidades. Como solución a esto, se desarrolla un *script* que convierte los documentos Word a TXT.

Así, la exportación de los documentos se lleva a cabo en dos formatos, DOCX (Word) y TXT (texto plano).

6.1. Exportación a TXT

Esta opción es la más básica; las líneas del documento se recogen de manera idéntica en un archivo de texto plano. Transkribus no ofrece ningún tipo de personalización para la exportación de este tipo.

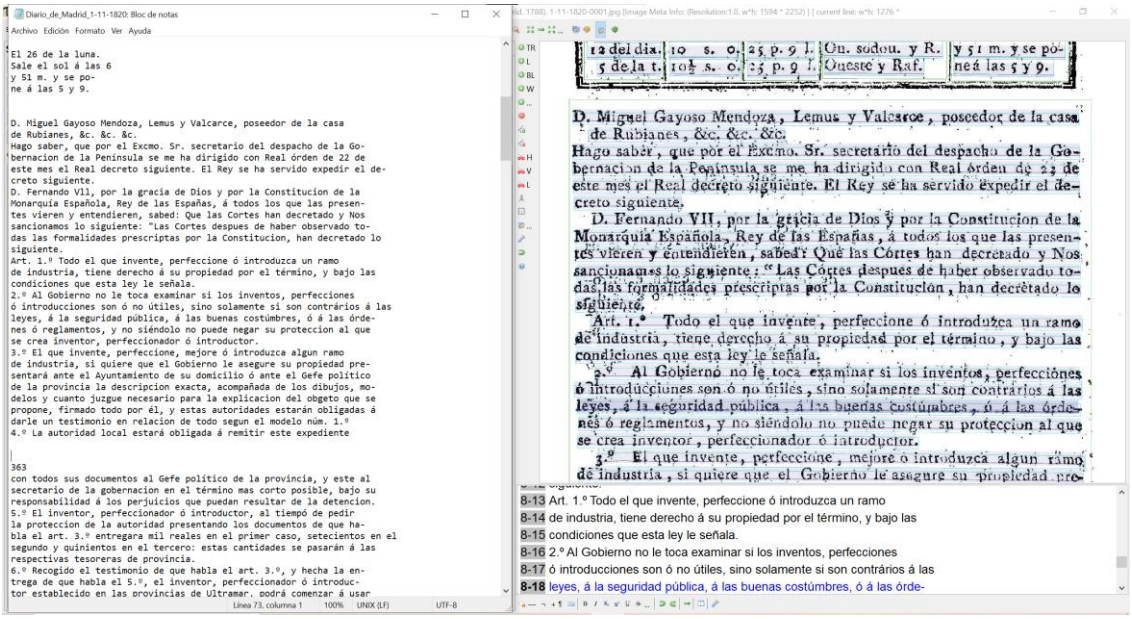


Figura 28. Exportación a texto plano en Transkribus.

6.2. Exportación a DOCX

La exportación a Word ofrece más posibilidades, entre ellas, la de elegir si conservar los saltos de línea tal y como aparecen en el documento, o ignorarlos, de manera que las palabras separadas al final de una frase en el documento original se muestran unidas en la exportación.

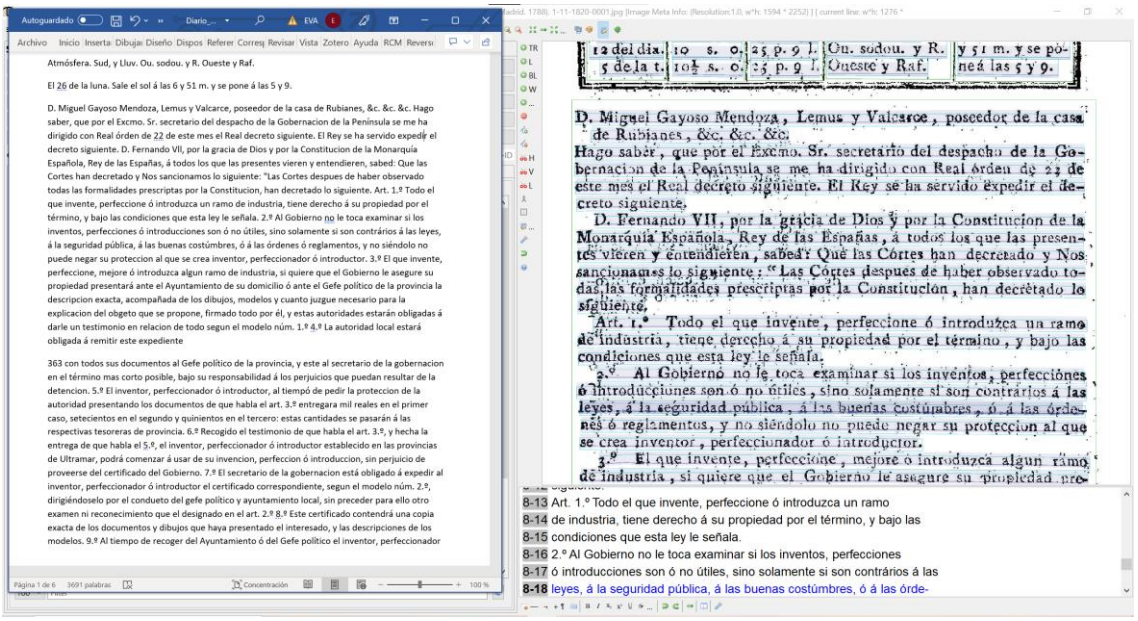


Figura 29. Exportación a DOCX en Transkribus.

6.3. Conversión de Word a TXT

Los técnicos del proyecto (Víctor Sánchez, Ricardo García y Andrés Rodríguez) desarrollan un *script* para convertir los documentos Word a TXT y poder aplicar una herramienta de etiquetado de textos. El código es el siguiente:

```
import os
import docx2txt
path = os.getcwd() #Ruta en la que esta el .py y los .docx
path2 = os.path.join(path, "TXTs") #Creamos (si no existe) una carpeta
                                   #para guardar los .txt
if not os.path.exists(path2):
    os.mkdir(path2)
listDOC = os.listdir(path) #Obtenemos el nombre de todos los archivos
                           #en el directorio
for doc in listDOC:
    if(doc.endswith(".docx")): #Solo trabajamos con los .docx
        txt = docx2txt.process(doc)
        fullName = os.path.join(path2, doc.split(".")[0] + ".txt")
#fullName = C:\Users\garsa\Desktop\Script\TXTs\Diario_de_Madrid_1-1-
1788.txt
        with open(fullName, "w", encoding="utf-8") as text_file:
            print(txt, file=text_file)
```

Para usarlo basta seguir los siguientes pasos:

1. Se coloca el archivo de código en la misma carpeta en la que se encuentran los documentos de Word.

2. Se abre el CMD y se dirige a la ruta en que se encuentra lo anterior.

3. Posiblemente antes del siguiente paso haya que instalar el modulo de Python "docx2txt". Para hacerlo, escribir “pip install docx2txt”.

4. Se ejecuta el script mediante “py DOCXtoTXT.py” y se espera unos segundos.

5. Se habrá creado una carpeta (TXTs) con todos los documentos convertidos a texto plano.

Apéndice IV. CLARA-DM: Propuesta de guía de estilo para el etiquetado de entidades nombradas

Lo que se presenta a continuación es la primera propuesta de guía de anotación de entidades para el corpus CLARA-DM con la herramienta Tagtog.

1. Tipos y subtipos de entidades

- Localizaciones
    - Localización general (loc): Mención que hace referencia a un lugar geográfico general que no forma parte del distrito de Madrid. Es decir, que no sea sensible de ser colocado en un mapa interactivo de Madrid. También incluye cualquier localización que no termine de pertenecer a ninguna del resto de categorías relacionadas con una localización (Calles, Plazas, Hospitales, Edificios Religiosos, Establecimientos, Edificios administrativos,...).

      Ejemplos: Madrid, España, Zaragoza, Talavera, Sevilla, India

    - Edificios religiosos (loc_relig): Pertenecen a esta categoría aquellas menciones que hacen referencia a edificios religiosos: conventos, iglesias, capillas, congregaciones, religiosas...

      Ejemplos: Iglesia de San isidro, capilla provisional del Prado, congregación de María Santísima, religiosas de la Encarnación.

    - Calles, Plazas, Puertas y direcciones (loc_direc): Categoría creada para las menciones de calles, lugares emblemáticos de Madrid, Plazas, Puertas, etc.

      Ejemplos: Rivera de Curtidores, calle de San Bernardo, portillo de Embajadores, Embarcadero del Canal,...

    - Establecimientos (establec): Categoría que abarca las menciones relacionadas con lugares donde se realiza una actividad comercial, administrativa, de ocio, u otros (establecimientos, organizaciones, administraciones, etc).

Ejemplos: teatro de la Cruz, posada de la Cruz, Ayuntamiento, librería de Paz y Dávila, Imprenta de Thevin.

- Colegios (loc_cole): Menciones relativas a colegios.

  Ejemplo: Real Academia de San Fernando

- Fuentes (loc_fuente): Menciones relativas a fuentes.
- Hospitales (loc_hosp): Aquellas menciones referentes a un Hospital o similar que aparecen en las noticias.

  Ejemplos: Hospital de Monserrat, hospital general, ...

Nota: Las tres últimas categorías (Colegios, Fuentes y Hospitales) apenas tienen menciones en una primera exploración. Muy posiblemente después de analizar se unificarán con la categoría loc_direc.

- Personas
  - Personas (pers): Menciones que hacen referencia a nombres de personas en general.

    Ejemplos: D. Luis de Garro, Loriery, Velarde

  - Casas de señores, Cargos importantes (pers_señores): Aquellas menciones que hacen referencia a personas importantes de la época: Marqueses, Condes, Jueces, Obispos. Suelen ir acompañadas siempre del cargo, título o denominación (Sr, Excma, ...)

    Ejemplos: Sr. D José Martinez Moscoso, Sr. D. Manuel Fernandez Gamboa, Excma. Sra. Doña Gerónima de Espinosa.

- Profesiones
  - Oficios y Profesiones (prof): Aquellas menciones referentes al oficio de una determinada persona, su cargo o posición dentro de la sociedad.

    Ejemplos: Sacerdote, guardas, secretario, sacristan, trompeta, sargento...

- Decoración
  - Adornos: Elementos que sirven de adorno tanto para las personas, como decoración de calles o casas.

    Ejemplos: alabastro, estatua, estampa, jaspe, mampara, mantelería, paño, cubo

    - Mobiliario: Elementos que sirven de mobiliario en una calle, edifico o casa.
- Venta y pérdida
    - Objeto en venta: Aquellos objetos, pertenencias y propiedades que se encuentran en venta en la sección de anuncios.

      Ejemplos: berlina, casa, pulsera, broche

    - Objeto perdido: Aquellos objetos, pertenencias y propiedades que se han perdido y salen anunciados en la sección PÉRDIDAS.

      Ejemplos: cristales de un rejido de pelo, perra doga, perra cachorra.

2. Criterio general de Anotación

Por defecto, las entidades deben incluir un nombre propio para todas las categorías, a excepción de elementos de decoración, oficios y profesiones, y productos. Algunas características comunes a los nombres propios son: utilización de mayúsculas, aparición en diccionarios enciclopédicos pero no en diccionarios léxicos, ausencia de significado (p. ej. Marqués de Fuentesol). Se descarta una definición más estricta de los nombres propios para que sea la intuición del anotador la encargada de decidir si una determinada mención debe ser etiquetada.

Se permite la etiquetación de frases largas.

> Ejemplo: Exequias fúnebres que la *congregración de María Santísima de la buena dicha*

Sólo las menciones más específicas se deben etiquetar, incluyendo descriptores y modificadores de la entidad (etiquetar la cadena de palabras más grande). Si un determinado adjetivo muestra duda, se recomienda visitar los listados creados por los historiadores como ejemplos de entidades. Modificadores que indican un ámbito general no deben ser incluidos (próximo, junto a, siguiente)

> Ejemplo: *calle de Embajadores*, frente á la *confiteria, núm 9, cuarto 2.º*

2.1. Límites de la anotación

No se anotan artículos (el, la...), signos de puntuación (puntos, comas...) ni números cardinales (dos, tres...) al comienzo o al final de una entidad.

> Ejemplos: en esta corte en la *Rivera de Curtidores*,
>
> patriotas madrileños sitiados en *Cadiz*;

El comienzo y final de una mención no puede tener solapamiento con otra diferente. Por lo tanto, si una mención interna se puede etiquetar a varias, sólo se elegirá una de ellas.

- NO ANOTAR: La ilustre [congregacion de [Maria] Santísima del Carmen]
- ANOTAR: La ilustre *congregacion de Maria Santísima del Carmen*

En las menciones de las entidades se excluyen:

- Oraciones subordinadas.
- Inserciones externas que dividen una mención. En este caso se etiquetan dos menciones separadas.
- Determinantes.

Se incluyen los modificadores anteriores y posteriores en la etiquetación de las menciones.

> Ejemplos: En el proveido por el *Sr. Dr. D Joaquín Barbagero, visitador eclesíastico interior* de esta corte.
>
> En la *calle de San Bartolomé, núm. 20*

2.2. Entidades anidadas

No se puede etiquetar entidades anidadas.

2.3. Anotación de entidades coordinadas

Entidades pertenecientes a listado se etiquetan de forma individual.

> Ejemplos: Señores *Ponce*, *Lopez*, *Diez*, *Campos* y *Alverá*.

2.4. Entidades con errores tipográficos

Menciones que contengan errores se etiquetan igualmente.

> Ejemplos: A las propias horas. *Madridr* 1º Julio de 1821.
>
> *Iglesia 575 parroquial de san Mantin*

2.5. Abreviaciones y acrónimos

Por defecto, se etiquetan las abreviaciones, a excepción de que contenga errores. En caso de estar rodeadas de paréntesis, estos no se etiquetan.

> Ejemplos: *La novéna de ánimas fundada por la Excma. Sra. Doña Gerónima de Espinosa*

La hermandad del santo Rosario cantado de Nuestra Señora de la Almudena (*Madrid*)

En ocasiones no aparece la palabra calle o teatro y aparece directamente "de la Palma". Se etiqueta a partir de la preposición *de.*

Ejemplo: TEATROS. En el *del Príncipe...*

2.6. Entidades anafóricas

No se etiquetan pronombres, determinantes o adjetivos que se refieren a otras menciones que aparecen en el texto.

3. Criterio para ambigüedades

Cuando existan ambigüedades entre dos determinadas categorías, se elegirá una de las dos por defecto para asignarlas aquellas menciones dudosas entre las dos.

- Persona-Profesión. Cuando existe ambigüedades entre si una mención es una persona o una profesión, se debe etiquetar como Persona. En principio, la entidad Persona será más específica y tiene mayor peso en el proyecto que las profesiones.

4. Reglas de anotación según el tipo de Entidad

- Personas. No se incluyen aquellas menciones que no contengan algún nombre propio.
- Localizaciones. Se incluyen los números de calle o edificio en la mención. Ejemplos: Que vive en la *calle de la Comadre, núm. 16.*

  se acudirá á la *calle de Leganitos, casa núm. 8, cuarto principal*

  No se incluye "id" cuando indica que se trata de una dirección mencionada anteriormente.

  Ejemplo: D. Antonio Madiano, *calle de Postas, núm. 10, media casa.* D. Francisco Cartresano, id, n. 11, tienda.

  No se etiquetan lugares comunes que no llevan asociado un punto específico dentro de la ciudad o que no realizan una función administrativa o social determinada.

  Ejemplos: villa, país

- Casas de señores y particulares. Los santos (san Pedro, san Juan) se etiquetan con pers_señores
- Calles, Plazas y Puertas. La categoría incluye nombres de barrios.

Ejemplo: barrios de san Lorenzo.

- Oficios y Profesiones. Cuando una profesión va seguida el nombre de la persona, se incluye todo en la entidad de persona.
  Ejemplos: *poeta Moliere el Hipócrita*
  maestro Tirso de Molina