



# **Actividad 1: Corpus**

CE Inteligencia Artificial y Big Data  
Modelos de Inteligencia Artificial  
2024/2025

Daniel Marín López  
Guadalupe Luna Velázquez

# Índice

1. Introducción .....	3
2. Corpus Lingüísticos en Diferentes Idiomas .....	3
2.1. Inglés.....	3
2.2 Alemán .....	3
2.3 Italiano.....	4
2.4 Japonés.....	4
3. Herramientas de etiquetado .....	5
3.1. Generales .....	5
3.2 Por idioma .....	7

## 1. Introducción

El objetivo de esta práctica es llevar a cabo una investigación profunda sobre las últimas tendencias y avances en el campo del NLP multilingüe. A través de la exploración de diversas metodologías y herramientas, buscaremos identificar las mejores prácticas y oportunidades para innovar en el desarrollo de modelos de lenguaje. Esta investigación nos proporcionará una base sólida para abordar desafíos complejos en el procesamiento de textos y contribuir al avance de la inteligencia artificial.

## 2. Corpus Lingüísticos en Diferentes Idiomas

Los idiomas con los que trabajaremos serán **inglés, alemán, italiano y japonés**.

### 2.1. Inglés

El [British National Corpus \(BNC\)](#) es uno de los corpus lingüísticos más importantes y representativos del inglés británico contemporáneo. Este corpus fue compilado entre 1991 y 1994 y contiene una amplia muestra de la lengua inglesa en diferentes contextos y géneros. Su principal objetivo es servir como un recurso representativo y equilibrado del uso del inglés moderno, abarcando tanto la lengua escrita como la hablada.

El BNC incluye más de 100 millones de palabras, procedentes de una variedad de fuentes que reflejan la diversidad del inglés británico. Estas fuentes se dividen en categorías como literatura, periodismo, textos académicos, conversación espontánea, materiales profesionales y más.

### 2.2 Alemán

El [Corpus of Contemporary German \(Deutsches Korpus der Gegenwartssprache\)](#), también conocido como "DTA" (Digitales Textarchiv) es un corpus muy valioso con más de 500 millones de palabras, está gestionado por la Biblioteca Digital de la Universidad de Leipzig y proporciona acceso libre a varios textos y recursos lingüísticos.

El DTA contiene textos de ficción, literatura, periodismo y otros géneros del alemán contemporáneo, con un enfoque en el alemán estándar y es una excelente herramienta para quienes estudian el alemán contemporáneo y desean realizar investigaciones lingüísticas en este idioma.

The screenshot shows the homepage of the Deutsches Textarchiv (DTA). At the top, there is a navigation bar with links for 'Anmelden (DTA)', 'DWDS', 'dlexDB', and 'CLARIN-D'. Below this is a search bar with a 'suchen' button. The main content area is titled 'Deutsches Textarchiv (DTA)' and contains a description of the corpus, its scope, and its use. It mentions that the DTA is a platform for interdisciplinary and cross-genre collections of German texts, serving as a reference corpus for the modern German language. It also notes that the DTA includes a large number of titles (around 1500) and a vast amount of text (over 500 million words). The interface is clean and professional, with a focus on providing access to a large and diverse collection of German texts.

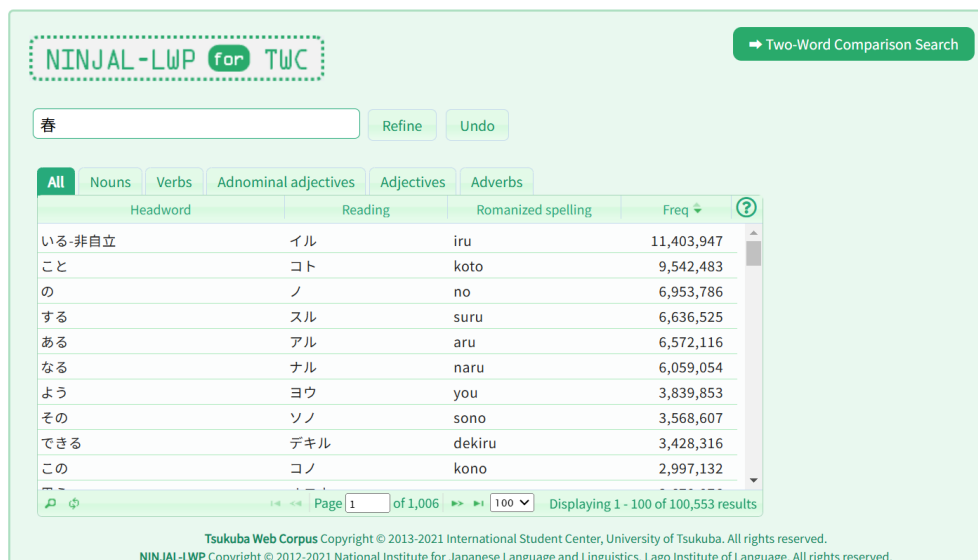
## 2.3 Italiano

El [Corpus PAISÀ](#) consiste en una amplia colección de textos auténticos contemporáneos en italiano extraídos de la web. Este corpus fue creado con el objetivo de proporcionar recursos para el aprendizaje del italiano mediante el estudio de materiales textuales auténticos, contiene aproximadamente 250 millones de palabras.

Los textos están anotados lingüísticamente, con lematización, etiquetado de partes del discurso y relaciones de dependencia sintáctica. Es una herramienta útil para el análisis del italiano moderno y está disponible de manera accesible en línea.

## 2.4 Japonés

El [NINJAL-LWP](#) (National Institute for Japanese Language and Linguistics - Linguistic Data Consortium) es un corpus lingüístico de referencia para el estudio del japonés, creado y mantenido por el Instituto Nacional de Lengua Japonesa (NINJAL).



The screenshot shows the NINJAL-LWP for TWC interface. At the top, there is a search bar containing the Japanese character '春' (haru). To the right of the search bar are buttons for 'Refine' and 'Undo'. Below the search bar is a table with columns: 'All', 'Nouns', 'Verbs', 'Adnominal adjectives', 'Adjectives', and 'Adverbs'. The table has four main columns: 'Headword', 'Reading', 'Romanized spelling', and 'Freq'. The table lists various Japanese words and their frequencies. At the bottom of the table, there is a pagination bar showing 'Page 1 of 1,006' and 'Displaying 1 - 100 of 100,553 results'.

All	Nouns	Verbs	Adnominal adjectives	Adjectives	Adverbs
Headword	Reading	Romanized spelling	Freq		
いる-非自立	イル	iru	11,403,947		
こと	コト	koto	9,542,483		
の	ノ	no	6,953,786		
する	スル	suru	6,636,525		
ある	アル	aru	6,572,116		
なる	ナル	naru	6,059,054		
よう	ヨウ	you	3,839,853		
その	ソノ	sono	3,568,607		
できる	デキル	dekiru	3,428,316		
この	コノ	kono	2,997,132		

Page 1 of 1,006 Displaying 1 - 100 of 100,553 results

Copyright © 2013-2021 International Student Center, University of Tsukuba. All rights reserved.  
NINJAL-LWP Copyright © 2012-2021 National Institute for Japanese Language and Linguistics, National Institute of Language. All rights reserved.

Este corpus contiene una amplia variedad de textos en japonés, tanto escritos como hablados, y está diseñado para reflejar el uso del idioma en contextos contemporáneos. Incluye textos de más de 50 millones de palabras, abarcando géneros como literatura, noticias, transcripciones de discursos y conversaciones, así como otros registros formales e informales.

Es especialmente útil para investigaciones sobre la morfología, sintaxis y pragmática del japonés y está anotado con información detallada, incluyendo etiquetado de partes del habla (POS), análisis de dependencias sintácticas y marcadores semánticos.

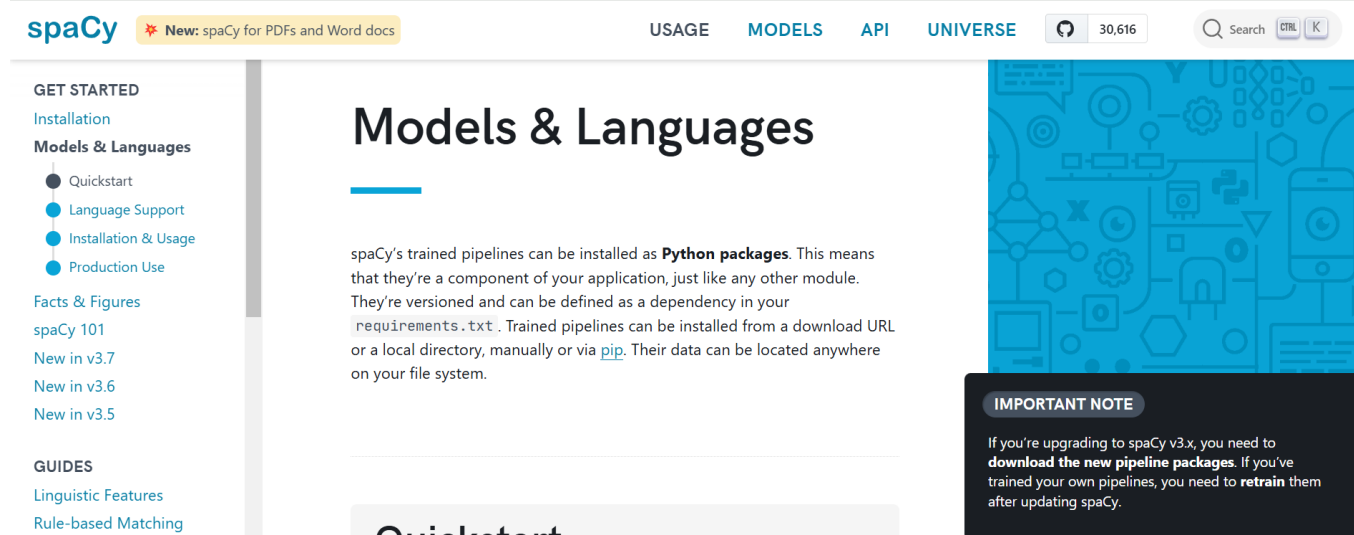
## 3. Herramientas de etiquetado

### 3.1. Generales

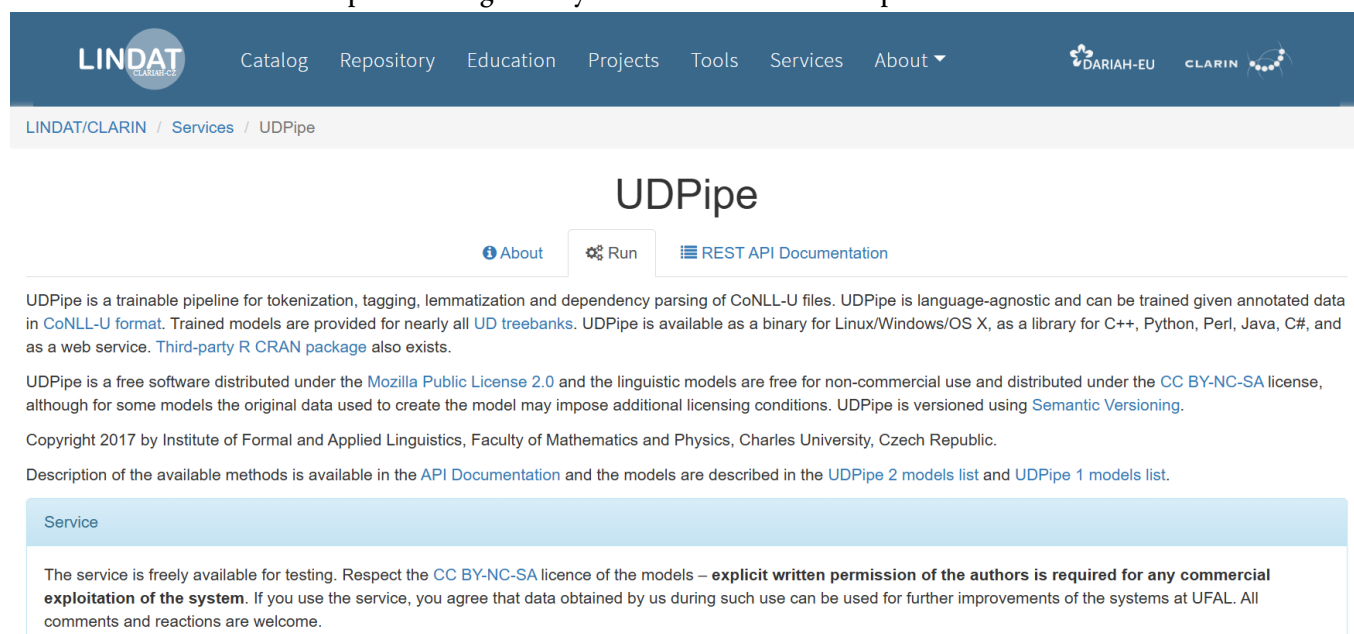
Algunas herramientas que soportan varios idiomas son:

- [Stanford NLP \(Stanza\)](#) es una biblioteca de procesamiento de lenguaje natural de última generación que ofrece un conjunto completo de herramientas para analizar texto en múltiples idiomas. Además de su capacidad para realizar etiquetado gramatical, análisis de dependencias y reconocimiento de entidades nombradas (NER), Stanza también puede llevar a cabo tareas como la tokenización, la lematización y la segmentación de oraciones. Estas funcionalidades permiten a los investigadores y desarrolladores extraer información valiosa de grandes volúmenes de texto, lo que la convierte en una herramienta esencial para aplicaciones como la búsqueda de información, la traducción automática y el análisis de sentimientos.

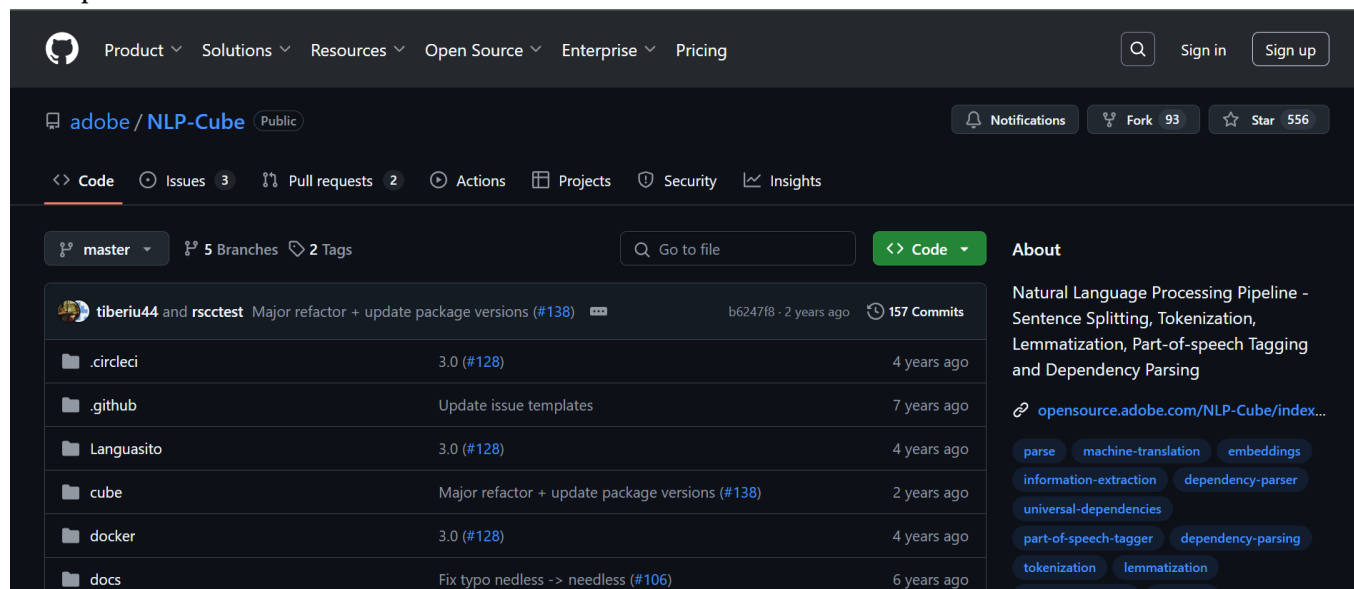
- [spaCy](#) es una biblioteca de Python de procesamiento de lenguaje natural (PLN) diseñada para ser extremadamente rápida y eficiente. Es ideal para construir aplicaciones de PLN en producción, gracias a su interfaz intuitiva y modelos preentrenados de alta calidad para inglés, alemán e italiano. spaCy destaca por su capacidad para realizar tareas como tokenización, lematización, reconocimiento de entidades nombradas (NER), análisis de dependencias y clasificación de textos de manera precisa y veloz. Además, su comunidad activa y su documentación detallada lo convierten en una excelente opción para desarrolladores de todos los niveles.



- [UDPipe](#) es una herramienta de procesamiento de lenguaje natural (PLN) altamente versátil que se basa en el proyecto Universal Dependencies. Gracias a su diseño multilingüe, UDPipe es capaz de analizar morfosintácticamente textos en una amplia variedad de idiomas. Este análisis, que descompone las oraciones en sus componentes más básicos (palabras y sus funciones gramaticales), es fundamental para una gran cantidad de tareas de PLN, como la traducción automática, la extracción de información y la generación de resúmenes. La capacidad de UDPipe para manejar múltiples idiomas y su precisión lo convierten en una herramienta esencial para investigadores y desarrolladores en el campo del PLN.



- [NLP-Cube](#) es una innovadora herramienta de procesamiento del lenguaje natural (NLP) diseñada para analizar de manera eficiente grandes volúmenes de texto en múltiples idiomas. Aprovechando el poder de las redes neuronales, NLP-Cube es capaz de extraer información valiosa de documentos, identificar patrones, analizar sentimientos y mucho más. Su arquitectura avanzada permite procesar textos complejos y ambiguos, ofreciendo resultados precisos y confiables. Además, su interfaz intuitiva y su capacidad para integrarse con otras herramientas lo convierten en una solución ideal para investigadores, empresas y cualquier persona que necesite extraer insights de sus datos textuales.



### 3.2 Por idioma

Aquí tenemos una tabla donde hemos clasificado varias herramientas que están enfocadas para un idioma concreto:

Idioma	Herramientas
Inglés	<a href="#">NLTK (Natural Language Toolkit)</a> , <a href="#">CoreNLP</a>
Alemán	<a href="#">GermaNER</a>
Italiano	<a href="#">Tint (The Italian NLP Tool)</a>
Japonés	<a href="#">MeCab</a> , <a href="#">Kuromoji</a> , <a href="#">GiNZA</a> (extensión de SpaCy)