1. **Con respecto al estado del arte de dataset que se han usado en el dominio medico en español, cuales tengo y cuantos me faltan.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Definidos para incporar al dataset de español. | Usados en MedLexSep | Usados en [1] | En [2] | [3] | En [4] |
| - **Wikipedia**  **- CodiEsp corpus.**  **- Cantemist corpus.**  **- The Chilean Waiting List Corpus.**  **- CT-EBM-SP .**  **- MedlinePlus Spanish (National Library of Medicine, NLM).**  - PharmaCoNER .  - **The Spanish Clinical Case Corpus (SPACCC).**  - **DisTEMIST.**  **- Spanish Drug Effect database.**  **Unified Medical Language System (UMLS).**  **- Spanish Biomedical Crawled Corpus**  **- MeSpEn\_Parallel-Corpora**  **- eHealthKD**[[1]](#footnote-1)  **- European Clinical Case Corpus.**  **- Biomedical Abbreviation Recognition and Resolution 2nd Edition (BARR2)**  **- CARES.**  **- IULA-SCRC**  **- IxaMed-GS**[[2]](#footnote-2)  **- LivingNER**  **- MEDDOCAN**  **- NUBes**  **- MESINESP2**  **- SocialDisNER** | - ~~Cancer Text Mining Shared Task (~~**~~CANTEMIST~~**~~)~~ (Biomedical Text Mining Unit): https://temu.bsc.es/cantemist/  - **~~Chilean Waiting List Corpus~~** ~~(Universidad de Chile)~~: https://zenodo.org/record/3926705  - ~~Clinical Case Coding in Spanish Shared Task (~~**~~CODIESP~~**~~) shared task (Biomedical Text Mining~~ Unit): https://temu.bsc.es/codiesp/  - ~~Clinical Trials for Evidence-Based Medicine in Spanish (~~**~~CT-EBM-SP~~**~~)~~ corpus: https://zenodo.org/record/6059737  - Diccionario de términos médicos (**Real Academia Nacional de Medicina de España**): https://dtme.ranm.es  - ~~MedlinePlus Spanish (~~**~~National Library of Medicine,~~ NLM**): https://medlineplus.gov/spanish/  - National Cancer Institute (**NCI**) Dictionary of Cancer Terms, Spanish version: https://www.cancer.gov/publications/dictionaries/cancer-terms  - Nomenclátor de Prescripció́n (AEMPS): https://listadomedicamentos.aemps.gob.es/prescripcion.zip  - Orphadata: Free access data from Orphanet: http://www.orphadata.org  - ~~Pharmacological Substances, Compounds and proteins and Named Entity Recognition (PharmaCoNER) challenge (Biomedical Text Mining~~ Unit): https://temu.bsc.es/pharmaconer/  - ~~Spanish Drug Effect database: https://github.com/isegura/ADR~~  - ~~Unified Medical Language System:~~ https://www.nlm.nih.gov/research/umls/index.html | **- CodiEsp corpus.**  **- Cantemist corpus.**   * **- The Chilean Waiting List Corpus.** * **- CT-EBM-SP** * - PharmaCoNER   **- The Spanish Clinical Case Corpus (SPACCC).**  - **DisTEMIST**  - **eHealthKD**  **- European Clinical Case Corpus**  **- Biomedical Abbreviation Recognition and Resolution 2nd Edition (BARR2)**  **- CARES**  **- IULA-SCRC**  **- IxaMed-GS**  **- LivingNER**  **- MEDDOCAN**  - **NUBes**  **- SocialDisNER** | - [SymTEMIST](https://zenodo.org/records/8413866) (Corpus of symptoms, signs and findings mentions and normalization to SNOMED CT, same document collection) | **Spanish Biomedical Crawled Corpus**  Clinical cases misc.  Clinical notes/documents  [Scielo](https://github.com/PlanTL-SANIDAD/SciELO-Spain-Crawler)  [BARR2\_background](https://temu.bsc.es/BARR2/downloads/background_set.raw_text.tar.bz2)  **Wikipedia\_life\_sciences**  Patents  [EMEA](http://opus.nlpl.eu/download.php?f=EMEA/v3/moses/en-es.txt.zip)  [mespen\_Medline](https://zenodo.org/record/3562536#.YTt1fH2xXbR)  Integra informacion de **Scielo** **y Pubmed**.  PubMed  <https://github.com/PlanTL-GOB-ES/lm-biomedical-clinical-es?tab=readme-ov-file>  <https://huggingface.co/PlanTL-GOB-ES/longformer-base-4096-biomedical-clinical-es>  https://huggingface.co/PlanTL-GOB-ES/longformer-base-4096-biomedical-clinical-es | BioAsq\_es: https://temu.bsc.es/mesinesp/index.php/datasets/ |

**Corpus que faltarían:**

* MedLexSep
* [EMEA](http://opus.nlpl.eu/download.php?f=EMEA/v3/moses/en-es.txt.zip)
* [mespen\_Medline](https://zenodo.org/record/3562536#.YTt1fH2xXbR)
* BioAsq\_es: <https://temu.bsc.es/mesinesp/index.php/datasets/>
* https://huggingface.co/datasets/LenguajeNaturalAI/casos\_clinicos\_tratamiento?row=21
* https://huggingface.co/datasets/LenguajeNaturalAI/casos\_clinicos\_diagnostico
* https://huggingface.co/datasets/head\_qa

1. Tengo un esquema de pre-entrenamiento con BioMistral y una idea de Finetuning con el mismo programado.
2. Falta el caso de Meditron en este esquema.
3. Falta usar el dataset de entrenamiento que busca los datos para el español usado por Meditron.
4. Falta escribir el articulo para el LatinX.
5. Falta ver como logro el esquema de comparación con los dataset que aparecen en HugginFace para el español.

https://huggingface.co/PlanTL-GOB-ES/longformer-base-4096-biomedical-clinical-es

**Referencia:**

[1] G. G. Subies, Á. B. Jiménez, y P. M. Fernández, «A Survey of Spanish Clinical Language Models», *ArXiv Prepr. ArXiv230802199*, 2023.

[2] M. A. Shaaban, A. Akkasi, A. Khan, M. Komeili, y M. Yaqub, «Fine-Tuned Large Language Models for Symptom Recognition from Spanish Clinical Text», *ArXiv Prepr. ArXiv240115780*, 2024.

[3] C. Aracena, N. Rodríguez, V. Rocco, y J. Dunstan, «Pre-trained language models in Spanish for health insurance coverage», en *Proceedings of the 5th Clinical Natural Language Processing Workshop*, 2023, pp. 433-438.

[4] A. V. Serrano, D. B. Sánchez, A. Segurado, G. G. Subíes, y Á. B. Jiménez, «BioMedIA: A Complete Voice-to-Voice Generative Question Answering System for the Biomedical Domain in Spanish».

1. No es muy sencillo. [↑](#footnote-ref-1)
2. Hay que contactar con los autores porque no es público. [↑](#footnote-ref-2)