The analysis of historical newspapers from the 18th, 19th, and early 20th centuries requires a certain quality of digitized sources and the use of specific domain or language resources. Any approach using current technologies finds that most of the NLP models available for transcription or entity recognition are trained with texts in "current languages". If, in addition, the challenge consists of extracting information from historical newspapers in Spanish, the complexity increases since the no

The analysis of historical newspapers from the 18th, 19th, and early 20th centuries requires a certain quality of digitized sources and the use of specific domain or language resources. Any approach using current technologies finds that most of the NLP models available for transcription or entity recognition are trained with texts in "current languages". If, in addition, the challenge consists of extracting information from historical newspapers in Spanish, the complexity increases since the n

BNE Digital Hemeroteca, Diario de Madrid (1788-1825)[1]

BNE Digital Hemeroteca, Diario de Madrid (1788-1825)[1]

INTELE

Infraestructura de Tecnologías del Lenguaje

Digital Humanities and Text Simplification Tasks: The CLARA-HD Project

Ana García Serrano, Antonio Menta, Eva Sánchez Salido

ETSI Informática, UNED

{agarcia, evasan}@lsi.uned.es, amenta@invi.uned.es

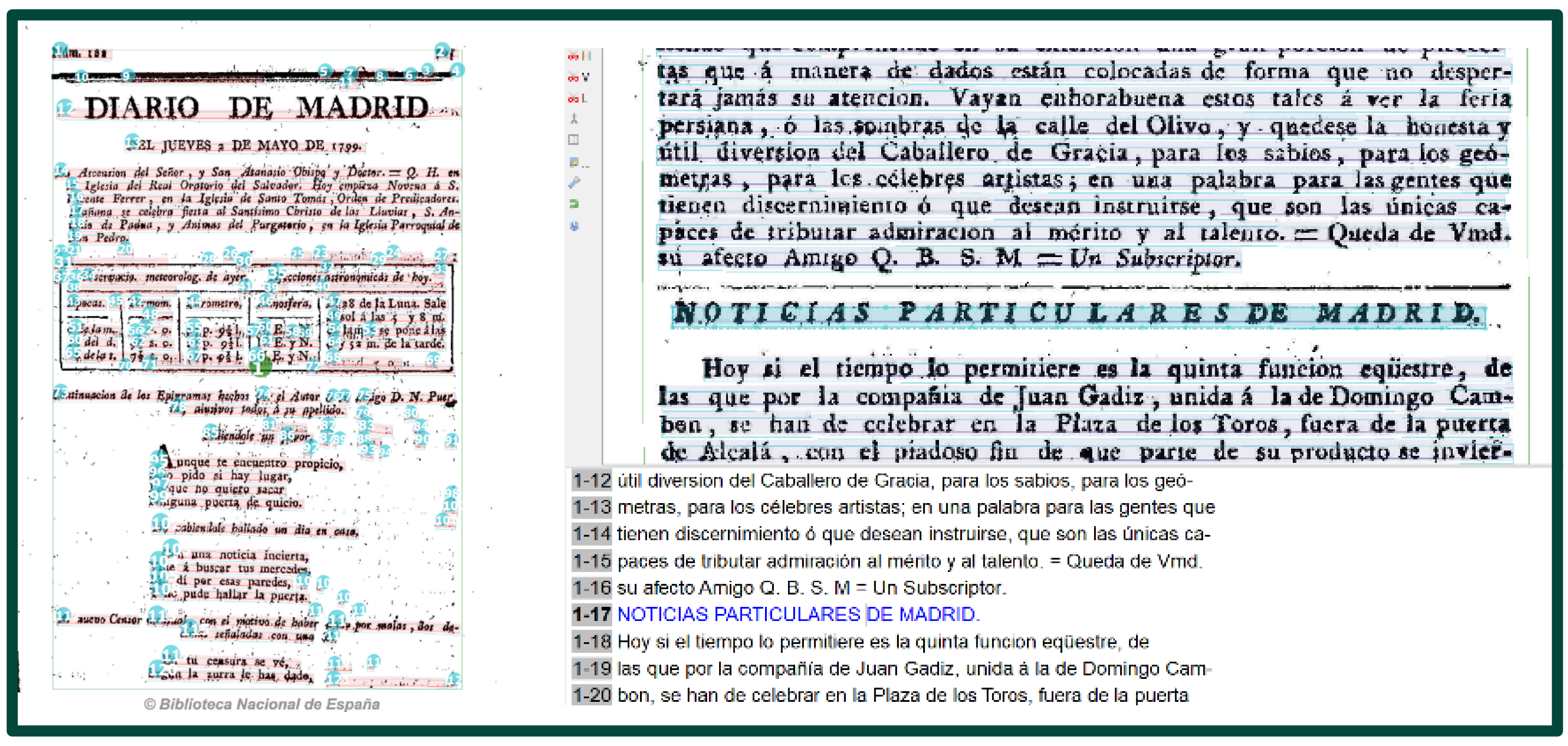
DESCRIPTION

The CLARA-NLP (www.clara-nlp.uned.es) integrates the experience of researchers from three institutions, UAM, UNED and CSIC, who have collaborated for more than a decade in different NLP projects. On this occasion **we explore the applications of automatic simplification of speciality text into easier-to-read text.** Therefore, the common basis will be the evaluation framework and the difference will be in the domain: financial (PID2020-116001RB-C31), ***digital humanities (PID2020-116001RB-C32)*** and medical (PID2020-116001RA-C33), where each group has a contrasting experience [1, 3, 4, 5, 8].

The project will develop a **software framework and infrastructure** to have easily available all the needed NLP tools and algorithms needed for the scientific work, and we will decide the inclusion in an international infrastructure as CLARIN or DARIAH. The experience and results will be described in detail in a MOOC developed by the three subprojects [9,10].

WORK DONE

The NLP&IR research group during this first year of the project is working in the domain of HD [2] with two main goals. First is the development of an annotated corpus from the Diario de Madrid [7] and its related interface to the support of historians partner works and second one is the design and experiment with different models to text simplification [6].



tas que á manera de dados están colocadas de forma que no despertará jamás su atencion. Vayan enhorabuena estos tales á ver la feria persiana, ó las sombras de la calle del Olivo, y quedese la honesta y útil diversion del Caballero de Gracia, para los sabios, para los geómetras, para los célebres artistas; en una palabra para las gentes que tienen discernimiento ó que desean instruirse, que son las únicas capaces de tributar admiracion al mérito y al talento. = Queda de Vmd. sú afecto Amigo Q. B. S. M. = Un Subscriptor.

NOTICIAS PARTICULARES DE MADRID.

Hoy si el tiempo lo permitiere es la quinta funcion eqüestre, de las que por la compañia de Juan Gadiz, unida á la de Domingo Camben, se han de celebrar en la Plaza de los Toros, fuera de la puerta de Alcalá, con el piadoso fin de que parte de su producto se invier-

1-12 útil diversion del Caballero de Gracia, para los sabios, para los geó-
1-13 metras, para los célebres artistas; en una palabra para las gentes que
1-14 tienen discernimiento ó que desean instruirse, que son las únicas ca-
1-15 paces de tributar admiración al mérito y al talento. = Queda de Vmd.
1-16 su afecto Amigo Q. B. S. M = Un Subscriptor.
1-17 NOTICIAS PARTICULARES DE MADRID.
1-18 Hoy si el tiempo lo permitiere es la quinta funcion eqüestre, de
1-19 las que por la compañía de Juan Gadiz, unida à la de Domingo Cam-
1-20 bon, se han de celebrar en la Plaza de los Toros, fuera de la puerta

SEPLN Conference 2022

The analysis of historical newspapers requires a certain quality of digitized sources and the use of specific domain resources. Any approach using current technologies finds that most of the models available for transcription or entity recognition are trained with texts in "modern languages". In our case, working with "Diario de Madrid 1788-1825", the complexity increases since the normalization of Spanish is relatively "modern". The steps followed using https://readcoop.eu/transkribus/ for the automatic transcription using a developed model reach 99% of performance.

REFERENCES

[1] Campillos-Llanos, L., *A. Terroba, S. Zakhir, A. Valverde and A. Capllonch,* **Building a comparable corpus and a benchmark for Spanish medical text simplification.** *Procesamiento del Lenguaje Natural* 69, 2022.

[2] Garcia Serrano, A.; Menta Garuz A. (2022) **La inteligencia artificial en las Humanidades Digitales: dos experiencias con corpus digitales** *Revista de Humanidades Digitales,* v.7, 19-39, ISSN 2531-1786. https://doi.org/10.5944/rhd.vol.7.2022.30928

[3] Garcia Serrano, Ana; Castellanos, Ángel (2017) **Representación y organización de documentos digitales: detalles y práctica sobre la ontología DIMH.** *Revista de Humanidades Digitales,* v.1, p.314-344, oct. 2017. ISSN 2531-1786. http://revistas.uned.es/index.php/RHD/article/view/17155

[4] Lara-Clares, Alicia; Garcia-Serrano, Ana; Rodrigo, Covadonga (2017). **Enrichment of Accessible LD and Visualization for Humanities: MPOC Model and Prototype**. *Research Conference on Metadata and Semantics Research.* Springer, 327–332.

[5] Lastra-Diaz, JJ; Goicoetxea, J; Taieb, MAH; Garcia-Serrano, A; Aouicha, MB; Agirre, E; Sanchez, D (2021) **A large reproducible benchmark of ontology-based methods and word embeddings for word similarity**. *Information Systems,* 96. PP: 1-17. ISSN: 0306-4379. https://doi.org/10.1016/j.is.2020.101636

[6] Menta-Garuz, Antonio; Garcia-Serrano, Ana; (2022). **Controllable Sentence Simplification Using Transfer Learning.** Simple Text Task at PAN - CLEF 2022 - Conference and Labs of the Evaluation Forum. CEUR-WS, vol 31-80, pp: 1-8. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http://ceur-ws.org/Vol-3180/paper-240.pdf

[7] Menta-Garuz, Antonio; Sanchez Salido, Eva; Garcia-Serrano, Ana (2022) **Transcripción de periódicos históricos: aproximación CLARA-HD.** *Demo en Conferencia SEPLN 2022.* CEUR-WS SEPLN-PD, pp: 1-4.

[8] Moreno-Sandoval, A., A. Gisbert and H. Montoro: "**Fint-esp: a corpus of financial reports in Spanish**" en *Multiperspectives in Analysis and Corpus Design,* Granada, Editorial Comares, 2020, pp. 89-102.

[9] Rodrigo, Covadonga; Iniesto, Francisco; García-Serrano, Ana (2020) **Reflections on Instructional Design Guidelines from the MOOCification of Distance Education: A Case Study of a Course on Design for All.** In book: UXD and UCD Approaches for Accessible Education. IGI Global, 2020. 21-37. Web. 7 Feb. 2020. https://www.igi-global.com/gateway/chapter/247870

[10] Rodrigo, Covadonga & Garcia-Serrano, Ana & Sánchez-Elvira Paniagua, Angeles. (2017). **Narrative design combined with a TAM survey to achieve a multisensorial museum user-experience for people with disabilities.** "The Online, Open Flexible Higher Education Conference. Higher Education for the future: Accelerating and strengthening innovation".Milton Keynes, UK. EADTU.

CLEF 2022

We propose the use of a pre-trained Deep Learning language model in a simplification task using its transfer learning capabilities. Because of the calculated features: text compression, word length, lexical and syntactic complexity, and the level of paraphrasing, the model has been able to simplify and obtain similar results to previous work, even without being trained directly on the domain data.

Our approach enables us to control the simplification result by selecting specific values for each of the features previously trained. This increases the flexibility of the system, as it is possible to generate different simplified versions. Code is released with an open-source license at https://github.com/Hisarlik/simpleTextCLEF.