DHQ: Digital Humanities Quarterly
Preview
2023
Volume 17 Number

Seeking Information in Spanish Historical Newspapers: The Case of *Diario de Madrid* (18th and 19th Centuries)

Eva Sánchez-Salido <evasan_at_lsi_dot_uned_dot_es>, ETSI Informática, UNED https://orcid.org/0000-0001-8665-3018
Antonio Menta <amenta_at_invi_dot_uned_dot_es>, ETSI Informática, UNED https://orcid.org/0000-0002-3172-2829
Ana García-Serrano <agarcia_at_lsi_dot_uned_dot_es>, ETSI Informática, UNED https://orcid.org/0000-0003-0975-7205

Abstract

New technologies for seeking information are based in machine learning techniques such as statistical or deep learning approaches that require a large number of computational resources as well as the availability of huge corpora to develop the applications that, in this concrete sub-area of Artificial Intelligence, are the so-called *models*. Nowadays, the reusability of the developed models is approached with fine-tuning and transfer learning techniques. When the available corpus is written in a language or domain with scarce resources, the accuracy of these approaches decreases, so it is important to address the start of the task by using state-of-the-art techniques.

This is the main problem tackled in the work presented here, coming from the art historians' interest in an image-based digitized collection of newspapers called *Diario de Madrid* (DM) from the Spanish press between 18th and 19th centuries, which is freely available at the Spanish National Library (BNE). Their focus is on information related to entities such as historical persons, locations as well as objects for sale or lost and others, to obtain geo-localization visualizations and solve some historical riddles. The first step needed technically is to obtain the transcriptions of the original digitalized newspapers from the DM (1788-1825) collection. After that, the second step is the development of a Named Entity Recognition (NER) model to label or annotate automatically the available corpus with the entities of interest for their research. For this, once the CLARA-DM corpus is created, a sub-corpus must be manually annotated for the training step in current Natural Language Processing (NLP) techniques, using human effort helped by selected computational tools. To develop the necessary annotation model (CLARA-AM), an experimentation step is carried out with state-of-the-art Deep Learning (DL) models and an already available corpus, which complements the corpus that we have developed.

A main contribution of the paper is the methodology developed to tackle similar problems like that of art historians' digitized corpus: selecting specific tools when available, reusing developed DL models to carry out new experiments in an available corpus, reproducing experiments in the art historians' own corpus and applying transfer learning techniques within a domain with few resources. Four different resources developed are described: the transcribed corpus, the DL-based transcription model, the annotated corpus and the DL models developed for the annotation using a specific domain-based set of labels in a small corpus. The CLARA-TM transcription model learned for the DM is accessible from January 2023 at the READ-COOP website under the title “Spanish print XVIII-XIX - Free Public AI Model for Text Recognition with Transkribus” (https://readcoop.eu/model/spanish-print-xviii-xix/).

1. Introduction

Corpora construction is a laborious task ([Gruszczyński et al. 2021]; [Ruiz Fabo et al. 2017]; [Wissler et al. 2014]), since it is desirable that the corpora may be used by different people and serve various purposes. Corpora are classified according to different parameters, such as subject matter, purpose, discourse modality and others, some of the most important being: balance, representativeness, or transparency ([Davies and Parodi 2022]; [Gebru et al. 2021]; [Torruella Casañas 2017]). The process of corpus construction is divided into different phases, ranging from the definition of its boundaries and purpose, pre-processing, storage, annotation, to name a few ([Aldama et al. 2022]; [Nakayama 2021]; [Calvo Tello 2019]; [Kabatek 2013]; [Rojo 2010]). 1

Corpus-based research has been dominated by statistical and neural models ([Moreno Sandoval 2019]; [Nieuwenhuijsen 2016]; [Rojo 2016]) until the end of the last decade, when Transformer-based models appeared [Vaswani et al. 2017], requiring the availability of very large corpora for training. Such massive corpora are scarcely available in the field of Humanities mainly because the corpora need to be annotated according to the interests of the corpus end-users, so it happens that the types of entities often vary according to the origin, language, domain, or purpose of the dataset. The Named Entity Recognition (NER) discipline, dealing with entities of interest, has evolved a lot since its beginnings in the first competitive NER task in 1996 [Grishman and Sundheim 1996], as many other tasks and 2 datasets have been created for evaluation.

Early NER systems made use of algorithms based on hand-built rules, lexicons and gazetteers, orthographic features or ontologies, among others, with the human cost of producing domain-specific resources that this entails [Li et al.2022]. Later, these systems were followed by those based on feature engineering and machine learning [Nadeau and Sekine 2007], which were the dominant technique in the task of named entity recognition until the first decade of the 2000s, that is, until the NLP revolution with the advent of neural networks. The most common supervised machine learning systems of this type that were used for NER include Hidden Markov Models (HMMs) [Bikel et al. 1997], Support Vector Machines (SVM) [Asahara and Matsumoto 2003], Conditional Random Fields (CRF) [McCallum and Li 2003], Maximum Entropy models (ME) [Borthwick et al. 1998], and decision trees [Sekine 1998]. 3

Starting with [Collobert et al. 2011], neural network-based systems are interesting because they do not require domain-specific resources such as lexicons or ontologies and are therefore more domain-independent [Yadav and Bethard 2018]. In this context, several neural network architectures were proposed, mostly based on some form of recurrent neural network (RNN) on characters [Lample et al. 2016], and embeddings of words or word components [Akbik, Blythe, and Vollgraf 2018]. Finally in 2017 the Transformer architecture was introduced in the paper *Attention Is All You Need* [Vaswani et al. 2017] which is here to stay, and its derivatives such as BERT [Devlin et al. 2019] and RoBERTa [Liu et al. 2019], which we make use of today, as well as the Spanish-based MarIA[1] or RigoBERTa[2]. 4

In the domain of Digital Humanities (DH), applying NER models to historical documents poses several challenges, as shown by other works in the domain ([Chastang, Torres Aguilar, and Tannier 2021]; [Kettunen et al. 2017]). This is a fertile field within NLP, since cultural institutions are carrying out digitization projects in which large amounts of images containing text are obtained ([Piotrowski 2012]; [Terras 2011]). One of the challenges is the margin of error still present in optical character recognition (OCR) systems [Boros et al. 2020], since historical documents are generally very noisy, contain smudges, and have different typographies that are generally unknown to the systems, which makes character recognition even more difficult. 5

Another challenge is related to the transfer of knowledge to new domains or languages, in this case to adapt NER models trained with datasets in current languages to old languages ([Baptiste et al. 2021]; [Bollmann 2019]; [De Toni et al. 2022]). Transfer learning techniques have become common practice in a wide range of tasks in NLP ([Hintz and Biemann 2016]; [Pruksachatkun et al. 2020]; [Zoph et al. 2016]), including the domain of DH such as works presented in the workshop on Natural Language Processing for Digital Humanities at NLPAI 2021 [Blouin et al. 2021] [Rubinstein and Shmidman 2021]). 6

Recent initiatives have been launched, such as the creation and annotation of datasets ([Ehrmann et al. 2022]; [Neudecker 2016]) and holding competitive events such as HIPE (*Identifying Historical People, Places and other Entities*) [Ehrmann et al. 2020a] addressing NER in historical texts. The general goals of HIPE include improving the robustness of systems, enabling comparison of the performance of NER systems on historical texts, and, in the long term, fostering efficient semantic indexing in historical documents. The HIPE2020 corpus shares objectives with the CLARA-DM corpus, since it involves the recognition of entities in historical newspapers, and has specific types of entities, that are slightly different from the general NER ones of localizations (LOC), persons (PER), organizations (ORG) and miscellanea (MISC). It is common practice in this domain that Historians, Linguists and Computer Science researchers collaborate to seek information on the domain-dependent entities and scenarios in which the historical research is focused ([Rivero 2022]; [Merino Recalde 2022]; [García-Serrano and Castellanos 2016]; [Calle-Gómez, García-Serrano, and Martínez, 2006]). We propose the use of the HIPE project approach with the aim of taking advantage of these resources and gaining perspective on the approach for our domain, a task for which we have little annotated data. 7

This paper is organized starting with a section dedicated to the corpus creation, transcription and sub-corpus manual annotation. Available tools are evaluated according to the properties of the original DM newspaper collection, the total amount of data contained and offered and required software to be installed. Afterwards, a section is devoted to the experimentation setting to perform the CLARA-DM corpus automatic annotation using deep learning models for NER. First, we carry out some experiments with a resource on multilingual historical newspapers such as the HIPE2020 corpus containing historical OCR texts in French, English and German [Ehrmann et al. 2020b]. One main goal is to identify which type of process adaptation and Transformer-based models will obtain the best results for the CLARA-DM corpus, based on the experience with HIPE2020. Then, we conduct some experiments on CLARA-DM dataset, in a transfer learning set-up (zero-shot and few-shot learning) between (1) three different sets of entity types: the general NER ones, the one used in the HIPE2020 experiments, and the domain specific ones used in CLARA-DM, and (2) languages (Spanish, English and German). Finally, in the last section we outline the conclusions drawn from the experiments and conclude the paper with some plans for future work. 8

2. Creation of the corpus CLARA-DM

A major goal of the CLARA-HD[3] collaborative project involving art historians, linguists, and computer scientists, is to speed up the art historians' research, through the transcription of the PDF newspapers' pages and the development of a robust model for recognition of named entities. The work started with a comparison of the performance of other NER systems on historical texts, following the 9

IMPRESSO[4] project and the HIPE series of shared tasks whose objective is to foster efficient semantic indexing on historical documents. The CLARA-HD project is based on the joint research of art historians, who are keen on investigating what is done and where it is done in the city of Madrid ([Molina Martín 2021]; [Cámara, Molina, and Vázquez 2020]; [Molina and Vega 2018]); the technicians, who are experts in Natural Language Processing and Deep Learning Models in applied research ([Menta and García-Serrano 2022]; [Sánchez-Salido 2022]; [Menta, Sánchez-Salido, and García-Serrano 2022]; [García-Serrano and Menta-Garuz 2022]), and the linguists, well-known experts in corpus creation, analysis, and exploitation ([Campillos-Llanos et al. 2022]; [Moreno Sandoval et al. 2018]).

The starting point of the CLARA-HD project was the analysis of the art historians' interest in the digitized collection of newspapers called *Diario de Madrid*, published between the 18th and 19th centuries and readily available at the BNE[5]. The focus of art historians is on information related to historical persons, locations but also objects for sale or lost, so the first step is the construction of a historical corpus from the original newspapers, the so-called CLARA-DM corpus, which involves designing a style guideline for transcription and a second style guideline for the annotation of named entities, which entails identifying the domain-based set of labels (semantic categories). The next step is the manual annotation of a sub-corpus to serve as the training corpus for an application development using current DL techniques. Finally, the development of an efficient model to recognise the specific named entities is tackled to annotate the CLARA-DM corpus automatically. 10

There are different tools to deal with any process in corpus management, from the document transcription, storage and search phase, annotation, analysis, and even corpus exploitation using Python libraries for information extraction. In this work we use Transkribus[6] for transcription, Tagtog[7] for annotation, and HuggingFace[8] for the implementation of Deep Learning models. In the remainder of this section the first two steps are detailed, and the third one will be tackled in the following section. 11

2.1. Transcription of the digitalized newspapers

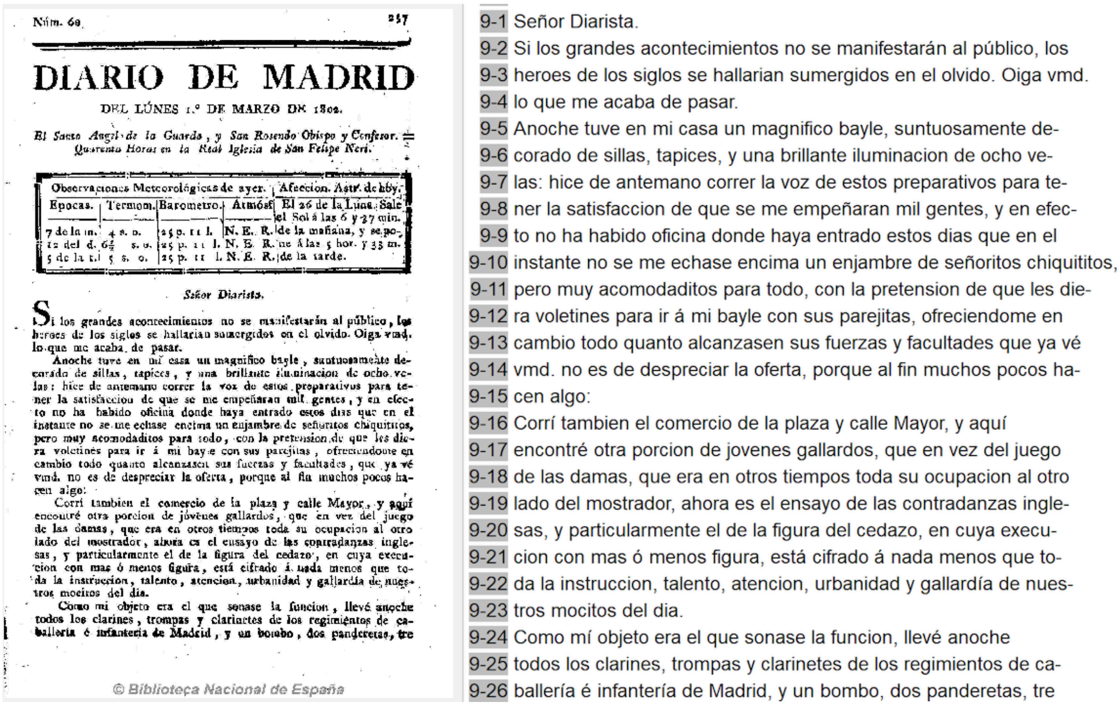
The first step building the corpus is the transcription of the texts, which we carried out using the Transkribus tool [Menta, Sánchez-Salido, and García-Serrano 2022]. The tool was selected once compared with commercial and open-source transcription tools as Amazon Textract[9] and Google's Tesseract[10], as the accuracy for old typography was slightly better, no fee for basic functionalities of the tool was required and has been used by other projects in ancient languages ([Kirmizialtin and Wrisley 2022]; [Aranda García 2022]; [Ayuso García 2022]; [Bazzaco et al. 2022]; [Alrasheed, Rao, and Grieco 2021]; [Derrick 2019]). Also, it provides functionalities both on the web browser and in the client version, and has proven to be very helpful for small DH research groups [Perdiki 2022]. 12

Transcription using the Transkribus tool starts by requiring the layout recognition, which consists of detecting the text regions and lines of text within the documents. This automatic process is not perfect, since the model does not recognise only text but also recognises some regions such as lines or spots that we have to remove manually. Furthermore, the main problem with newspapers is that there are tables or columns, and the model is not able to recognise the order in which it should read the pages. That is why we have to carry out a manual task to sort the text. It should be noted that we do this process because we make use of models that use sentences for training (Transformers for NER), and we also want the corpus to be used in the future for semantic and syntactic analysis. Note that the layout is not so relevant when only a word-based analysis is wanted. 13

Once we have carried out the structure recognition of the newspaper pages, the so-called layout recognition, we can move on to transcribing the text of the pages, either manually or using a public model, since the Transkribus tool contains public models that are trained with historical texts in different languages. We explored some of them, but they were still unable to recognise the text with some quality. The "Spanish Golden Age Theatre Prints 1.0" model ([Cuéllar and Vega García-Luengos 2021]; [Cuéllar 2021]) especially failed to recognise numbers or capital letters in our documents. So, what we propose to do is to develop our own transcription model (CLARA-TM) for the transcription of the documents in the CLARA-DM corpus. The first question was to find out how many pages we needed for a model in Transkribus to learn to transcribe automatically. The more training data, the better, but according to Transkribus guidelines, it is possible to start training the model with at least 25-75 pages manually transcribed, or less when working with printed instead of handwritten texts [11]. 14

We carried out several tests. In the first one we trained the model with 37 manually transcribed newspaper pages and obtained an error rate in character recognition (CER) of 4% in the validation set. This error decreases in successive tests with more amount and homogeneous training data. We realised that most errors in the transcription were caused by a lack of uniformity in the manual transcriptions, since they were carried out by different people. In order to make a more reliable transcription model, the transcribed pages were reviewed manually and a standardisation process wase carried out. This includes aspects such as the unification of the way fractions are transcribed (1/2 o ½), the inclusion or omission of symbols such as "=" (used before an author's signature), "&" or "§§", or the correction or not of typos. After this process, there is a high degree of homogeneity in the data that the model will see in its training, and therefore a higher probability of successful learning. After several tests (performed by changing the model parameters (number of epochs, learning rate and documents selected for training and validation) we trained the model with 193 transcribed pages and obtained 15 less than a 1% error, so we have kept it as our own transcription model for the CLARA-DM corpus. The obtained CER is a good one, slightly better than that obtained from public models that use similar resources to ours[12]. The pages used for training correspond to 37 different newspapers randomly chosen (between the ones downloaded, which covered the first day of every month between 1788-1825), since they all have a similar structure: 4-8 pages, the first one containing a table and both capital and lower case letters, and pages divided in columns at the end of the newspaper.

16 From now, the model can be applied to transcribe text automatically from any *Diario de Madrid* newspaper (see Figure 1) with a layout recognition carried out manually. The original collection has 13,479 newspapers (59,424 pages) and the CLARA-DM corpus currently has 589 newspapers with layout recognition done (2,474 pages), 37 newspapers manually transcribed (201 pages), 143 newspapers automatically transcribed (657 pages, containing an average of 1% of errors in the transcriptions), and 10 newspapers manually annotated (53 pages/24,843 tokens).



Núm. 60. 257

DIARIO DE MADRID

DEL LÚNES 1.º DE MARZO DE 1802.

El Santo Angel de la Guarda, y San Rosendo Obispo y Confesor. = Quarenta Horas en la Real Iglesia de San Felipe Neri.

Observaciones Meteorológicas de ayer.				Afeccion. Astr. de hoy.
Epocas.	Termom.	Barometro.	Atmósf.	El 26 de la Luna: Sale el Sol á las 6 y 37 min. de la mañana, y se pone á las 5 hor. y 33 m. de la tarde.
7 de la m.	4 s. o.	25 p. 11 l.	N. E. R.	
12 del d.	6½ s. o.	25 p. 11 l.	N. E. R.	
5 de la t.	5 s. o.	25 p. 11 l.	N. E. R.	

© Biblioteca Nacional de España

9-1 Señor Diarista.
9-2 Si los grandes acontecimientos no se manifestarán al público, los
9-3 heroes de los siglos se hallarian sumergidos en el olvido. Oiga vmd.
9-4 lo que me acaba de pasar.
9-5 Anoche tuve en mi casa un magnifico bayle, suntuosamente de-
9-6 corado de sillas, tapices, y una brillante iluminacion de ocho ve-
9-7 las: hice de antemano correr la voz de estos preparativos para te-
9-8 ner la satisfaccion de que se me empeñaran mil gentes, y en efec-
9-9 to no ha habido oficina donde haya entrado estos dias que en el
9-10 instante no se me echase encima un enjambre de señoritos chiquititos,
9-11 pero muy acomodaditos para todo, con la pretension de que les die-
9-12 ra voletines para ir á mi bayle con sus parejitas, ofreciendome en
9-13 cambio todo quanto alcanzasen sus fuerzas y facultades que ya vé
9-14 vmd. no es de despreciar la oferta, porque al fin muchos pocos ha-
9-15 cen algo:
9-16 Corrí tambien el comercio de la plaza y calle Mayor, y aquí
9-17 encontré otra porcion de jovenes gallardos, que en vez del juego
9-18 de las damas, que era en otros tiempos toda su ocupacion al otro
9-19 lado del mostrador, ahora es el ensayo de las contradanzas ingle-
9-20 sas, y particularmente el de la figura del cedazo, en cuya execu-
9-21 cion con mas ó menos figura, está cifrado á nada menos que to-
9-22 da la instruccion, talento, atencion, urbanidad y gallardía de nues-
9-23 tros mocitos del dia.
9-24 Como mí objeto era el que sonase la funcion, llevé anoche
9-25 todos los clarines, trompas y clarinetes de los regimientos de ca-
9-26 ballería é infantería de Madrid, y un bombo, dos panderetas, tre

Figure 1. Hardcopy of an original page (left) and its automatic transcription with CLARA-TM (right).

17 The human and computational resources spent for the development of our transcription model are quite important in the project planning and, given that the project is currently funded by the Spanish Government, the model is already published (January 2023) for free use at the Transkribus tool and website[13].

2.2. Sub-corpus manual annotation

18 Once we have the transcribed corpus, we can move on to the manual annotation step, but why do we need to do this? The datasets used for a NER task can be annotated or not, depending on whether the system to be developed is supervised or unsupervised. Since we are going to use Transformer-based systems for leading the current state of the art, we need annotated data to train them. On the other hand, a general NER system usually includes the general entities categories (tags/labels) Person, Place, Organisation and Others. However, in the CLARA-DM corpus of historical newspapers we need to define a different set of tags according to the needs of art historians, so that is the second reason why we have to carry out the task of manual annotation first to "teach" (train) the model. Furthermore, the presence of one or more types of encoding and annotations is a very important aspect in the possibilities of corpus exploitation [Rojo 2010].

19 To decide which annotation tool to use, an evaluation was made between four current leading annotation tools: Prodigy[14], Doccano[15], Brat[16] and Tagtog[17]. The comparison of their characteristics is shown in Table 1 (general functionalities; whether or not there are user management facilities; whether inter-annotator agreement is automatically calculated; if it is possible to automate tagging; input and output formats; operating system availability; availability of a web version; whether or not Python programming tools need to be installed locally, and finally if is for free, open-source and if user programming skills are required). The Tagtog tool was chosen because it is the only one that integrates a metric for viewing the agreement between annotators. In addition, it allows us to visualize it as it is annotated, which makes the annotation flow much more dynamic.

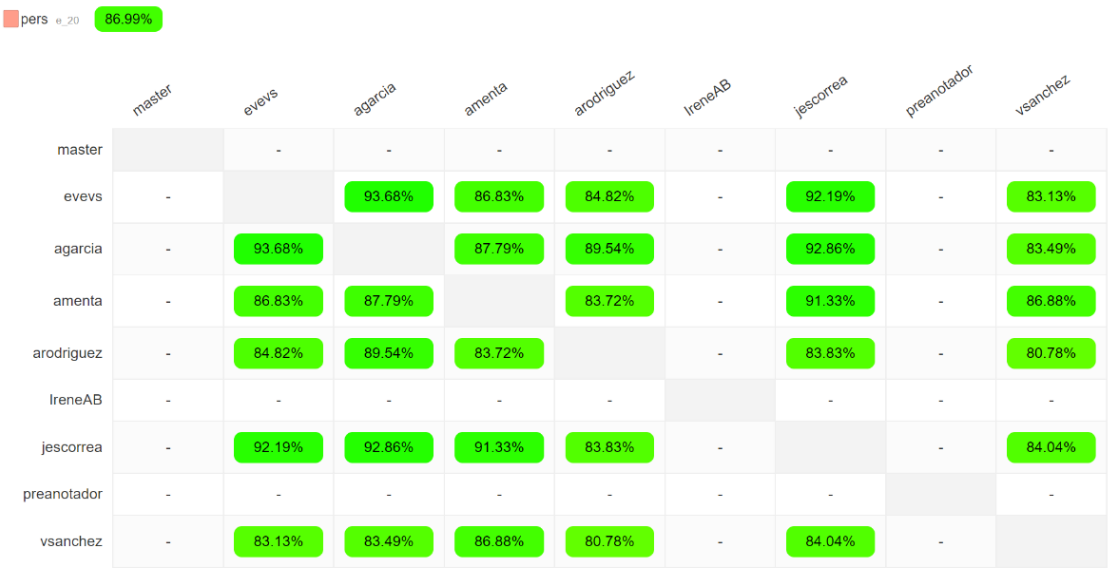
	Prodigy	Doccano	Brat	Tagtog
General functionalities	tagging text, images and videos and train models with the tagged data	text classification, sequence labelling, sequence-to-sequence tasks	entity and relationship annotation, lookups and other NLP derived tasks	entity and relationship annotation, document classification
User management	no	yes	yes	yes
Metrics for inter-annotator agreement	no	no	no	yes
Possibility to automate tagging	yes	no	no	yes (under subscription)
Input formats	TXT, JSONL, JSON, CSVand others	TXT, JSONL, CoNLL	TXT	TXT, CSV, source code files, URLs and others
Output formats	JSONL	JSONL	.ann	TXT, HTML, XML, CSV, ann.json, EntitiesTsv, others
Operating system	Windows, Mac and Linux	Windows, Mac and Linux	Mac or Linux (on Windows it is recommended a virtual machine)	Windows, Mac and Linux
Requires interacting with the command line	yes	yes	yes	no (on the web version)
Needs Python installed	yes (3.6+)	yes (3.8+)	yes (2.5+)	no
Programming skills required	familiarity with Python is desirable	no	knowledge of Linux and Apache servers is required	no (on the web version)
Open source	Partially	yes	yes	yes
Free	no	yes	yes	yes (different subscriptions)

Table 1. Comparison of tagging tools.

The process of the manual annotation of a corpus involves the construction of an “annotation guideline” in which the types of labels/tags are defined and specifications are given on what to annotate and what not to annotate, from where and to where to annotate, etc ([Campillos-Llanos et al. 2021]; [Moreno Sandoval et al. 2018]). Achieving a high annotation agreement enables the number of annotators in the future to be increased. 20

The process also includes the definition of the set of labels, that were defined together with art historians in several turns. On the one hand, the art historians know the information they need, but they do not have the perspective of the computer scientist who knows what kind of categories the model is able to learn to generalise. It is therefore a delicate task that requires an effort of understanding on both sides. When deciding on the set of labels, these can describe broad categories, such as persons, places, organisations, etc, or finer categories, such as streets and squares within the place category or nobles and lords within the person category. It is more convenient to start with a finer or more granular set of labels, as converting these categories into their corresponding broad versions is simpler than carrying out the reverse operation. 21

Based on the dialogue with the art historians and the documents they provided us with, a first proposal of taxonomy of entities was drawn up. It contained a wide set of entities and sub-entities (such as different kinds of religious places or houses) that is expected to be reduced in order to increase the quality of the automatic annotation. These entities are dumped into the annotation tool and the first annotation cycle was carried out. 22



🟥 pers e_20 86.99%

	master	evevs	agarcia	amenta	arodriguez	IreneAB	jescorrea	preanotador	vsanchez
master		-	-	-	-	-	-	-	-
evevs	-		93.68%	86.83%	84.82%	-	92.19%	-	83.13%
agarcia	-	93.68%		87.79%	89.54%	-	92.86%	-	83.49%
amenta	-	86.83%	87.79%		83.72%	-	91.33%	-	86.88%
arodriguez	-	84.82%	89.54%	83.72%		-	83.83%	-	80.78%
IreneAB	-	-	-	-	-		-	-	-
jescorrea	-	92.19%	92.86%	91.33%	83.83%	-		-	84.04%
preanotador	-	-	-	-	-	-	-		-
vsanchez	-	83.13%	83.49%	86.88%	80.78%	-	84.04%	-	

Figure 2. Inter-Annotator Agreement for Person entity class automatically calculated by the tool.

23 The first sub-corpus contains five newspapers with 28 pages, 928 sentences and 15,145 tokens, which are annotated by four annotators using a blind annotation process, that is, each person annotates the document independently without consulting the others, following the guidelines. The tool then calculates the Inter-Annotator Agreement (IAA) for each entity (see Figure 2). With these metrics, the quality of the labels is re-evaluated, and the taxonomy is adjusted. In the second round the tag taxonomy is as shown in Figure 3.

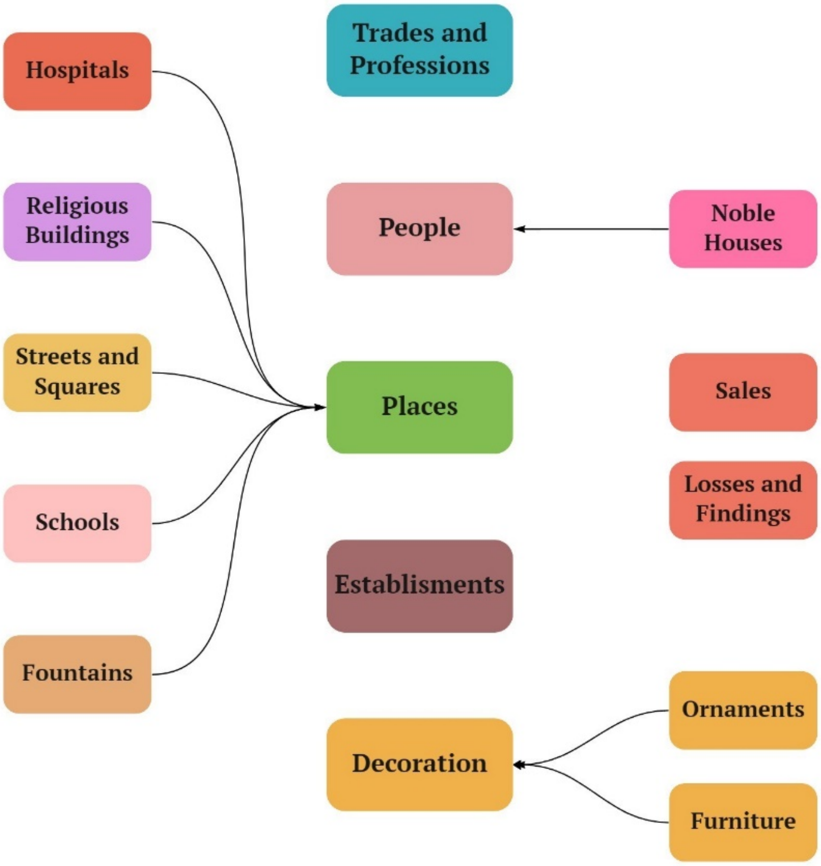
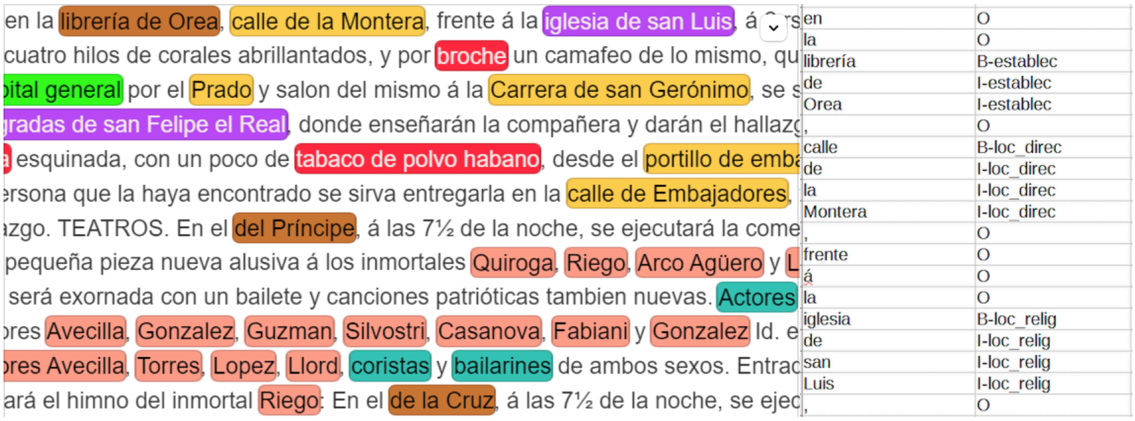


Figure 3. Taxonomy of entities.

24 Figure 4 shows an annotated text using the Tagtog tool that shares the colour legend classes shown at Figure 3: *iglesia de san Luis* is a

(multiword) entity denoting a religious building; *actores, coristas, bailarines* are professions.



en la librería de Orea, calle de la Montera, frente á la iglesia de san Luis, á
cuatro hilos de corales abrillantados, y por broche un camafeo de lo mismo, qu
ital general por el Prado y salon del mismo á la Carrera de san Gerónimo, se s
radas de san Felipe el Real, donde enseñarán la compañera y darán el hallaz
esquinada, con un poco de tabaco de polvo habano, desde el portillo de emba
rsona que la haya encontrado se sirva entregarla en la calle de Embajadores,
zgo. TEATROS. En el del Príncipe, á las 7½ de la noche, se ejecutará la come
pequeña pieza nueva alusiva á los inmortales Quiroga, Riego, Arco Agüero y L
será exornada con un bailete y canciones patrióticas tambien nuevas. Actores
res Avecilla, Gonzalez, Guzman, Silvostri, Casanova, Fabiani y Gonzalez Id. e
res Avecilla, Torres, Lopez, Llord, coristas y bailarines de ambos sexos. Entrac
ará el himno del inmortal Riego: En el de la Cruz, á las 7½ de la noche, se ejec

en	O
la	O
librería	B-establec
de	I-establec
Orea	I-establec
,	O
calle	B-loc_direc
de	I-loc_direc
la	I-loc_direc
Montera	I-loc_direc
,	O
frente	O
á	O
la	O
iglesia	B-loc_relig
de	I-loc_relig
san	I-loc_relig
Luis	I-loc_relig
,	O

Figure 4. Tagtog tool: hardcopy of a manually annotated text (left) and an excerpt codified in the IOB format (right).

Finally, once all the occurrences in the sub-corpus of historical newspapers are manually annotated/tagged, the Tagtog tool provides different export formats, including ann.json. After, the annotations output is converted into the IOB format in order to train the machine learning models (as described in the next section). To obtain the documents in IOB format, a new script is designed to transform the Tagtog EntitiesTsv output format into IOB format. At this moment the development of the deep learning system can start (the so-called DM model). 25

3. Towards an automatic annotation model for CLARA-DM

Once we have some CLARA-DM manually annotated (previously transcribed) newspapers (sub-corpus), some experiments to perform a NER task automatically reusing or transforming into a new one previously developed models in similar annotation settings can be carried out. The question is whether we can use one of the existing NER models or not, and how. 26

A brief explanation of the methodology and terminology of the experiments is as follows: 27

- First, we experiment with the HIPE2020 dataset (notice that we can use the terms corpus or dataset interchangeably, although the latter has a more computational nuance). We seek to evaluate the performance of monolingual and multilingual models (being the latter larger and trained in several languages), and to see if knowledge transfers between languages, through monolingual and/or multilingual models in a fine-tuning setup. Then we look at knowledge transfer between tasks, that is, we evaluate whether it is beneficial for a different task to use models trained with datasets for the general NER task (and therefore have different labels than HIPE2020).
- In a similar approach, we then experiment with the CLARA-DM dataset. First, we carry out some experiments without training the selected models (zero-shot experiments)we carry out some zero-shot experiments, that is, we evaluate on CLARA-DM dataset some models trained for general tags in a NER task or for NER on historical newspapers, and compare which setup achieves better results. After that, we train models with the CLARA-DM dataset and see if adding more historical training data improves the results. We include a preliminary qualitative error analysis on the results obtained for this first evaluation step based on HIPE2020 and CLARA-DM datasets.
- Then we carry out a second evaluation step, in which we evaluate several aspects. The first one is to study whether the method of adjudication for the final version of the manually annotated documents plays a role in the performance of the models (that is, when there are several annotators, there are different versions of the annotations and it is necessary to decide which label is the final one). The second one is a measurement of the performance based on the development of the annotation guideline versions, that is, the way in which the documents are annotated, and the availability of more documents annotated with the latest guidelines to see the gain in performance.

All the previous steps imply the selection of different available DL models to decide justifiably if we have to develop our own model, as we did for transcription. 28

The models used for the experiments are based on RoBERTa (monolingual) [Liu et al. 2019] and XLM-RoBERTa[18] (multilingual) [Conneau et al.2020]. Among the monolingual models we experiment with: 29

- two Spanish: RoBERTa-BNE[19] from the MarIA project [Gutiérrez-Fandiño et al. 2022] and BERTin-RoBERTa[20] from the BERTin project [De la Rosa et al. 2022],

- one English: DistilRoBERTa[21] [Sanh et al. 2019],
- one French: DistilCamemBERT[22] [Delestre and Amar 2022] and
- one German: GottBERT[23] [Scheible et al. 2020].

On the other hand, we use models that have been trained for a general set of NER tags: 30

- one monolingual for Spanish: RoBERTa-BNE-NER-CAPITEL[24],
- and two multilingual ones, one for Spanish (XLM-RoBERTa-NER-Spanish[25]) and another one trained in 10 languages with high resources (XLM-RoBERTa-NER-HRL[26]).

The working environment is a Google Colaboratory notebook, which provides a NVIDIA Tesla T4 GPU with 16GB of RAM and CUDA version 11.2. In addition, Transformers 4.11.3, Datasets 1.16.1, HuggingFace Tokenizers 0.10.3 and Pytorch 1.12.1+cu113 libraries are installed for running the experiments. 31

In the following sub-sections the subsequent experiments and related analysis of the results are described: 32

1) Using the HIPE2020 dataset, that contains different multilingual sub-corpus and entities annotated with specific tags, different from the general ones. Two different strategies are studied. The first one is the fine-tuning to observe both whether the monolingual training in French and German transfers to English, and if the multilingual model trained only with French or German improves in other languages not trained with. The second one is the evaluation of the transfer of knowledge from models using general tags to models with a different set of tags. 33

2) As the CLARA-DM dataset uses its specific set of tags, different from the HIPE2020 ones, two strategies are used for the first set of experiments: on the one hand, the use of models trained with external NER datasets (generalist or specific) on a zero-shot setup, and on the other hand, training with the CLARA-DM labelled data on a few-shot learning set-up. Some experiments use the CAPITEL dataset[27] from IberLEF2020 (the task is a general NER for Spanish). 34

3) After the discussion and conclusions on the first evaluation step, a new step for experiments is planned in order to evaluate (a) the method of adjudicating the final version of the manually annotated newspapers, (b) different aspects of the annotation guidelines (the way of annotating the classes and the total amount of tags), and (c) the amount of training data. 35

3.1. Experiments with HIPE2020

The HIPE2020 (*Identifying Historical People, Places and other Entities*) competitive event held at the CLEF conference, shares several objectives with the work presented as it focuses on the evaluation of NLP, information extraction and information retrieval systems. The HIPE2020 corpus [Ehrmann et al. 2020b] made available for experimentation in this competition is a collection of digitized historical documents in three languages: English, French and German. The documents come from the archives of different Swiss, Luxembourg and American newspapers. The dataset was annotated following the HIPE annotation guidelines [Ehrmann et al. 2020c], which in turn was derived from the Quaero annotation guidelines [Rosset, Grouin, and Zweigenbaum 2011]. The corpus uses the IOB format, providing training, test and validation sets for French and German, and no training corpus for English. The goal was to gain new insights and prospectives into the transferability of entity recognition approaches across languages, time periods, document types, and annotation tag sets. 36

The HIPE2020 corpus is annotated with the labels of Person, Place, Organization, Time, and Human Production. It contains 185 German documents totalling 149,856 tokens, 126 English documents totalling 45,695 tokens, and 244 French documents with 245,026 tokens. In total, they make up a corpus of 555 documents and 440,577 tokens. The dataset is pre-processed to recover the phrases that make up the documents and to be able to pass them to the models together with the labels, obtaining a total of 7,887 phrases in French (of which 5,334 correspond to the training -166,217 tokens-, 1,186 to validation and 1,367 to test), 5,462 sentences in German (of which 3,185 correspond to training -86,444 tokens-, 1,136 to validation and 1,141 to test), and 1,437 sentences in English (938 in validation and 499 in test). Note that the French training set is considerably larger than the German training set. 37

3.1.1. Fine-tuning

The first experiments consist of fine-tuning on the HIPE2020 dataset. When training a machine learning model there are several hyperparameters to be configured. The “number of epochs” is the number of times that the algorithm is going through the whole training dataset. The “batch size” is the number of training examples (in this case, sentences) used in one iteration. And the “learning rate” determines the pace at which an algorithm updates or learns the values of the parameters. The models’ hyperparameters are configured for a training in 3 epochs and a batch size of 12 in both the training and validation set, and a 5e-5 learning rate. The rest of the model configuration is set by default using the AutoConfig, AutoModel and AutoTokenizer classes of the Huggingface Transformers library. First, we fine-tune three monolingual models, which are shown in the first three rows of Table 2, and then the multilingual model, whose 38 results are shown in the last three rows. In both cases we train first with the French dataset, then with the German dataset, and thirdly with French and German jointly, since the English sub-corpus has no training dataset.

39 The evaluation metrics are based on precision, recall and F1. Briefly explained, precision is the relationship (fraction) of relevant instances among the retrieved instances, whilst recall is the fraction of relevant instances that were retrieved. The F1 measure is the harmonic mean of the precision and recall.

40 The objective of the first experiment is to analyse whether the knowledge learned on the NER training with the historical texts in French and German transfers to English. Secondly, whether the multilingual training with one language improves the performance in the other languages. We find that both claims hold true. The performance annotating the English sub-corpus of a model trained jointly in French and German improves compared to the performance of the models trained only with French or German (both when using the monolingual English model and the multilingual model) as shown with the results in the third row, that are better than the ones in the first and second rows, as well as the results in the sixth row, that are better than those in the fourth and fifth ones. Also, when training the multilingual model only with French or German, the result improves in the languages in which it has not been trained. For example, when training DistilCamemBERT with the French sub-corpus, the F1 in the German sub-corpus is 0.19, whilst when training XLM-RoBERTa with French, the F1 in German is 0.63.

41 Moreover, it is noteworthy that the multilingual model manages to equal or even improve the results of the monolingual models.

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

Table 2. Experiments with monolingual and multilingual models on French, German and English HIPE2020 datasets.

3.1.2. Transfer of knowledge with general NER datasets

42 In view of the usefulness of the multilingual model in the previous results, in the following experiments we use the multilingual model trained for NER in 10 languages with high resources.

	FR			DE			EN	
	P	R	F1	P	R	F1	P	R
XLM-R-ner-hrl	0.4	0.6	0.56	0.53	0.56	0.54	0.46	0.54
XLM-R-ner-hrl-**fr**	0.77	**0.82**	0.79	0.67	0.7	0.68	0.56	0.63
XLM-R-ner-hrl-**de**	0.71	0.73	0.72	0.73	0.77	0.75	**0.64**	0.57
XLM-R-ner-hrl-**fr+de**	**0.78**	0.68	**0.8**	**0.76**	**0.8**	**0.78**	0.6	**0.68**

Table 3. Evaluation on HIPE2020 of the model XLM-RoBERTa trained on 10 high resource languages for NER

43 The model is first evaluated on HIPE2020 without training (the so-called zero-shot learning), which is shown in the first row of the table. Then the model is trained with the HIPE2020 datasets, first French, then German and finally with both. Briefly, the zero-shot transfer learning means that we are taking a model trained for a specific or general task, and directly applying it on a different task or dataset for which the model has not been trained.

44 The results are slightly better than those obtained in the previous experiments, but in return we are evaluating on a general set of labels (Person, Location, Organization, and Dates), different from the one that HIPE2020 uses.

3.2. Experiments with CLARA-DM

45 With the insights we have extracted from the results of previous experiments, we move on to carrying out experiments with our developed dataset. To carry out these experiments we have the manually annotated sub-corpus of 5 newspapers, which have been obtained from the annotations of between 3 and 4 annotators and merged using the majority vote method. After a pre-processing phase

carried out using the spaCy[28] package to delimit the sentences that make up the newspapers, a dataset of 928 sentences is obtained, with a total of 15,145 tokens. The annotation guidelines were in a preliminary version, and the inter-annotator agreement was still to be improved. Therefore, at this point we tackle different experiments.

The CLARA-DM dataset has a large and original set of labels. This implies that, in order to obtain a specific NER model for the dataset, it will be necessary to have enough training data. We will adopt two strategies to carry out the first experiments, on the one hand, making use of models trained with external NER datasets (generalist or specific), and on the other hand, training with the available labelled data. 46

The set of labels of the dataset is extensive: it includes the generic labels of person, place, establishment, profession, ornaments, furniture, sales and losses or findings, and also the sub-labels person_lords (for nobles, high officials, etc), place_address (for streets, squares, gateways), place_religious (convents, parishes), place_hospital, place_college and place_fountain. In total, they make up a set of 14 tags, which when duplicated in the IOB format and together with the empty tag "O", add up to a total of 29 tags. This increases the complexity for the models to learn, and therefore the need for sufficient training data. On the other hand, in order to apply zero-shot learning, the names of the labels must be changed and simplified so that they are the same as those of the training datasets of the models, thus losing the more specific labels (such as religious places, ornaments or objects for sale) and drastically reducing the size of the set of labels, with the loss of information and efforts made during the labelling process that this entails. 47

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

Table 4. Label distribution in CLARA-DM.

The distribution of the tags in these documents is shown in Table 4. The class with by far the most examples is person, followed by address, profession, establishment, place, and lords. The classes of religious places, losses, schools, hospitals, and ornaments are notably underrepresented, and those of fountains and furniture do not even appear in the dataset (because they might not have been annotated by at least two people and therefore do not appear in the final version). 48

In the following we describe the experiments carried out to study the benefits of the transfer of knowledge between tasks and languages on our CLARA-DM dataset. 49

3.2.1. Zero-shot using CLARA-DM as test set

In these first two experiments we are not training with CLARA-DM but using it as a test set. This means that we are not evaluating on the set of labels of CLARA-DM but on the ones of the datasets that the models have been trained on. In Table 5 we evaluate two models trained for NER in Spanish, one multilingual and one monolingual. In Table 6, models trained with HIPE2020 (specific labels in historical newspapers) or with the CAPITEL dataset from IberLEF2020 (general NER for Spanish) are evaluated in CLARA-DM. 50

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

Table 5. Evaluation on CLARA-DM of general NER models for Spanish.

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

Table 6. Evaluation on CLARA-DM with models trained with HIPE2020.

We find that in both cases the monolingual models are slightly better. In the case of training with HIPE2020, it turns out to be more beneficial to train only with French, than to add English and German, since it is the most similar language to Spanish. Moreover, it is better to train with the French sub-corpus of HIPE2020, that tackles the same task (NER in historical newspapers), than training with 51

CAPITEL, which tackles general NER for Spanish, so the task influences the model.

3.2.2. Few-shot/fine-tuning on CLARA-DM

In these experiments we train with the few data we have manually annotated (that is why the experiments are few-shot learning and not zero-shot learning) using 3 newspapers as training, 1 as validation and 1 as test (containing approximately 700 sentences for training, 120 for validation and 110 for test). On Table 7 the results of fine-tuning several models are shown, the best one being the Spanish monolingual BERTin model. 52

	P	R	F1
XLM-R**-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

Table 7. Fine-tuning on CLARA-DM.

Even if we consider the corpus very small in comparison with the HIPE2020 one, (700 sentences versus 7,900 in the French sub-corpus) it turns out that with only 3 newspapers for training, similar results are achieved to those of directly evaluating the models trained with HIPE2020 shown in Table 6. 53

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

Table 8. Training and evaluation in CLARA-DM of models trained with HIPE2020 and CAPITEL.

Lastly, in Table 8 the training with CLARA-DM has been combined with training with HIPE2020 (only the French part) and CAPITEL. Again, we have obtained the best results of all experiments until now, but at the cost of evaluating on a different set of labels than that of CLARA-DM and therefore wasting the annotation effort. 54

This is the only case in which the performance of monolingual and multilingual models are very similar. Here it is interesting to note that, again, it gives better results to train with the French HIPE2020, which contains historical newspapers, than with CAPITEL, which is Spanish for generic NER in general domains. CAPITEL[29] contains texts after 2005 on the following topics: Science and technology; Social sciences, beliefs and thought; Politics, economy and justice; Arts culture and shows; Current affairs, leisure and daily life; Health and Others. 55

At this point, it is worth noting that the labels that each model had in its (first) training in each experiment are as follows: 56

in Table 5, the tags for the experiment XLM-R-ner-spanish are general person, location, organization, and miscellaneous, and those of the experiment RoBERTa-bne-ner-capitel are those of CAPITEL, that is person, location, organization and others (in BIOES format instead of BIO),

in Table 6 all the experiments have the labels of HIPE2020, except for RoBERTa-bne-ner-capitel-fr, which has those of CAPITEL,

in Table 7 the labels are those of CLARA-DM and

in Table 8 the labels are those of HIPE2020 or CAPITEL, whichever comes first.

Since we have seen that models trained with a different set of labels (HIPE2020 or CAPITEL ones) are able to predict with some quality the labels that they have in common with the CLARA-DM dataset, we can do the opposite experiment. In order not to lose the wide range of labels in CLARA-DM, we can first train the models adding fictitious tags, so that in the second training with CLARA-DM we include all the classes. 57

For example, regarding the XLM-RoBERTa model: when fine-tuning it first with French HIPE2020 and after with CLARA-DM, we obtained metrics of around 60% (first row of Table 8), but we were evaluating only on the tags Person, Place and Organization, which are the ones HIPE2020 has in common with the CLARA-DM corpus. If we change the classes in the first training with HIPE2020 by 58 adding the ones present in CLARA-DM, and fine-tune first with HIPE2020 and then with CLARA-DM, we get the metrics shown in Table 9. Performance drops by 14% on average, but in return we are evaluating the corpus on the whole CLARA-DM tagset, with a model trained with both CLARA-DM and HIPE2020.

	P	R	F1
XLM-R**-fr-clara**	0.44	0.51	0.47

Table 9. Evaluation on CLARA-DM of a model trained first with HIPE and then with CLARA-DM, with the tags of CLARA-DM corpus.

Comparing this performance with that obtained when only training with CLARA-DM (Table 7) (since here we are training with both HIPE2020 and CLARA-DM), the results are quite similar, so apparently there is no real added value when adding the first training with HIPE2020, that is, using more resources. 59

3.2.3. Qualitative analysis

As a preliminary qualitative analysis, or error analysis, Table 10 shows the performance per label of the best model fine-tuned with CLARA-DM, which was BERTin (Table 7). 60

It is interesting to note that the results are consistent with the current state of the annotation guidelines, since entities such as persons, locations and religious places have a high degree of inter-annotator agreement, above 70%, being those that obtain the best metrics, while others such as establishments or objects for sale still need to be revised through the guideline and the model also has a harder time identifying them correctly. This seems to be even more relevant than the inner imbalance of the dataset, since classes such as religious places do not have many appearances or occurrences, but the model recognises them with a high degree of accuracy. 61

		P	R	F1	support
Establishment	establec	0.23	0.31	0.26	29
Place or Location	loc	0.79	0.71	**0.75**	21
Place - College	loc_cole	0.00	0.00	0.00	1
Place - Address	loc_direc	0.42	0.78	0.55	23
Place - Religious	loc_relig	0.80	0.67	**0.73**	6
Losses or Findings	perdida_hallazgo	0.00	0.00	0.00	0
Person	pers	0.53	1.00	0.70	8
Person - Lords	pers_señores	0.56	0.64	0.60	14
Trades and Professions	prof	0.61	0.67	0.64	21
Sales	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

Table 10. Metrics of each label with BERTin model fine-tuned on CLARA-DM.

3.3. First evaluation step discussion

In the experiments with HIPE2020 we have observed that the use of multilingual models is beneficial when we have datasets in several languages for the NER task, as they allow the knowledge to be transferred to languages with fewer or no resources. Furthermore, we have seen that including domain-generic datasets slightly improves the results, but at the cost of evaluating on a different set of labels, and therefore wasting the efforts of the tagging procedure. 62

In spite of not having a particularly robust or sophisticated model, and although very precise results were not one main goal of the work, results of around 80% on the F1 measure for French and German, and 65% for English, which has no training data, have been achieved (previous Section 3.1). 63

As regards the experiments with CLARA-DM, in general, better results have been obtained with the monolingual models in Spanish, except when we have trained jointly with CLARA-DM and HIPE2020 datasets, in which the multilingual model has been on a par with the monolingual ones. It has also been shown that the similarity between Spanish and French favours the transfer of knowledge within the same domain, and that this transfer is even better than training with a generic NER dataset in the same language. 64

An important conclusion for corpora of languages or domains with scarce resources has also been initially contrasted: the importance of the inter-annotator agreement over the dataset imbalance. 65

In addition, as with three annotated newspapers, the results of fine-tuning with CLARA-DM achieve similar results than evaluating in CLARA-DM a model trained on HIPE2020, even if, the joint training with HIPE2020 and CLARA-DM has not given rise to a great improvement in results. That is shown in Table 6 (zero-shot) were models trained with HIPE2020 are evaluated in CLARA-DM with results of around 50% F1, while in Table 7 (fine-tuning in CLARA-DM) a 50% F1 is also achieved just by training with only 3 newspapers. So, it is for sure that with more annotated documents, good results can be expected. 66

The results are susceptible to improvement, since the quality of the annotation guidelines is still to be enhanced, and so the inter-annotator agreement, which will lead to have more quality and homogeneous data. 67

From this analysis, the plan for the second evaluation step is to try to improve the models obtained in this first evaluation step, and to measure the gain in performance, once we have more robust annotation guidelines, and more annotated newspapers. As will be described in the next section, the experiments are based on the three models used for fine-tuning in CLARA-DM (Table 7) since it has been shown that training with more added datasets is not so beneficial, but it is better to have more data in CLARA-DM. 68

3.4. Second evaluation step

In order to confirm the previous results, in this series of experiments the following parameters are evaluated: the method of adjudicating the final version of the manually annotated newspapers, aspects of the annotation guidelines (the way of annotating the classes and the total amount of tags), and the amount of training data. 69

In Tagtog, when several users annotate the same document, as a result, there are different annotation versions. Adjudication is the process of resolving inconsistencies between these versions before promoting a version to master (final version). In other words, the different annotators' versions are merged into one, (using various strategies). Adjudication can either be manual (when a reviewer promotes a version to master) or automatic[30], based on different adjudication methods such as the IAA (or Best Annotators) or the Majority Vote. Automatic adjudication based on Best Annotators means that for each single annotation task, the annotations of the user with the best IAA are promoted to master. The goal is to have the best annotations available for each annotation task in master version. Furthermore, automatic adjudication by Majority Vote means that for each single annotation, it is promoted to master only if it was annotated by over 50% of the annotators. 70

First, an experiment is carried out with the same documents as before but obtained with a different adjudication method. While in the first experiments the final version was obtained by the Majority Vote method, in this case the Best Annotators method is used. 71

Then, the progress of the annotation guidelines is evaluated, as well as the gain in performance with a bigger number of annotated newspapers. 72

In these experiments we will limit ourselves to carrying out experiments exclusively with the CLARA-DM dataset (not HIPE2020, CAPITEL, etc) and with the models used in the previous few-shot experiments. The newspapers to be used will be those of the first experiments and also new annotated newspapers. 73

3.4.1. Evaluation of the adjudication method

By carrying out the experiments in Table 7, that is, fine-tuning with CLARA-DM, but instead with the final annotations obtained by the Best Annotators method, we get the results shown in Table 11. 74

	P	R	F1
XLM-R**-clara**	0.47	0.53	0.50
RoBERTa-bne**-clara**	0.49	0.55	0.52
BERTin-R**-clara**	0.37	0.47	0.41

Table 11. Fine-tuning with CLARA-DM (the same experiments as in Table 7), but with final version of the annotations obtained with the Best Annotators method.

With RoBERTa-BNE model the F1 measure improves from 0.46 to 0.52, accuracy from 0.42 to 0.49, and recall from 0.50 to 0.55. And with XLM-RoBERTa, the F1 measure improves from 0.46 to 0.5, precision from 0.41 to 0.47 and recall from 0.52 to 0.53. However, with the BERTin model, which achieved the best results in Table 7, the F1 measure has decreased from 0.52 to 0.41, accuracy from 0.48 to 0.37, and recall from 0.58 to 0.47. 75

That is, by changing the adjudication method, the best performing method has changed, even though the F1 measure of 0.52 is still not surpassed. 76

3.4.2. Progress of annotation guidelines and availability of more training data

The annotation guidelines are adjusted in several turns, by analysing both the Inter Annotator Agreement and the performance of the models. 77

Eleven new newspapers were annotated in accordance with the new guidelines. In particular, the place_fountains entity is deleted, since we consider it from now on included within the furniture. Also, place_college and place_hospital tags are deleted (and included in establishments), since the three entities had very few mentions in the newspapers. Finally, the category Organization (administrative bodies) is created, to differentiate it from that of Establishments (commerce, leisure, services and others) and to be in line with other common annotation guidelines. 78

All in all, we get a set of 12 classes: two for people (person_general, person_lord) three for places (place_general, place_address, place_religious), establishments, organizations, professions, and four for objects (ornaments, furniture, sales, losses/findings), having the taxonomy shown in Figure 5. 79

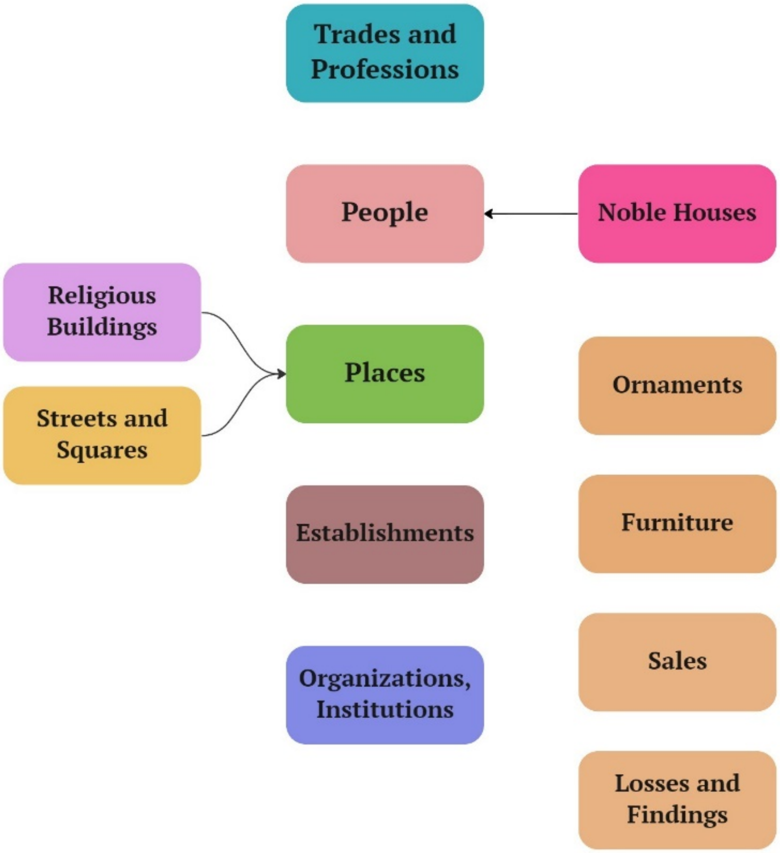


Figure 5. Final label taxonomy for CLARA-DM.

In intermediate steps, some different ways of labelling were evaluated. For example, at some point we agreed to annotate the profession of a person within the *person* tag whenever they appeared contiguously (as in *Sr. D. Josef de la Cruz y Loyola, Gobernador de dicho Real Sitio*). However, this proves ambiguous, and we confirm that is it better to label more concrete and nuclear entities, since it is clearer for annotators, and thus improves the IAA, and in turn leads to better performance of the models on the classes with better IAA. 80

The results of fine-tuning the models with the 11 new annotated newspapers, annotated with the final guidelines and the final version obtained with the Best Annotators method, are shown in Table 12. In this case we used 7 newspapers for training, that contained 1,228 sentences, which nearly doubles the number of sentences that we had for the first experiments. 81

	P	R	F1
XLM-RoBERTa-clara	0.74	0.79	0.76
RoBERTa-bne-clara	**0.75**	**0.80**	**0.78**
BERTin-R-clara	**0.75**	0.77	0.76

Table 12. Training and evaluation in CLARA-DM with more newspapers and new guidelines.

While the results of the first fine-tuning had a performance of around 50% in all the metrics (Table 7), with the updated annotation guidelines and double the number of sentences for the training, metrics of more than 75% have been achieved. It is also observed that when the IAA improves for a specific class, so the models get to predict it better, even when there are fewer examples of the class. 82

		P	R	F1	support
Places or Locations	loc	0.89	0.84	0.87	50
Places — Streets and Squares	loc_direc	0.82	0.94	0.87	111
Places – Religious Buildings	loc_relig	0.64	0.62	0.63	26
Organizations, Institutions	org_adm	0.53	0.63	0.58	27
Establishments	org_establec	0.67	0.60	0.63	50
Persons	pers	0.81	1.00	0.89	216
Persons - Lords	pers_señores	0.58	0.40	0.47	55
Ornaments	prod_ador	0.80	0.57	0.67	7
Furniture	prod_mobil	0.00	0.00	0.00	1
Losses or Findings	prod_perdida-hallazgo	0.82	0.75	0.78	12
Sales	prod_venta	0.62	0.57	0.59	14
Trades and professions	prof	0.66	0.67	0.66	78
	Micro avg	0.75	0.80	0.78	647
	Macro avg	0.65	0.63	0.64	647
	Weighted avg	0.74	0.80	0.77	647

Table 13. Metrics of each label with RoBERTa-BNE.

If we take a look at the performance per entity class as in Table 13, one noticeable aspect is that entities that do not contain proper names are usually harder to predict, such as products or professions. We might consider tagging these making use of predefined lists instead of the NER model. Furthermore, sometimes there are very few occurrences of these classes (such is the case of the furniture in Table 13) and this affects the performance as a whole. 83

On the other hand, the results are also very conditioned by the choice of training and test data, since we do not yet have enough examples. Even so, it is shown that both stronger annotation guidelines and the availability of more documents improves the performance of the models. 84

4. Proposed methodology

As a summary of the methodology used in this work, the following is a list of the necessary steps necessary to carry out a process of recognition of named entities in historical documents.As a summary of the methodology used in this work, the steps followed are listed below. 85

- **Digitization:** The first step is to digitize the historical documents if they are not already in a digital format. This can be done using scanners or digital cameras. This step has already been carried out by the BNE in this work.
- **Optical Character Recognition (OCR):** Once the documents have been digitized, OCR software is used to convert the text-based images into machine-encoded text. This stage may involve manual error correction to deal with the inaccuracies of the OCR process, particularly with historical documents that may have faded or smudged ink or unusual typography. As explained in previous sections, we used a layout recognition model and trained our own model in Transkribus to obtain the text.
- **Annotation guidelines:** The initial stage in training a NER model involves establishing rules for accurate entity identification. It's crucial to explicitly define the types of named entities that the model is expected to recognize.
- **Annotation:** The entities to be recognized by the NER process are then annotated. This involves marking up the text with

tags that indicate the type of entity. This is usually done manually by human annotators, using annotation software.

- **Validation:** The annotated text is then validated to ensure the accuracy of the entity recognition and annotation processes. This can involve a review by human annotators, or the use of validation software that compares the annotated text to a gold standard or benchmark. In this work, we used Tagtog for annotation and validation.
- **Training the model:** The final step is to train a named entity recognition model from the annotated and validated text. This involves the separation of the data into a training set, a validation set and a test set. This is followed by the selection of the most appropriate model and finally by training the model several times with different hyperparameters.
- **Evaluation:** The performance of the NER process should be evaluated using appropriate metrics such as precision, recall, and F1-score. This helps to identify areas for improvement and guide future work.

The methodology employed in this research can be extended to other domains and languages, facilitating advancements in various fields. For instance, by leveraging multilingual models and adapting them to other languages, similar NER tasks in historical newspapers from different countries can be accomplished. Additionally, applying the developed annotation guidelines and expanding the dataset to include newspapers from diverse regions and time periods would enable the automatic annotation and prediction of entities in a broader range of Spanish newspapers, and potentially other languages. This adaptability and transferability of the methodology make it a valuable resource for historians and researchers working on various textual collections beyond historical newspapers, such as ancient manuscripts, legal documents, or literary works in different languages. By scaling up the annotated data and training the models with more diverse samples, the accuracy and robustness of the NER models can be further enhanced, fostering more efficient and accurate analyses in the digital humanities and beyond. 86

5. Conclusions and future work

The characteristics of digitized newspapers and the research interests of historians justify the need to develop the CLARA-DM corpus, a model for transcription, and a specific model for named entity recognition in these texts. 87

As regards named entity recognition, we have seen that cross-language and cross-domain knowledge is transferred not only with multilingual models, but also with monolingual ones. On the other hand, having corpora or datasets for the same specific task (NER in historical newspapers in this case) in other languages might be useful and is even better than having generic datasets in the same language. 88

In the developed CLARA-DM dataset, the monolingual models and the use of a dataset of historical newspapers in French have stood out because it is a language close to Spanish. 89

Nevertheless, the use of external datasets could not compete with having more annotated data of our own corpus. In the final experiments we found that an improvement in the annotation guidelines and an increase in the labelled data significantly improves the performance of the models. In addition, it is verified that the models are sensitive to choices such as the method of adjudication for the final version of the annotations, or the choice of data for training and testing. 90

We believe that the entity recognition system can be improved in future research by using the Simple Knowledge Organisation System (SKOS) and creating an ontology. Based on the trained models, and after reviewing the discovered entities by historians, we propose to create a taxonomy. This will be used to improve the identification of new mentions in other newspapers. In addition, storing information in an ontology will allow more complex queries at run time by reducing the ambiguity of entities. For example, having all the mentions of a particular location in a single URI would allow you to queringqueryingy all the people who have traded there. 91

In summary, from a collection of newspapers in PDF format it has been possible to obtain a model for its transcription, with an accuracy of 99%, and a model for the prediction of the entities in the transcribed newspapers, with an accuracy of more than 75%. This last result will be improved in the future as we trained the model with only 7 newspapers. Furthermore, as we already have better-defined annotation guidelines we hope to speed up the annotation process and get an even more robust final NER model trained with more data to annotate automatically any other Spanish newspaper published in the same period.In conclusion, the methodology employed in this research can be extended to other domains and languages, facilitating advancements in various fields. For instance, by leveraging multilingual models and adapting them to other languages, similar NER tasks in historical newspapers from different countries can be accomplished. Additionally, applying the developed annotation guidelines and expanding the dataset to include newspapers from diverse regions and time periods would enable the automatic annotation and prediction of entities in a broader range of Spanish newspapers, and potentially other languages. This adaptability and transferability of the methodology make it a valuable resource for historians and researchers working on various textual collections beyond historical newspapers, such as ancient manuscripts, legal documents, or literary works in different languages. By scaling up the annotated data and training the models with more diverse samples, the accuracy and robustness of the NER models can be further enhanced, fostering more efficient and accurate analyses in the digital humanities and beyond. 92

Acknowledgements

This work has been supported by the Spanish Government and funded under the CLARA-HD project (PID2020-116001RB-C32, https://clara-nlp.uned.es/home/dh/). 93

We would like to acknowledge the support of the UNED art history professors Álvaro Molina and Alicia Cámara, research directors of the CARCEM[31]project. Finally, warm thanks to Andrés Rodríguez-Francés and Víctor Sánchez-Sánchez members for a while of UNED research group NLP&IR > DH. 94

Appendix: Brief introduction to Deep Learning

Artificial Intelligence (AI) is the field of study that focuses on the creation of computer systems and software capable of performing tasks that require human intelligence. This covers areas ranging from speech recognition and computer vision capabilities, to complex decision making, machine learning and problem solving. 95

There are different approaches within AI, such as rule-based AI, which uses a set of predefined instructions and rules to make decisions, and machine learning, which is based on algorithms and models that allow machines to learn from examples and data. 96

Machine learning is a sub-discipline of AI based on the idea of building mathematical or statistical models that can learn from data. These models are trained using a training data set, where examples are provided with their respective labels or expected results. The machine learning algorithm analyses the data and adjusts its internal parameters to find patterns and correlations between input features and output labels. 97

Once the model has been trained, it can be used to make predictions or decisions about new data that have not been used during training. The goal of machine learning is to generalise the knowledge acquired during training so that it can be applied to new and unknown situations. 98

There are several types of machine learning, each focusing on different approaches and techniques to address specific problems. There are three main types: supervised and unsupervised learning, and reinforcement learning. 99

In supervised learning, the algorithm is provided with a training data set consisting of input examples and the corresponding outputs, and the goal for the algorithm is to learn to map the inputs to the correct outputs. In unsupervised learning, the algorithm is confronted with a set of unlabelled training data. The objective is to find patterns, structures or intrinsic relationships in the data. In reinforcement learning, the algorithm interacts with a dynamic environment and receives feedback in the form of rewards or punishments based on its actions, and learns through trial and error, adjusting its behaviour to maximise rewards over time. 100

Deep Learning is a branch of machine learning that relies on artificial neural networks to learn and extract high-level representations from complex, unstructured data. 101

Two essential stages in deep learning are pre-training and training/fine-tuning. Pre-training involves training a model on a related task or a large dataset to learn general features and patterns. For example, large language models are pre-trained in huge corpora such as Wikipedia. Then, fine-tuning follows, where the model's parameters are adjusted on a smaller labeled dataset related to the specific target task (i.e. NER). This process allows the model to leverage prior knowledge from pre-training and adapt to the target task, leading to improved performance. This is a popular approach in Deep Learning called transfer learning ([Malte and Ratadiya 2019]; [Ruder et al. 2019]), especially when dealing with limited labeled data, besides the usual supervised, unsupervised and reinforcement learning approaches. 102

In the context of transfer learning, zero-shot and few-shot learning are approaches that leverage pre-trained models to address limited data scenarios. Zero-shot learning aims to recognize new classes unseen during training by utilizing semantic relationships or embeddings learned from related classes. This allows the model to make predictions on entirely novel categories without any fine-tuning on specific examples. On the other hand, few-shot learning focuses on learning from a few examples of each new class. The model adapts its knowledge from pre-training to recognize and generalize to new classes with only a small amount of labeled data. These techniques significantly enhance the capabilities of transfer learning, enabling models to excel in situations with minimal labeled data and effectively tackle new and previously unseen tasks. 103

Hyperparameters, such as epochs and learning rate, are crucial settings in deep learning models that are not learned from the data during training. Instead, they are set before training begins and can significantly impact the model's performance. "Epochs" represent the number of times the model iterates through the entire dataset during training. Increasing epochs can allow the model to see the data more times but may risk overfitting. "Learning rate" controls the step size for updating the model's parameters during training. A high learning rate can lead to faster convergence, but it might cause overshooting and instability. Balancing these hyperparameters is essential to achieve optimal training and ensure the model generalizes well to new, unseen data. 104

105 Two types of systems make use of Deep Learning in this paper: OCR and NER. Optical Character Recognition (OCR) is a technology that utilizes neural networks and computer vision techniques to automatically recognize and extract text from images or scanned documents. Deep learning models, such as Convolutional Neural Networks (CNNs), are employed to learn the complex features of characters and words, enabling accurate text recognition. Named Entity Recognition (NER) systems are a type of natural language processing (NLP) technology that uses deep learning and machine learning techniques to automatically identify and classify named entities in text. NER systems employ models, such as recurrent neural networks (RNNs) or transformer-based architectures like BERT, to learn the patterns and context of words in sentences, allowing them to recognize and label named entities accurately.

Notes 106

[1] https://www.bne.es/es/noticias/1111-el-primer-sistema-masivo-de-inteligencia-artificial-de-la-lengua-espanola-maria

[2] https://www.iic.uam.es/inteligencia-artificial/procesamiento-del-lenguaje-natural/modelo-lenguaje-espanol-rigoberta/

[3] CLARA-HD project (PID2020-116001RB-C32) is funded by MCIN - AEI (AEI/10.13039/501100011033).

[4] https://impresso-project.ch

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

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

[7] https://tagtog.com

[8] https://huggingface.co

[9] https://aws.amazon.com/es/textract

[10] https://github.com/tesseract-ocr/tesseract

[11] https://readcoop.eu/transkribus/howto/how-to-transcribe-documents-with-transkribus-introduction/

[12] See *German Fraktur 19th-20th century* (https://readcoop.eu/model/german-fraktur-19th-20th-century/), *French newspapers late 18th century – midth of 20th century* (https://readcoop.eu/model/french-newspapers-late-18th-century-midth-of-20th-century/), *Transkribus Print Multi-Language* (https://readcoop.eu/model/print-multi-language-danish-dutch-german-finnish-french-latin-swedish/) or *Dutch newspapers 17th century* (https://readcoop.eu/model/dutch-newspapers-17th-century/).

[13] https://readcoop.eu/model/spanish-print-xviii-xix

[14] https://prodi.gy

[15] https://doccano.herokuapp.com

[16] http://brat.nlplab.org

[17] https://www.tagtog.com/

[18] https://huggingface.co/xlm-roberta-base

[19] https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne

[20] https://huggingface.co/bertin-project/bertin-roberta-base-spanish

[21] https://huggingface.co/distilroberta-base

[22] https://huggingface.co/cmarkea/distilcamembert-base

[23] https://huggingface.co/uklfr/gottbert-base

[24] https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne-capitel-ner

[25] https://huggingface.co/MMG/xlm-roberta-large-ner-spanish

[26] https://huggingface.co/Davlan/xlm-roberta-base-ner-hrl

[27] https://sites.google.com/view/capitel2020

[28] https://spacy.io

[29] https://plantl.mineco.gob.es/tecnologias-lenguaje/comunicacion-formacion/eventos/eventosinfoday2019/Aspectos%20destacados%20del%20Plan%20TL/corpus-anotado-Jordi-Porta.pdf

[30] https://docs.tagtog.com/collaboration.html#automatic-adjudication

[31] https://dimh.hypotheses.org/author/dimh

Works Cited

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

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

Alrasheed, Rao, and Grieco 2021 Alrasheed, N., Rao, P. and Grieco, V. (2021) “Character Recognition Of Seventeenth-Century Spanish American Notary Records Using Deep Learning”, *DHQ* 15.4. Available at: http://www.digitalhumanities.org/dhq/vol/15/4/000581/000581.html

Aranda García 2022 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. Available at: https://doi.org/10.13136/2284-2667/1107

Asahara and Matsumoto 2003 Asahara, M., and 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. Available at: https://aclanthology.org/N03-1002

Ayuso García 2022 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. Available at: https://doi.org/10.13136/2284-2667/1102

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

Bazzaco et al. 2022 Bazzaco, S., Ruiz, A. M. J., Ruberte, Á. T., and 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. Available at: https://doi.org/10.13136/2284-2667/1190

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

Blouin et al. 2021 Blouin, B., Favre, B., Auguste, J., & Henriot, C. (2021). Transferring Modern Named Entity Recognition to the Historical Domain: How to Take the Step? *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, 152-162. https://aclanthology.org/2021.nlp4dh-1.18

Bollmann 2019 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. Available at: https://doi.org/10.18653/v1/N19-1389

Boros et al. 2020 Boros, E., Hamdi, A., Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Sidere, N., and 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. Available at: https://doi.org/10.18653/v1/2020.conll-1.35

Borthwick et al. 1998 Borthwick, A., Sterling, J., Agichtein, E., and 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. Available at: https://aclanthology.org/M98-1018

Calle-Gómez, García-Serrano, and Martínez, 2006 Calle-Gómez, Javier; García-Serrano, Ana and Martínez, Paloma. (2006). “Intentional processing as a key for rational behaviour through Natural Interaction”, *Interacting With Computers*, Vol: 18 Nº: 6, pp:1419-1446 10.1016/j.intcom.2006.05.002

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

Campillos-Llanos et al. 2021 Campillos-Llanos, L., Valverde-Mateos, A., Capllonch-Carrión, A. et al. (2021) “A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine.” BMC Med Inform Decis Mak 21, 69 Available at: https://doi.org/10.1186/s12911-021-01395-z

Campillos-Llanos et al. 2022 Campillos-Llanos, L., Terroba Reinares, A. R., Zakhir Puig, S., Valverde-Mateos, A., and Capllonch-Carrión, 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)*.

Chastang, Torres Aguilar, and Tannier 2021 Chastang, P., Torres Aguilar, S., and Tannier, X. (2021). “A Named Entity Recognition Model for Medieval Latin Charters.” *DHQ* 15.4. Available at: http://www.digitalhumanities.org/dhq/vol/15/4/000574/000574.html

Collobert et al. 2011 Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011) “Natural Language Processing (Almost) from Scratch.” *Natural Language Processing*, 45.

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

Cuéllar 2021 Cuéllar, Álvaro. (2021). “Spanish Golden Age Theatre Manuscripts (Spelling Modernization) 1.0”. Transkribus.

Cuéllar and Vega García-Luengos 2021 Cuéllar, Álvaro and Vega García-Luengos, Germán. (2021) “ETSO. Estilometría aplicada al Teatro del Siglo de Oro.” etso.es.

Cámara, Molina, and Vázquez 2020 Cámara, Alicia; Molina, Álvaro y Margarita A. Vázquez. (2020). Manassero (eds.). “La ciudad de los saberes en la Edad Moderna”, Gijón, Ediciones Trea, 296 pp. Available at: http://e-spacio.uned.es/fez/view/bibliuned:404-Amolina-1011.

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

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

De la Rosa et al. 2022 De la Rosa, J., Ponferrada, E. G., Villegas, P., Salas, P. G. de P., Romero, M., and Grandury, M. (2022) “BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling” (arXiv:2207.06814). ArXiv. Available at: http://arxiv.org/abs/2207.06814

Delestre and Amar 2022 Delestre, C., and Amar, A. (2022) “DistilCamemBERT: A distillation of the French model CamemBERT” (arXiv:2205.11111). ArXiv. Available at: https://doi.org/10.48550/arXiv.2205.11111

Derrick 2019 Derrick, T. (2019) “Using Transkribus For Automated Text Recognition of Historical Bengali Books” *British Library Digital Scholarship Blog*. Available at: https://blogs.bl.uk/digital-scholarship/2019/08/using-transkribus-for-automated-text-recognition-of-historical-bengali-books.html

Devlin et al. 2019 Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019) “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” (arXiv:1810.04805). ArXiv. Available at: http://arxiv.org/abs/1810.04805

Ehrmann et al. 2020a Ehrmann, M., Romanello, M., Fluckiger, A., and Clematide, S. (2020a). “Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers.” 38.

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

Ehrmann et al. 2020c Ehrmann, M., Watter, C., Romanello, M., and Clematide, S. (2020c). “Impresso Named Entity Annotation Guidelines”. Available at: https://doi.org/10.5281/zenodo.3604227

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

García-Serrano and Castellanos 2016 García-Serrano, A. and Castellanos, A. (2016) “Representación y organización de documentos digitales: detalles y práctica sobre la ontología DIMH”. *Revista de Humanidades Digitales*, v.1, 314-344, ISSN 2531-1786. Available at: https://doi.org/10.5944/rhd.vol.1.2017.17155

García-Serrano and Menta-Garuz 2022 García-Serrano, A., and Menta-Garuz, A (2022). “La inteligencia artificial en las Humanidades Digitales: dos experiencias con corpus digitales.” *Revista de Humanidades Digitales*, 7, 19-39.

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

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

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

Gutiérrez-Fandiño et al. 2022 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., and Villegas, M. (2022) “MarIA: Spanish Language Models”, 22.

Hintz and Biemann 2016 Hintz, G., and Biemann, C. (2016). “Language Transfer Learning for Supervised Lexical Substitution.” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 118-129. https://doi.org/10.18653/v1/P16-1012

Kabatek 2013 Kabatek, J. (2013) “¿Es posible una lingüística histórica basada en un corpus representativo?” *Iberoromania*, 77(1).

Available at: https://doi.org/10.1515/ibero-2013-0045

Kettunen et al. 2017 Kettunen, K., Mäkelä, E., Ruokolainen T., Kuokkala J. and Löfberg, L. (2017) “Old Content and Modern Tools – Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910.” *DHQ* 11.3. Available at: http://digitalhumanities.org:8081/dhq/vol/11/3/000333/000333.html

Kirmizialtin and Wrisley 2022 Kirmizialtin, Suphan and David Joseph Wrisley. (2022) “Automated Transcription of Non-Latin Script Periodicals: A Case Study in the Ottoman Turkish Print Archive.” *DHQ* 16.2. Available at: http://www.digitalhumanities.org/dhq/vol/16/2/000577/000577.html

Lample et al. 2016 Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016) “Neural Architectures for Named Entity Recognition” (arXiv:1603.01360). ArXiv. Available at: https://doi.org/10.48550/arXiv.1603.01360

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

Liu et al. 2019 Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019) “RoBERTa: A Robustly Optimized BERT Pretraining Approach” (arXiv:1907.11692). ArXiv. Available at: https://doi.org/10.48550/arXiv.1907.11692

Malte and Ratadiya 2019 Malte, A., and Ratadiya, P. (2019). *Evolution of transfer learning in natural language processing* (arXiv:1910.07370). arXiv. https://doi.org/10.48550/arXiv.1910.07370

McCallum and Li 2003 McCallum, A., and 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. Available at: https://aclanthology.org/W03-0430

Menta and García-Serrano 2022 Menta, A., and García-Serrano, A. (2022) “Controllable Sentence Simplification Using Transfer Learning.” Proceedings of the Working Notes of CLEF.

Menta, Sánchez-Salido, and García-Serrano 2022 Menta, A., Sánchez-Salido, E., and 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)*. Available at: https://ceur-ws.org/Vol-3224/paper17.pdf

Merino Recalde 2022 Merino Recalde, David. (2022) “El sistema de personajes de las comedias urbanas de Lope de Vega. Propuesta metodológica y posibilidades del análisis de redes sociales para el estudio del teatro del Siglo de Oro” Master Thesis, UNED. Facultad de Filología. Departamento de Literatura Española y Teoría de la Literatura. Available at: http://e-spacio.uned.es/fez/view/bibliuned:master-Filologia-FILTCE-Dmerino

Molina Martín 2021 Molina Martín, Á. (2021) “Cartografías del adorno en las residencias nobiliarias de la corte de Carlos IV: redes y modelos de buen gusto y distinción” *Magallanica. Revista de Historia Moderna*, 7(14), 205-235.

Molina and Vega 2018 Molina, Á., and Vega, J. (2018) “Adorno y representación: escenarios cotidianos de vida a finales del siglo XVIII en Madrid”, 139-166.

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

Moreno Sandoval et al. 2018 Moreno Sandoval, A., Díaz García, J., Campillos Llanos, L., and Redondo, T. (2018) “Biomedical Term Extraction: NLP Techniques in Computational Medicine”. Available at: https://doi.org/10.9781/ijimai.2018.04.001

Moreno Sandoval, Gisbert, and Montoro Zamorano 2020 Moreno Sandoval, Antonio, Gisbert, Ana and Montoro Zamorano, Helena. (2020) “FinT-esp: A corpus of financial reports in Spanish”.

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

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

Neudecker 2016 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. Available at: https://aclanthology.org/L16-1689

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

Perdiki 2022 Perdiki, Elpida. (2022) “Review of ‘Transkribus: Reviewing HTR training on (Greek) manuscripts’.” *RIDE 15*. doi: 10.18716/ride.a.15.6. Accessed: 21.12.2022. Available at: https://ride.i-d-e.de/issues/issue-15/transkribus/

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

Pruksachatkun et al. 2020 Pruksachatkun, Y., Phang, J., Liu, H., Htut, P. M., Zhang, X., Pang, R. Y., Vania, C., Kann, K., and Bowman, S. R. (2020). “Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work?” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5231-5247. https://doi.org/10.18653/v1/2020.acl-main.467

Rivero 2022 Rivero, Manuel. (2022) “Italian Madrid: Ambassadors, Regents, and Courtiers in the Hospital de San Pedro y San Pablo”, *Culture & History Digital Journal*, 11(1), e003. Available at: https://doi.org/10.3989/chdj.2022.003

Rojo 2010 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 2016 Rojo, G. (2016) “Los corpus textuales del español”, In book: *Enciclopedia lingüística hispánica*. Publisher: Routledge. Editors: Gutiérrez-Rexach. Available at: https://www.researchgate.net/publication/294407007_Los_corpus_textuales_del_espanol

Rosset, Grouin, and Zweigenbaum 2011 Rosset, S., Grouin, C., and Zweigenbaum, P. (2011) “Entités nommées structurées: Guide d’annotation Quaero.”. Available at: http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf

Rubinstein and Shmidman 2021 Rubinstein, A., and Shmidman, A. (2021). “NLP in the DH pipeline: Transfer-learning to a Chronolect.” *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, 106-110. Available at: https://aclanthology.org/2021.nlp4dh-1.12

Ruder et al. 2019 Ruder, S., Peters, M. E., Swayamdipta, S., and Wolf, T. (2019). “Transfer Learning in Natural Language Processing.” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 15-18. Available at: https://doi.org/10.18653/v1/N19-5004

Ruiz Fabo et al. 2017 Ruiz Fabo, P., Bermúdez Sabel, H., Martínez-Cantón, C. and Calvo Tello J. (2017) “Diachronic Spanish Sonnet Corpus (DISCO)”, Madrid. UNED. Available at: https://github.com/pruizf/disco

Sanh et al. 2019 Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019) “Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter”, ArXiv preprint, abs/1910.01108

Scheible et al. 2020 Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V., and Boeker, M. (2020) “GottBERT: A pure German Language Model”, (arXiv:2012.02110). ArXiv. Available at: https://doi.org/10.48550/arXiv.2012.02110

Sekine 1998 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. Available at: https://aclanthology.org/M98-1019

Sánchez-Salido 2022 Sánchez-Salido, Eva. (2022) “Reconocimiento de entidades en corpus de dominios específicos: experimentación con periódicos históricos”, Master Thesis (30 ECTS). ETSI Informática. UNED

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

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

Vaswani et al. 2017 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017) “Attention Is All You Need”, (arXiv:1706.03762). ArXiv. Available at: https://doi.org/10.48550/arXiv.1706.03762

Wissler et al. 2014 Wissler, L., Almashraee, M., Monett, D., and Paschke, A. (2014) “The Gold Standard in Corpus Annotation”, Available at: https://doi.org/10.13140/2.1.4316.3523

Yadav and Bethard 2018 Yadav, V., and 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. Available at: https://aclanthology.org/C18-1182

Zoph et al. 2016 Zoph, B., Yuret, D., May, J., and Knight, K. (2016). “Transfer Learning for Low-Resource Neural Machine Translation.” *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1568-1575. https://doi.org/10.18653/v1/D16-1163