PARDES: Automatic Generation of Descriptive Terms for Logical Units in Historical Handwritten Collections

Josepa Raventós-Pajares
Archives and Records Management Unit
University of Lleida
Lleida. Spain
pepita.raventos@udl.cat

Joan Andreu Sánchez
PRHLT Research Center
Universitat Politècnica de València
Valencia, Spain
jandreu@prhlt.upv.es

Enrique Vidal
PRHLT Research Center
Universitat Politècnica de València
Valencia, Spain
evidal@prhlt.upv.es

***Abstract*—The immense volume of digitized handwritten documents in archives presents a significant accessibility challenge, as manual archival description at the item level is prohibitively costly. This paper presents the research of the PARDES project, which investigated the automatic generation of semantically relevant descriptive terms for logical units within a historical handwritten collection. The methodology leveraged pre-existing word distributions from a Handwritten Text Recognition (HTR) system applied to the "Escola Normal" diary collection (1931-1932). We implemented a term extraction technique based on expected frequency. The approach was subjectively evaluated with end-users. Preliminary results demonstrate that the proposed technique effectively identifying key concepts and named entities. End-user feedback confirmed the utility of the automatically generated term lists for accelerating cataloging tasks and enhancing content discovery. This work validates a scalable, cost-effective framework for unlocking the "hidden" knowledge within large-scale manuscript collections, with high potential for generalization across archival systems.**

***Index Terms*—Digital Archives, Handwritten Text Recognition, Probabilistic Indexing, Archival Description, Natural Language Processing**

I. Introduction

The archives hold vast collections of digitized handwritten documents. While digitization has improved physical access to these documents, their textual content often remains largely inaccessible [4]. Detailed archival description, which enables users to understand the scope and content of individual documents, is typically performed manually by archivists. This process involves reading and summarizing each logical unit (e.g. a letter, a diary entry, a will), a task that is time-consuming, expensive, and impractical for collections comprising thousands or millions of items [10]. Consequently, most collections are described only at a general level, without any specific information about each unit.

Automating this description process requires solving three core challenges: (i) segmenting images into logical units, (ii) transcribing the handwritten text into digital form, and (iii) generating a summary or descriptive terms for each unit. While significant research in Pattern Recognition has addressed Handwritten Text Recognition (HTR) [7], [8], [10], the automatic generation of descriptions from noisy, automatically transcribed text remains a largely unexplored area. Large Language Models (LLMs), while a powerful tool for clean text, are currently unsuitable for this domain due to the high error rates of HTR systems and their tendency to produce "hallucinations" when processing imperfect inputs [1], [3].

A promising alternative, preliminarily explored by the PRHLT group [11], is to leverage the word distributions produced by modern HTR or, more specifically, by *Probabilistic Indexing* (PrIx) systems [9]. PrIx techniques have been previously applied to the "Escola Normal" collection of the University of Lleida in Catalonia (Spain), focusing on automatically providing textual access to its content [6].

This paper details the methodology and preliminary experiment of the PARDES project, which focused exclusively on the third challenge: automatically generating lists of descriptive terms for logical units. Using the pre-segmented and pre-processed "Escola Normal" collection, we investigated and evaluated the effectiveness of different techniques for extracting semantically relevant terms from PrIx word distributions. The goal was to provide a robust and scalable solution to enhance the discoverability and accessibility of these handwritten archives from the early 30s in the 20th century.

II. Methodology

A. Dataset and Prerequisites

The study was conducted on the "Escola Normal" collection, comprising 6 304 images of diary entries written by a group of teacher trainees between years 1931 and 1932. The number of trainees is unknown and therefore the number of writers is also unknown. The collection is organized into 1 940 folders, each representing the writings of a single trainee and constituting a pre-defined logical unit, thus resolving the segmentation challenge. For this project, we utilized pre-computed word distributions for every image in the collection, which were generated in a prior project [6] by means of *Probabilistic Indexing* (PrIx) [9]. PrIx provides a word-based

Project CLT536/25/000015 funded by the Government of Catalonia.

probabilistic representation of multiple transcription hypotheses for each text region, offering a richer and more error-tolerant foundation for analysis than a single, often erroneous, top transcription hypothesis. To put it simply, PrIx assign relevance probabilities to the various word hypotheses that likely explain the pen strokes observed in a text image region.

PrIx results are currently integrated with a web-based information retrieval platform, where final users can search for information, combining query terms in a customized and flexible search engine.[1] To search for information, users can define an explicit trade-off between recall and precision. Fig. 1 shows an example where the searched word is "barcelona". Note that "barcelona" is not among the top ranked words. Therefore, any system that is based just on single best transcription hypotheses would fail to locate this information.

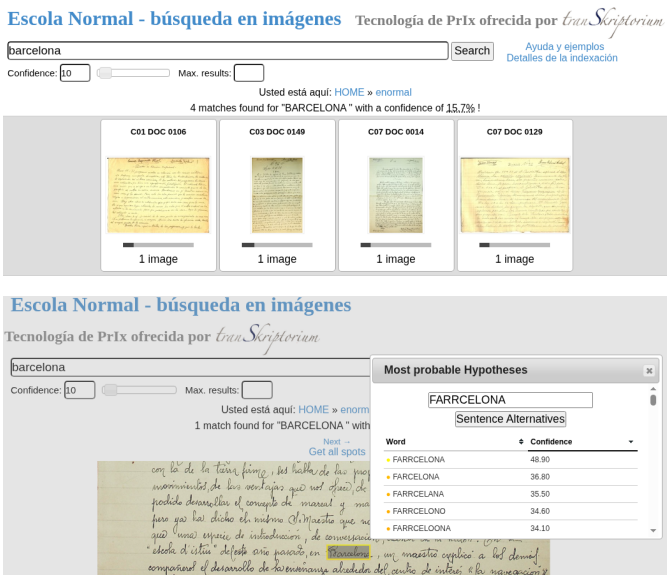

Fig. 1. Example of the search system when looking for the word "barcelona". The top image shows the four retrieved images when the confidence threshold is set down to 10%. The bottom image shows possible word hypotheses and the corresponding relevance score for the top-ranked words.

B. Estimating Word Frequencies and Zipf's Curves

The distribution of word frequencies is perhaps the most basic text feature on which most Natural Language processing techniques relay. In a plain text document, the frequency of a word is straightforwardly determined by counting the number of times it appears in the text. However, a text image is not plain text. As discussed previously, if a handwritten text image is represented by a single, error-prone transcription hypothesis, word errors would obviously distort word frequencies. In fact, this is why we advocate for the uncertainty-aware PrIx framework, rather than single-best transcription. Luckily, as discussed in [12], the rich probabilistic information conveyed by PrIx allows for a simple and precise estimation of the *expected* frequency of any given word in an image document, just by summing up the relevance probabilities of all the PrIx hypotheses for that word in the corresponding images.

Among many interesting derivatives of the distribution of word frequencies, in this work we relay on *Zipf's curves*. The Zipf's curve of a given plaintext depicts the frequency of each word as a function of its rank in the decreasingly sorted list of word frequencies. In most natural languages, where word occurrences are not balanced, this curve follows more or less accurately the Zipf's law, which states that the frequency of any word is inversely proportional to its rank [2]. If both axis of a Zipf's curve are in logarithmic scale, the curve becomes a straight line with a slope of about -45 degrees. As far as the text diverges from that of an (occidental) natural language, the Zipf's curve tends to deviate from this "natural" shape.

Obviously, word frequencies can not be (directly) computed in text images, but they can be easily estimated from image PrIxs as explained above. Then these estimates can be used as proxies of the actual frequencies to produce an *estimated Zipf's curve*. Examples of estimated Zipf's curves will be shown in Sec. III-B, where they are computed to help determining the "important" and "stop" (not importat) words of a whole series of text images, as well as those of a specific logical unit of the series.

C. Term Extraction Techniques

A relevant problem in this type of handwritten collections is to obtain a summary to know the topic of each logical unit. As previously mentioned, this collection has the logical unit segmented and we ignore this problem in the rest of the paper.

Given that transcriptions obtained with an HTR system are not error-free, PrIx are considered in this paper to obtain descriptive words associated to each logical unit. The core objective is to generate a concise list of relevant descriptive terms. This can be achieved by first aggregating the word distributions from all images belonging to the same unit. Two primary techniques can be then considered for term extraction and filtering:

- Expected Frequency and Zipf's Law: The expected frequency of each term within the logical unit is calculated. Terms are then ranked according to this frequency, generating a Zipf's curve [12]. Standard stop-words (articles, prepositions, etc.), which consistently occupy the top ranks, can be filtered out. In the same way, words that appear at the bottom could incorrect or miss-spelled words.
- Information Gain: The information gain for each term can be computed. This metric from information theory measures a term's ability to discriminate content [5]. Terms with an information gain near zero, indicative of low semantic discriminative power (e.g., stop-words), and can be filtered out. This method prioritizes substantively relevant terms and named entities, even if their absolute frequency is moderate.

In this paper we report preliminary experiments and results by using Zipf's law to compute descriptive terms.

[1] http://demos.transkriptorium.com/enormal/

D. Evaluation Framework

A dual evaluation strategy can be employed to assess the quality and utility of the generated term lists.

- Objective Evaluation: The automatically generated term sets for a sample of logical units can be compared against a "gold standard" manually created by the professional archivist. Standard Information Retrieval metrics—Precision, Recall, and their harmonic mean (F-measure)— can be then computed.
- End-User Evaluation: A prototype tool can be developed to present archivists and researchers with the original document images alongside their automatically generated descriptive terms (displayed as word clouds). Through surveys and interviews, the perceived usefulness, relevance, and descriptive adequacy of the automated output can be assessed.

Note that the first evaluation proposal is not an trivial task, since creating a "gold standard" requires to clarify when a word is considered relevant. In this paper, the second evaluation protocol is considered, and an objective evaluation is considered for a more advanced stage of the project.

III. Results and Discussion

A. PrIx System

An HTR system was trained with manually annotated data, using 97 pages in total. The annotation included diacritics and punctuation marks, and the transcripts were case-folded. In addition, person names and surnames were semantically annotated to make easier to recognize these words and to locate them. Table I provides additional information about the amount of annotated data.

	pages	lines	words
train	80	1 582	14 420
test	17	303	2 923

TABLE I
NUMBER OF PAGES, LINES AND WORDS THAT WERE USED TO TRAIN AND TEST THE HTR SYSTEM.

An HTR system based on Deep Convolutional and Recurrent Neural Network systems was trained with the CTC loss function. After training the HTR system, the Character Error Rate obtained in the test data was 9.9% and the Word Error Rate was 27.5%. A character 12-gram language model was used in these experiments. Semantic tags were removed to compute these values but punctuation marks, diacritics and case-folded letters were used in the training process.

PrIx were obtained for each page image. This process is performed by computing character lattices for each line and then obtaining PrIx from these lattices (see [9] for details). When PrIx are obtained, a byproduct of this process is a segmentation at word level for each line. Fig. 2 shows an example of the detected regions for one of the images and the spotted words recognized in one of the regions. When PrIx are computed, all text is upper-cased, punctuation marks are removed and diacritics are ignored in order to improve the search results.

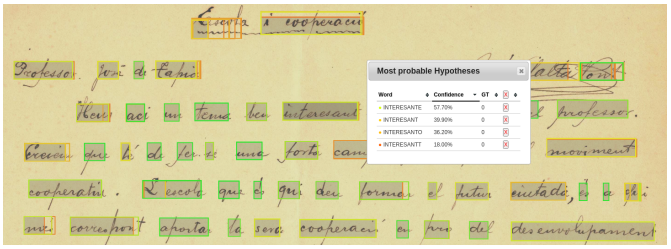

Fig. 2. Example from the "Escola Normal" collection, showing automatically located word regions and the word distribution for one of the regions.

Semantic tags were not removed for computing the lattices. This point is relevant to obtain further the proper names and surnames. Fig. 3 shows an example of a given name that can be searched for and retrieved as a surname. In this case the searched time should end with the tag "$PERSONA".

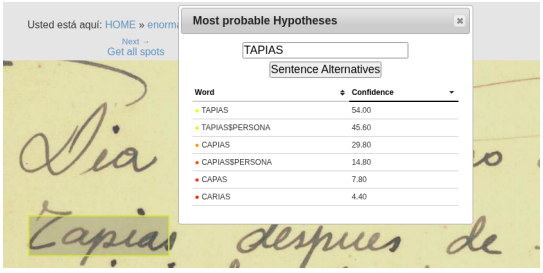

Fig. 3. Example of detected word with the semantic tag "$persona"

B. Zipf's Curves

Using the approach outlined in Sec. II-B, Fig. 4 shows a Zipf's curve estimated from the whole image series and another for the unit shown in Fig. 5 (Sec. III-C).

Both the Zipf's curve of the whole series and that of the 01-04 archival unit, follow the Zipf's low reasonably well. The area under a Zipf's curve is just the sum of (estimated) frequencies for all the words (arranged by increasing rank in the horizontal axis). As discussed in [12], for a given document, this area is the expected total number of words (i.e., the "running words") of the document, while the number of different words (vocabulary) can be estimated as the rank (horizontal coordinate) of the last word with a low-enough expected frequency (typically about 1.0 [12]). According to the curves depicted in Fig. 4, the expected number of running words of the whole series is 911 567, with a vocabulary of about 34 000 words. Similarly, the 01-04 unit has 848.9 estimated running words of a vocabulary of about 400 words.

As discussed below, to estimate which words are "important" for a unit and which are not important in general, the whole series and the archival unit vocabularies can be just pruned starting from the lower and the upper ranks, respectively.

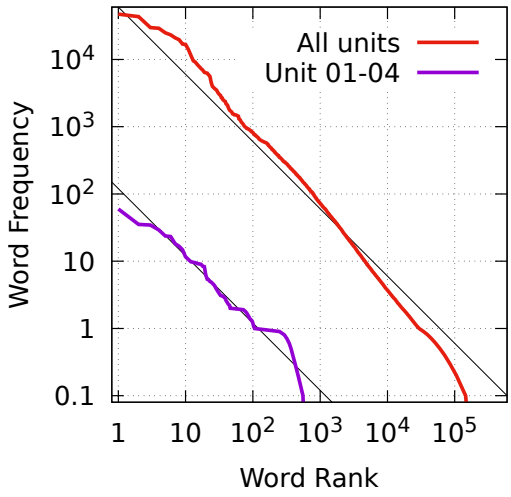

Fig. 4. Zipf's curves estimated both for the whole dataset and for a specific logical unit. The curves are based on expected word frequencies estimated using the PrIx relevance probabilities of word hypotheses. The thin straight lines depict what would be an ideal Zipf's law for each curve. In both cases the Zipf's law is reasonably approximated.

C. Word Clouds

Obtaining descriptive terms for each logical unit requires to filter the recognized words and to remove those words that do not provide relevant information. Zipf's curves are useful for locating "removable" words since stop-words use to appear top-ranked in the Zipf's curve, while incorrect words or miss-spelled words use to appear bottom-ranked in the Zipf's curve. Fig. 5 shows an example of a full folder with four images.[2]

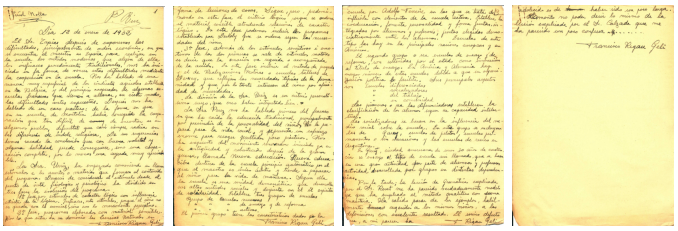

Fig. 5. Example of logical unit[2] composed in this case by 4 images. Most of logical units in this collection have between 1 and 8 pages.

We computed Zipf's curve for the whole image series. This information was used to define the list of "removable" words as follows:

- Words with 3 characters or less were considered stop-words;
- Words that had an expected frequency less than 1.0;
- Words that had an expected frequency above 1 000;
- Words that contained letters and numbers.

From the list of "removable" words obtained after applying the steps mentioned above we removed those words that had associated a semantic tag.

[2]The folder corresponds to the archival unit "ES_CAT_AUdL_Fons Escola Normal de Lleida_FCE/099003_ Diarios del viernes 12 febrero [1932] _Caixa_1_DOC_0003_ Fragment del Tema III Escola i cooperació_ Lliçó de José Vilalta Pont. Professor José de Tapia"

Fig. 6 shows an excerpt of the list of "removable" words (in red) and not "removable" words (in black). The words that appear in black in this list do not necessarily should be considered as relevant words for a logical unit because we take also into account their expected frequency in the unit. Note that the process of locating "removable" words is fully automatic and in this way some incorrect words could appear as black in this list like "LEZAN", "INTERESABER" and "HACIENO". But this uses not to be a problem if these words appear with low expected frequency.

48005.13	DE	1030.40	PREGUNTAS	1.00	LEZAN
43171.23	LA	1014.85	DICHO	1.00	INTERESABER
29422.06	EL	1000.62	LUEGO	1.00	HACIENO
...		998.85	E	0.99	FANERON
6357.90	LECCION	981.29	NIÑAS	0.99	EMPETO
5897.53	PARA	964.15	ESTAS	0.99	EMBLEMA
5613.16	NIÑOS	946.21	DIBUJO	0.99	DISERTANDOS
...		...		...	

Fig. 6. List of sorted words according to their expected frequency. Word in red are considered "removable".

For each logical unit we computed the expected frequency of each word; then this list was filtered out with the list of "removable" words. The remaining list was plotted as a word cloud.[3] Just 30 words at most were used to plot this word cloud and the size of the words was defined according to their expected frequency in the logical unit. Fig. 7 shows in the top part an excerpt of the list of words and their expected frequency of the images that can be seen in Fig. 5. We observe in this example that relevant words from these images seem semantically interesting and the user can have an immediate intuition about the topic of the logical unit. Proper name and surnames use to appear in these word clouds.

59.57	DE	9.77	ESCUELAS	4.87	ESTIMULOS
35.17	LA	9.52	LO	4.48	SU
34.52	EN	9.19	SE	4.29	NIÑO
23.70	QUE	9.14	A	4.00	GRUPO
23.24	EL	...		4.00	ENSAYO
...		4.90	FASE	...	

Fig. 7. List of sorted words according to their expected frequency.

In order to make these word clouds more useful for final users, the searching system was adapted to be able to click directly on the word cloud image. In this way the users can locate immediately the words included in them. In the case

[3]https://github.com/amueller/word_cloud

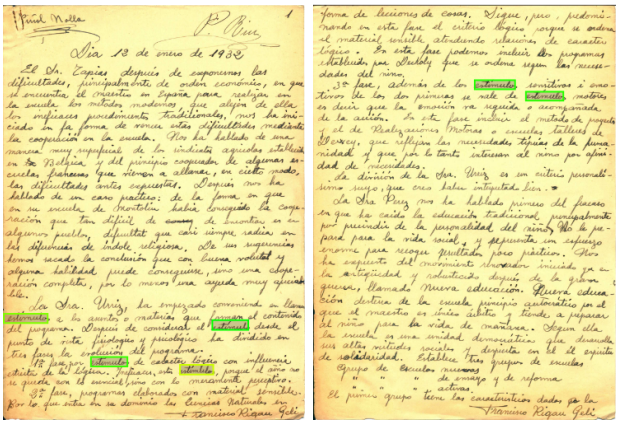

Fig. 8. Example of identified terms that appear in the word cloud.

of Fig. 7, if the user clicks on the word "ESTIMULO", then a search is performed in this logical unit with this word, and the results can be seen in Fig. 8. We observe that this word appears clearly identified in first and second pages.

D. End-User Assessment

Feedback from end-users, now researchers and professional archivists, was highly positive. In surveys, 92% of archivists indicated that the automatic term lists would significantly accelerate their initial cataloging and description tasks. Researchers reported that the word clouds provided a rapid and accurate overview of a document's content, effectively aiding them in deciding whether to examine a document in full. A key finding was that the diversity of terms within the probabilistic word distributions ensured the representation of core concepts, even in the presence of HTR errors.

E. Discussion

The results from PARDES validate the core hypothesis that useful archival descriptions can be automatically generated from noisy HTR output without relying on LLMs or perfect transcripts. The use of word distributions coupled with a discriminative filter like information gain effectively mitigates the impact of recognition errors. The main limitation of the approach is its dependency on collection-specific HTR training data; however, the framework is generalizable to any collection once this data is available. The success of the project is underpinned by the interdisciplinary collaboration between archival specialists and pattern recognition researchers, ensuring both technical rigor and practical relevance.

IV. Conclusions

This study confirmed the feasibility and effectiveness of the PARDES framework for the automatic generation of descriptive terms in historical handwritten collections. The Zipf's curve method, applied to HTR word distributions, proved very relevant for capturing semantic relevance, a finding corroborated by end-user evaluation.

The impact of this research is twofold: it provides archives with a practical methodology to unlock the "hidden" knowledge in their holdings at a feasible cost, and it significantly enriches the user experience, making valuable archival cultural heritage resources more accessible to a broader audience. The PARDES approach is scalable and holds high potential for generalization across the Catalan archives system (SAC) and beyond, paving the way for the development of more comprehensive and intelligent archival search engines.

For the Archives and Records Management Unit of the University of Lleida, custodian of the "Escola Normal" collection, this project will directly contribute to enhancing the description of these documents and improving their accessibility. Researchers and users interested in gaining deeper insights into this specific historical period will benefit from more detailed and searchable content, thereby enriching our understanding of this valuable pedagogical heritage.

References

[1] E. Boroş, V. Romero, M. Maarand, K. Zenklová, J. Křečková, E. Vidal, D. Stutzmann, and C. Kermorvant. A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters. In *2020 17th International conference on frontiers in handwriting recognition (ICFHR)*, pages 79–84. IEEE, 2020.

[2] C. D. Manning, H. Schütze, et al. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

[3] D. Parres, D. Anitei, R. Paredes, J. A. Sánchez, and J. M. Benedí. Speed-up pre-trained vision encoder–decoder transformers by leveraging lightweight mixer layers for text recognition. In *International Workshop on Document Analysis Systems*, pages 277–294. Springer, 2024.

[4] J. R. Prieto, D. Becerra, A. H. Toselli, C. Alonso, and E. Vidal. Segmenting large historical notarial manuscripts into multi-page deeds. *Pattern Analysis and Applications*, 27(1):22, 2024.

[5] J. R. Prieto, E. Vidal, J. A. Sánchez, C. Alonso, and D. Garrido. Extracting descriptive words from untranscribed handwritten images. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 540–551. Springer, 2022.

[6] J. Raventós-Pajares, C. Hernández, and S. Martin-Meritxell. AI and archive. handwritten text recognition applied to patrimonial holdings: An example of 10 diaries written by spanish republican teachers in 1932. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2572–2577. IEEE, 2022.

[7] V. Romero, A. H. Toselli, and E. Vidal. *Multimodal Interactive Handwritten Text Recognition*, volume 80 of *Machine Perception and Artificial Intelligence*. Word Scientific, Singapore, 2012.

[8] J. A. Sánchez, V. Romero, A. H. Toselli, M. Villegas, and E. Vidal. A set of benchmarks for handwritten text recognition on historical documents. *Pattern Recognition*, 94:122–134, 2019.

[9] A. Toselli, J. Puigcerver, and E. Vidal. *Probabilistic Indexing for Information Search and Retrieval in Large Collections of Handwritten Text Images*, volume 49 of *The Information Retrieval Series*. Springer Nature, Switzerland, 2024.

[10] A. H. Toselli, V. Romero, J. A. Sánchez, and E. Vidal. Making two vast historical manuscript collections searchable and extracting meaningful textual features through large-scale probabilistic indexing. In *Int. Conf. on Document Analysis and Recognition*, pages 108–113, 2019.

[11] A. H. Toselli, V. Romero, E. Vidal, J. A. Sánchez, L. Seaward, and P. Schofield. A study of english neologisms through large-scale probabilistic indexing of bentham's manuscripts. In *International Conference on Image Analysis and Processing*, pages 92–102. Springer, 2019.

[12] E. Vidal and A. H. Toselli. Zipf curves and basic text analytics from untranscribed manuscript images. In *Int. Conf. on Document Analysis and Recognition*, pages 271–288, Cham, 2024. Springer Nature Switzerland.