Unlocking Nahua cosmovision through machine learning

Javier Pereda

John Moores University – Liverpool, UK

Alexander Sanchez Diaz

Universidad de Alicante – Spain

Patricia Murrieta Flores

Lancaster University – UK

The tradition of writing in Mesoamerica was by painting. While most pre-Hispanic codices were destroyed during the Conquest, the tradition continued with colonial codices created by the tlacuiloque, where indigenous and Spanish techniques entangled while Indigenous scholars resisted, contested, and fought to maintain their own traditions. The shift from pre-Hispanic to colonial codices marked significant changes in aesthetics and purposes. Pre-Hispanic codices combined painting, oral, and performative traditions that served as mnemonic devices and held religious significance, connecting the diverse Indigenous knowledges and cosmovision. It is estimated that around 20 pre-Hispanic codices survive, and the tradition of writing by painting was substantially transformed while preserving their legal relevance. Nonetheless, the legacy continues with over 500 colonial codices documented (Valle, 1999), and more being discovered by the communities that created and sheltered them. Codices chronicle the cosmovision, history, and mythologies of Mexican people and many still constitute 'live documents', influencing modern social practices (Jiménez Padilla & Villela Flores, 1999).

Unlocking the colonial archive

Codices contain deep semantic richness with iconography and text. Traditionally, studying them required producing interpretations from analysing these documents, once scattered worldwide due to historical shifts, which was expensive and labour-intensive. Digital archives, supported by various institutions such as the National Institute of Anthropology and History of Mexico, the LLILAS Benson Latin American Collection, and projects such as Tlachia and Amoxcalli, among others, have eased access to this information by digitalising vast amounts of its content. Yet challenges persist, notably technological limitations of Western tools and barriers hindering Indigenous semantic understanding. We aim to dismantle these obstacles, enhancing access to knowledge within codices and maps and advancing the decolonisation of archives and computational techniques.

Our research group developed a Natural Language Processing (NLP) system for identifying and extracting large amounts of information from historical texts, creating a classification system, ontology, and Machine Learning (ML) model for automatic marking of historical information (Murrieta-Flores et al., 2022). This innovation streamlined the annotation of colonial Mexico's data, including the 12 volumes of the 16th-century Geographic Reports of New Spain, and identified 15,000 placenames from 16th-century Mexico (Murrieta-Flores et al., 2022). Our work led to the Digging into Early Colonial Mexico Geographical Gazetteer, a digital resource pioneering the study of this period (Favila-Vázquez et al., 2023).

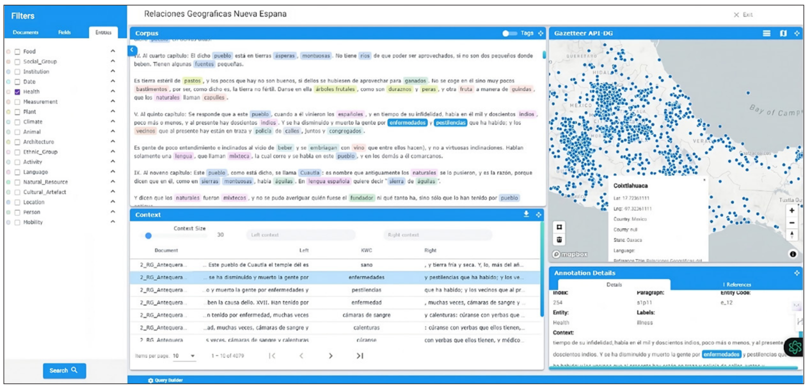

Figure 1. GTA Software. Query of all terms related to the entity of "health" in the corpus of the Relaciones del Arzobispado de México.

We built the Geographical Text Analysis (GTA) software, combining Corpus Linguistics, AI, and Geographic Information Science, to analyse information in large historical collections. The software allows the linguistic exploration of topics that might be of interest to the researcher and the identification of geographies related to these (see Figure 1).

We're now expanding our approach to include Mexican pictographic sources, using ML and NLP to make Indigenous visual languages more accessible. Our new approach, Visual Natural Language Processing (VNLP), combined with a domain ontology, enhances the machine and human interpretation of iconographic content, facilitating a deeper semantic understanding of the visual systems involved.

Visual NLP for Mexican codices

Decoding visual narratives unfolds across three layers: the pre-iconographical description, which involves identifying basic visual elements; the iconographical analysis, interpreting these elements as narratives and allegories; and the iconological interpretation, which unravels the symbolic and cultural significance (Panofsky, 1939). While ML through Computer Vision (CV) excels in the first two layers, it faces challenges with the complexity of the symbolic layer. Visual Natural Language Processing (VNLP) can help overcome this limitation, unlocking the layers of complex symbolism and unveiling the rich cosmovision within the multimodal complexity of the codices.

Specialised *tlacuiloque* in Mesoamerica crafted codices that showcased diverse contextual information through unique representational structures. The *Codice Xolotl* and the *Tira de Tepechpan*, both historical codices, showcase this variety; the former emphasises spatial historical narratives while the latter focuses on chronological sequences (Thouvenot, 2018). In religious context, the *Codice Fejervay-Mayer* uses figurative and pictographic writing to depict the natural and divine forces of human activities, in contrast to the *Codice Borbonico*, which merges European imagery with Nahua writings (Thouvenot, 2018). This differentiation in specialisation has significantly influenced the representation of information, as exemplified by the *Codice Borbonico's* integration of European and Nahua elements, facilitating pre-iconographic analysis through ML technologies. The challenge, however, lies in equipping ML technologies with the system of knowledge and understanding of figurative writing and reading sequencing of codices to fully grasp their depth.

To comprehend pictographic content, both machines and humans need to understand the underlying cosmovision and meaning. Mesoamerican iconography forms part of a sophisticated language system, embodying semiotic, cultural, and visual elements. For instance, the deity *Chalchiuhtlicue* is characterised by her connections to water and social rituals such as *temazcal* (steam bath) and childbirth ceremonies. This is represented through facial paint, nose ornaments, headdress and decorations, among others (Dehouve, 2020). Lopez Austin (2010) and others have shown how design elements convey complex concepts and extend to place and personal names (Kuehne Heyder, 1998; Valle, 1998). For example, the anthroponym *Xochicozcatl* is shown in the Cuauhtinchan map through a composition of "*xochitl*=flower" and "*cozcatl*=necklace", and the toponym *Texcallan*, or place of many cliffs, denoted by the depiction of a toponym (a mountain) and a cliff (see Figure 2).

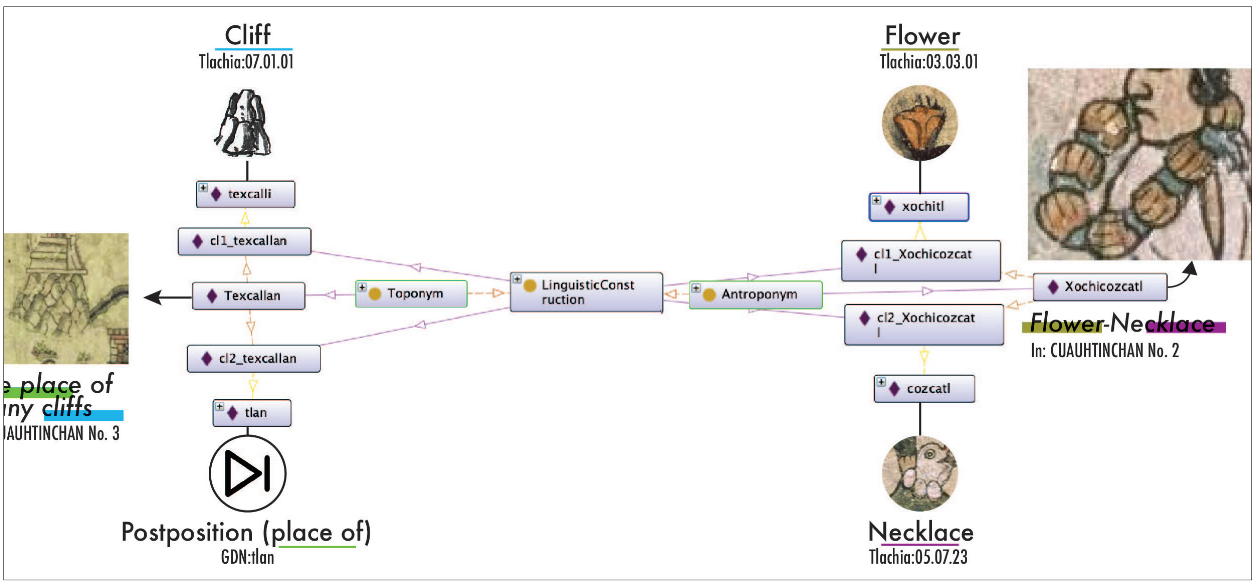

Figure 2. Ontologic representation of the linguistic construction of the toponym Texcallan (left) and the anthroponym Xochicozcatl (right).

Furthermore, understanding these documents requires grasping their internal logic, contextualising iconography within the context of the entire narrative. For example, the *Mapa Quinatzin*, shown in Figure 3, features a three-layered narrative: The outer layer list places (toponyms) subject to taxation, the middle layer depicts the government palace, and the innermost layer outlines the *Nezahualcoyotl* council (Mohar Betancourt, 1999). Analysing codices through VNLP requires engaging with the iconographic description as well as the layout and positioning of the glyphs.

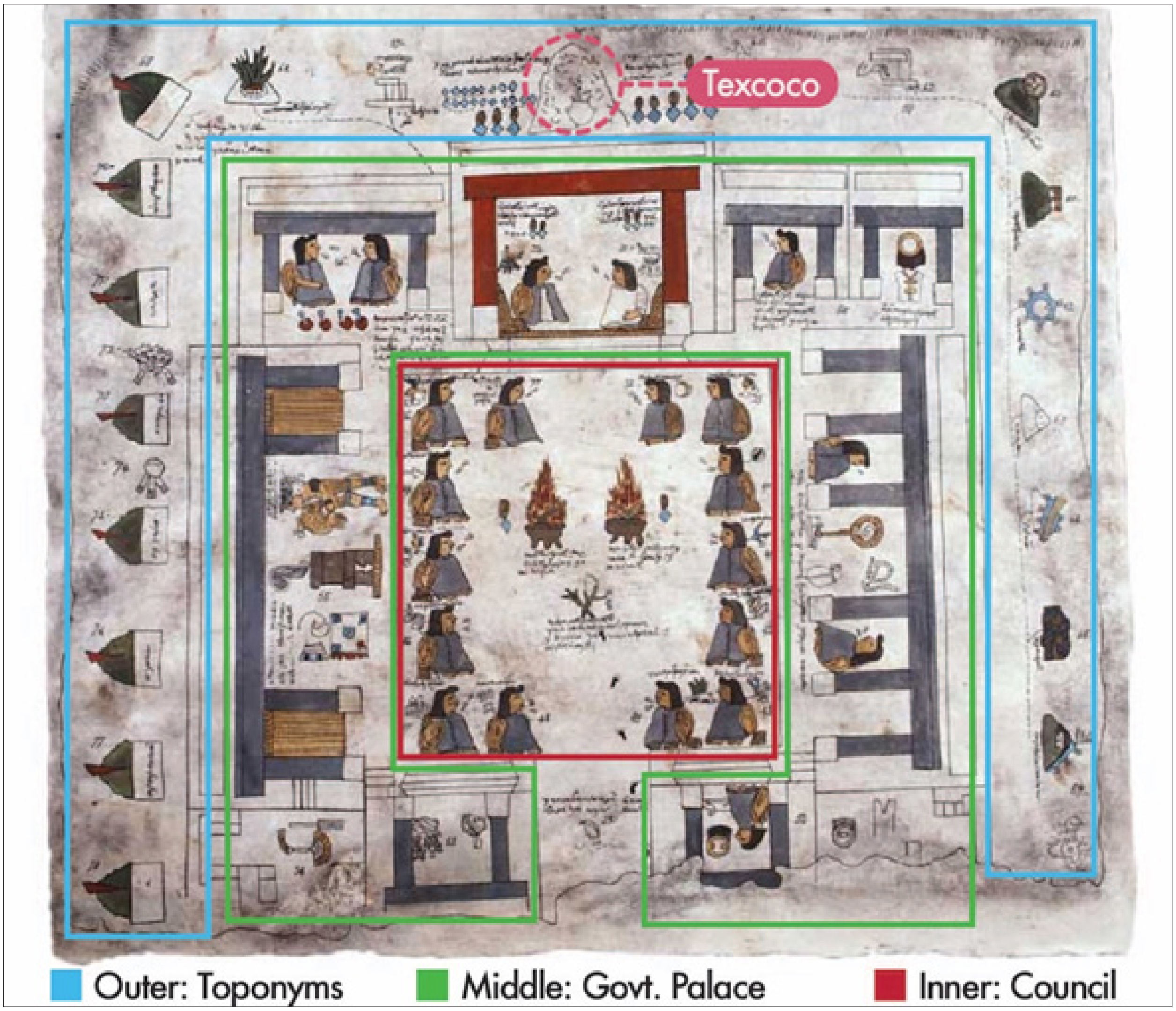

Figure 3. How to read the Mapa Quinatzin (CC0) adapted from Mohar Betancourt (1999), (Wikimedia, 2024)

We are creating a glyph taxonomy for AI use, focusing on a unified framework for organising and connecting Linked Data. This will facilitate access to the nuanced contexts of historical documents, as the absence of consistent metadata significantly hinders human and machine understanding, and the variability in image description practices aggravates the risk of misinterpreting critical data. The flexibility of Indigenous languages, such as Nahuatl, coupled with the multiplicity of interpretations a single image or codex can evoke, calls for a socio-technical solution that accommodates diverse insights, including those shaped by the Indigenous cultural and linguistic cosmovision, and the translation process of both Indigenous and Europeans. Our strategy is grounded in the Semantic Web principle of "Anyone can say Anything about Any topic" (AAA) (Allemang & Hendler, 2011), striving to merge diverse ontological representations through a combination of automated identification and manual collaborative refinement (Zhitomirsky-Geffet & Shalom Erez, 2014). This approach is further enriched by our inter-annotator agreement approach (Liceras-Garrido et al., 2019), inviting contributions from various perspectives and backgrounds to enrich our ontology.

This data is produced as Open Annotate Data to support the diverse interpretations and reuse of information aligned to our domain ontology. Our domain ontology for Nahuatl descriptions has been based on *Tlachia's* Codex Dictionary (Tlachia, 2012), the *Gran Diccionario Náhuatl* (GDN, 2012), and Joaquin Galarza's work on the *Mapa de Cuauhtinchan III* (Galarza & Yoneda, 1982). This effort aims to enhance searchability and analysis through knowledge graphs, standardised image descriptions, and improved access to iconographic knowledge, helping reduce and avoid misinterpretation.

Conclusion

Developing a domain ontology is opening new ways to explore and record information about the cosmovision and languages in codices through ML technologies. We are also connecting this to standard ontologies such as CIDOC-CRM and Europeana Data Model (Candela et al., 2023). Integrating cultural understandings and information of Indigenous cosmovision into VNLPs is crucial, as this requires considering how to integrate complex cultural knowledge with computational workflows. For this, we are creating a Web AI Lab to implement these pipelines, featuring ML-supported VNLP and a Text Annotation tool for historical documents.

References

Allemang, D. & Hendler, J. (2011). *Semantic web for the working ontologist: Effective modeling in RDFS and OWL*. Elsevier.

Candela, G., Pereda, J., Sáez, D., Escobar, P., Sánchez, A., Torres, A. V., . . . Murrieta-Flores, P. (2023). An ontological approach for unlocking the Colonial Archive. *ACM Journal on Computing and Cultural Heritage, 16*(4), 1-18.

Dehouve, D. (2020). The rules of construction of an Aztec deity: Chalchiuhtlicue, the goddess of water. *Ancient Mesoamerica, 31*(1), 7-28.

Favila-Vázquez, M., Liceras-Garrido, R., Bellamy, K., & Murrieta-Flores, P. (2023). De la cosmografía real a la cartografía digital: construyendo un diccionario geográfico de la Nueva España en el siglo XVI. In A. Trujillo Herrada (Ed.), *Sistemas de información geográfica para arqueólogos: Repensando el espacio en contextos arqueológicos mesoamericanos*. El Colegio Mexiquense.

Galarza, J., & Yoneda, K. (1982). *Mapa de Cuauhtinchan no. 3: (glifos: catalogo-diccionario)* (Vol. 3): Archivo General de la Nación.

GDN. (2012). Gran Diccionario Náhuatl. https://gdn.iib.unam.mx/

Jiménez Padilla, B., & Villela Flores, S. (1999). Vigencia de la territorialidad y ritualidad en algunos códices coloniales. *Arqueología mexicana, VII*(38), 58-61.

Kuehne Heyder, N. (1998). Iconografía de las deidades lunares huaxtecas. In B. Barba de Piña Chán (Ed.), *Iconografía mexicana I*. México: Instituto Nacional de Antropología e Historia.

Liceras-Garrido, R., Favila-Vázquez, M., Bellamy, K., Murrieta-Flores, P., Jiménez Badillo, D., & Martins, B. (2019). Digital approaches to historical archaeology: exploring the geographies of 16th century New Spain. *Open Access Journal of Archaeology and Anthropology, 2*(1).

Mohar Betancourt, L. M. (1999). Cómo leer un códice. El Mapa Quinatzin. *Arqueología Mexicana, VII*(38), 32-33.

Murrieta-Flores, P., Jiménez-Badillo, D., & Martins, B. (2022). Digital Resources: Artificial Intelligence, Computational Approaches, and Geographical Text Analysis to Investigate Early Colonial Mexico. In *Oxford Research Encyclopedia of Latin American History*.

Panofsky, E. (1939). *Studies in iconology: humanistic themes in the art of the Renaissance*. Routledge.

Thouvenot, M. (2018). Escritura figurativa del náhuatl: in tlacuilolli, in tlilli, in tlapalli. https://thouvenotmarc.com/textos/Tlacuilolli.html

Tlachia. (2012). Contenidos Pictográficos del Náhuatl. https://tlachia.iib.unam.mx/

Valle, P. (1998). Topónimos del Códice de Tlatelolco. In B. Barba de Piña Chán (Ed.), *Iconografía mexicana I* (pp. 157-165). México: Instituto Nacional de Antropología e Historia.

Valle, P. (1999). Memorias en imágenes de los pueblos indios. *Arqueología Mexicana, VII*(38), 6-13.

Wikimedia. (2024). Codex Quinatzin.

Zhitomirsky-Geffet, M. & Shalom Erez, E. (2014). Maximizing agreement on diverse ontologies with "wisdom of crowds" relation classification. *Online Information Review, 38*(5), 616-633.