Digitizing the Chicago Assyrian Dictionary: Accelerating Document Analysis with VLMs and HPC Infrastructure

Himanshi Yadav [1] Christian Borgen [2] Hervé Reculeau [2] Foy Scalf [2] Ross Hyman [1]

[1]Research Computing Center [2]Institute for the Study of Ancient Cultures

Motivation

- The Chicago Assyrian Dictionary (CAD), completed in **21+ volumes over 90 years (1921–2011)**, is a foundational lexical reference for the **Akkadian language and Assyriological studies**. [1]
- The dictionary volumes, available as scanned images of the print dictionary, limit scholarly use and computational analysis.
- Optical Character Recognition (OCR) tools enable conversion from images to text but are not designed for the typographic and linguistic density of Assyriological lexicography.
- Vision-language models (VLMs) have proven proficient on multi-language OCR tasks.
- Digitization of the CAD is an ethical and scholarly **text corpus development task at scale**, necessitating **multi-level domain expert data alignment** with **low-medium computational needs** for **machine learning model and pipeline development**.
- **Can VLMs accelerate OCR for low-resource, mixed-script, multi-language corpora? Can HPCs accelerate corpus development**?

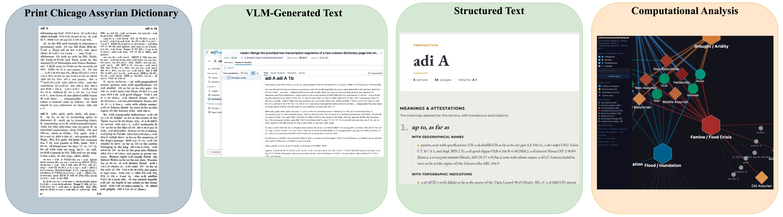


Figure 1. Print dictionary pages are transcribed using a VLM. VLM-generated transcriptions are converted to structured lexical entries to enable scholarly use and downstream computational analysis.

Challenges

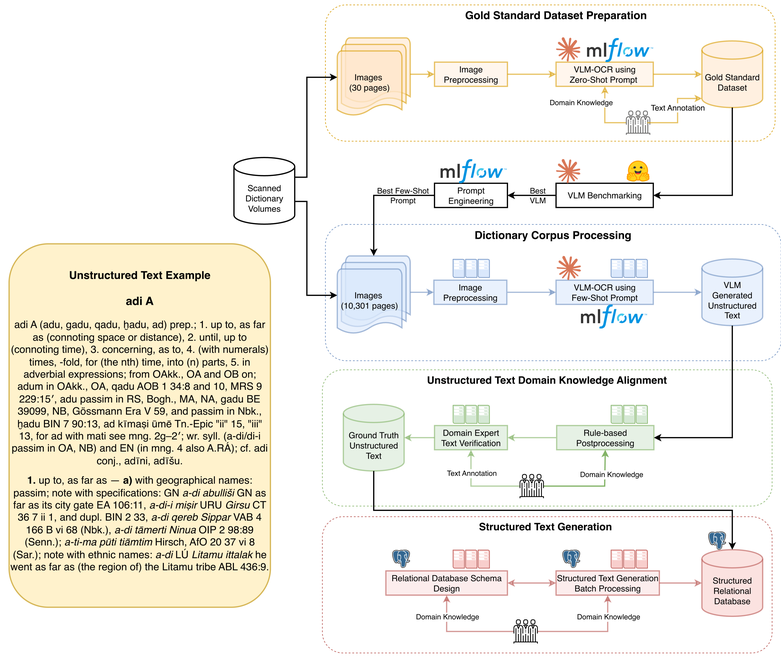
- **Lack of Ground Truth Data** Benchmarking tools with a lack of pre-existing corpora similar to the CAD.
- **High Lexicographic Density** The CAD includes English definitions, diacritic-rich Akkadian transliteration, Sumerian logograms, and bibliographic references to external cuneiform corpora.
- **Lack of VLM Interpretability** Inconsistent VLM errors have low visibility and lack traceability.
- **Data-Limited Optimal Prompting** Developing robust prompting techniques is constrained by limited availability of diverse development data.

Approach

We present a pipeline to convert scanned dictionary images into a structured relational text database.

1. **Gold Standard Dataset Preparation** VLM-assisted image-to-text generation and domain expert alignment to prepare a gold standard dataset.
2. **VLM Benchmarking** Best VLM selection by benchmarking open-weight and closed-weight VLMs on the gold standard dataset.
3. **Prompt Engineering** Optimal prompt design and selection, using best performing VLM, on the gold standard dataset.
4. **Rule-based Postprocessing** Rule-based postprocessing of VLM-generated dictionary text corpus, using domain expert-identified VLM errors and edge cases.
5. **Domain Expert Text Alignment** Domain experts and annotators make corrections to VLM-generated dictionary text corpus at scale.
6. **Structured Relational Database Generation** Relational database schema design and unstructured to structured text conversion with domain expert inputs.

Approach



Unstructured Text Example

adi A

adi A (adu, gadu, qadu, ḫadu, ad) prep.; 1. up to, as far as (connoting space or distance), 2. until, up to (connoting time), 3. concerning, as to, 4. (with numerals) times, -fold, for (the nth) time, into (n) parts, 5. in adverbial expressions; from OAkk., OA and OB on; adum in OAkk., OA, qadu AOB 1 34:8 and 10, MRS 9 229:15′, adu passim in RS, Bogh., MA, NA, gadu BE 39099, NB, Gössmann Era V 59, and passim in Nbk., ḫadu BIN 7 90:13, ad kīmaṣi ūmē Tn.-Epic "ii" 15, "iii" 13, for ad with mati see mng. 2g–2′; wr. syll. (a-di/di-i passim in OA, NB) and EN (in mng. 4 also A.RA); cf. adi conj., adīni, adīšu.

1. up to, as far as — **a)** with geographical names: passim; note with specifications: GN *a-di abulliši* GN as far as its city gate EA 106:11, *a-di-i miṣir* URU *Girsu* CT 36 7 ii 1, and dupl. BIN 2 33, *a-di qereb Sippar* VAB 4 166 B vi 68 (Nbk.), *a-di tāmerti Ninua* OIP 2 98:89 (Senn.); *a-ti-ma pūti tiāmtim* Hirsch, AfO 20 37 vi 8 (Sar.); note with ethnic names: *a-di* LÚ *Litamu ittalak* he went as far as (the region of) the Litamu tribe ABL 436:9.

Figure 2. HPC-enabled pipeline for converting scanned print dictionary volumes to a structured relational database.

Results

- **VLM Benchmarking** Claude Opus 4.6 produces best zero-shot performance on 30 pages of the gold standard dataset with 2.40% CER and 6.80% WER. In few-shot experiments, it achieves a 1.45% CER and 5.75% WER on the development set (for prompt design) and 1.49% CER and 4.73% WER on the validation set (for prompt design validation).

Table 1. Zero-Shot VLM Benchmarks: Average Character Error Rate and Word Error Rate (in %) on the Gold Standard Dataset

Model	CER	WER
Qwen3-VL-2B	45.02	68.02
OlmOCR-2-7B	8.25	21.62
Sonnet 4.5	7.25	21.95
Opus 4.6	**2.40**	**6.80**

CAD Website
https://isac-cad.uchicago.edu

VLM Benchmarks
https://github.com/himanshisyadav/VLM-OCR-CAD

Results

- **VLM-OCR Token Cost and Execution Time** The dictionary page images are split into halves locally before constructing the API requests, to address the high text density per page. Using a transcription prompt configuration, followed by a merge prompt configuration, to convert the images to markdown text, we arrive at the VLM-generated unstructured text corpus.

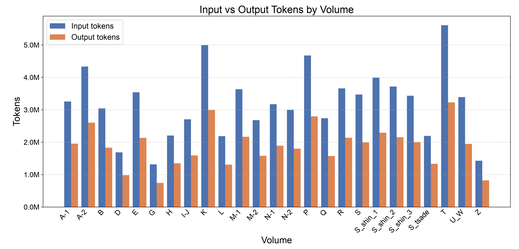


Figure 3. Total input and output token costs (closed-weight model) per CAD dictionary volume.

Figure 4. Total execution time per CAD dictionary volume.

Conclusion and Future Work

- **Conclusion** Work demonstrates that VLMs and HPC infrastructure can accelerate corpus development.
- **Ongoing** Domain expert verification of postprocessed VLM-generated text is underway. Comparing Volume Q's postprocessed text against domain expert aligned ground truth text, we achieved a 4.04% CER and 3.60% WER. We continue efforts on domain expert aligned relational database development.
- **Future Work** Leverage corpus to finetune SOTA VLMs to digitize the CAD master cards or master image corpus. Computational analysis to study ancient climates and social structures.

Acknowledgments

This work is supported by the Yelda Family Fund at ISAC and University of Chicago's AI Pillar initiative.

References

[1] Martha Roth. How we wrote the chicago assyrian dictionary. *JNES*, 69:1–21, 2010. doi:10.1086/654936.

hyadav@uchicago.edu