Journal of Open Humanities Data

# The Evolution of Written Chinese (1910–1945): A Computational Newspaper Similarity Dataset

DATA PAPER

LOUIS L. MERRIAM

ubiquity press


## ABSTRACT

This dataset supports a computational study of linguistic change in written Chinese during the late Qing and Republican era. The dataset comprises a machine-readable text corpus and pairwise similarity measures for seven regional Chinese newspapers published from 1910–1945, with images sourced from the Internet Archive. In total, 324 editions were converted to text using two independent Optical Character Recognition (OCR) pipelines, yielding a corpus of 11.6 million characters. Length-matched yearly corpora were compared using a TF–IDF weighted cosine similarity analysis, and the results of this analysis are included in the dataset. OCR accuracy was assessed against two independent gold standards totaling 2,045 characters, giving a 97.7–98.4% character capture rate. The deposit contains raw and normalized text, the yearly corpora, an edition-level manifest, a block-level index with quality metrics, similarity tables and scripts.



**CORRESPONDING AUTHOR:**

**Louis L. Merriam**

Sidwell Friends School, Washington, D.C., USA

louis.merriam@gmail.com



**KEYWORDS:**

Chinese language; historical linguistics; digital humanities; optical character recognition; cosine similarity; newspapers

**TO CITE THIS ARTICLE:**

Merriam, L. L. (2026). The Evolution of Written Chinese (1910–1945): A Computational Newspaper Similarity Dataset. *Journal of Open Humanities Data*, 12: 144, pp. 1–6. DOI: https://doi.org/10.5334/johd.615

# (1) OVERVIEW

## REPOSITORY LOCATION

Zenodo: https://doi.org/10.5281/zenodo.20777803 (Merriam, 2026; concept DOI, all versions; this paper describes v2.0, https://doi.org/10.5281/zenodo.22286183), released under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

## CONTEXT

During the period between 1910 and 1945, written Chinese shifted from Classical Chinese, the literary standard used for over two thousand years, toward a vernacular (baihua) form closer to the spoken language. This was a gradual, regionally uneven transition that is difficult to quantify (Chen, 1999). Newspapers are ideal for tracking this change as they appeared regularly and often reflected local usage. The seven titles chosen for the analysis sample dialectally distinct regions: Wu (Shanghai), Northern Mandarin (Beijing), Southwestern Mandarin (Sichuan and, in part, Hunan), Yue (Guangzhou) and Southern Min (Xiamen). The dataset was produced to measure the pace and regional variation of this change. Sample analyses conducted on it are described in Merriam (2025) and summarized in the Reuse Potential section.

Historical newspapers are an established and excellent substrate for measuring language change. Buntinx, Bornet and Kaplan (2017) tracked lexical change across 200 years of Swiss newspapers, and Hengchen, Ros, Marjanen and Tolonen (2021) modeled changing vocabularies across four different national-scale collections. Comparable Chinese studies include the ENP-China project, which addressed word segmentation in transitional Chinese (Blouin, Huang, Henriot, & Armand, 2023). Ren (2026) studied the Northwest-frontier press for discourse analysis, and Henke and Arnold (2022) built a high-accuracy OCR pipeline for one entertainment newspaper. However, none have released a dialectally stratified, multi-title corpus, nor one with matched-length yearly units and per-edition metrics from this era.

# (2) METHOD

## STEPS

Seven newspapers were selected to represent distinct dialect regions: Shenbao (Shanghai), Shangwu Ribao (Chongqing), Beijing Ribao (Beijing), Yuehua Bao (Guangzhou), Hunan Guomin Ribao (Changsha), Xinxin Xinwen (Chengdu) and Jiangsheng Bao (Xiamen). Issues were retrieved from the Internet Archive collections and rendered at approximately 600 dots per inch. The primary OCR pipeline used to convert the images into machine-readable text was the Google Cloud Vision Document Text Detection, configured to recognize Traditional Chinese. As an independent validation, a Tesseract pipeline (Smith, 2007), using the chi_tra and chi_sim models, was also run after OpenCV pre-processing. Recognized text was filtered to Han (CJK) characters and analyzed at the level of individual characters rather than segmented words, because word segmentation is unreliable on mixed Classical–vernacular text (Blouin et al., 2023). For each newspaper and year, issues were concatenated toward a target of 50,000 characters to produce length-matched yearly corpora. The yearly corpora were represented as character-level term vectors, weighted with TF–IDF, and compared pairwise using cosine similarity (Manning, Raghavan, & Schütze, 2008). This similarity measure is fairly common, but the adaptations made for this material are the character-level units, the length-matching of the yearly corpora, which removes sensitivity of the measure to sample size, and the dual-pipeline design, which allows us to account for OCR quality by comparing results across the two OCR pipelines (Hill & Hengchen, 2019). Advertisements and other non-news matter were not excluded from the yearly corpora or from the similarity analysis but were instead flagged so that users can filter them.

## SAMPLING STRATEGY

Coverage spans 1910 to 1945 across 136 newspaper-years. In total 325 editions were scanned (324 produced OCR output), comprising 11,555,879 characters (9,472,962 of them Chinese). The 50,000-character target was reached for 133 of 136 newspaper-years. Three sparsely digitized years fell short because too few or too thin issues were available: Beijing Ribao 1912 (19,417 characters), Beijing Ribao 1920 (29,972) and Jiangsheng Bao 1939 (25,216). Because editions were sampled from the start of each year, New Year issues are over-represented.

## QUALITY CONTROL

Only Traditional characters were recognized to avoid the accidental recognition of Simplified forms introduced after the study period. Agreement between the two OCR pipelines served as the first quality check: the yearly similarity trends derived from the Vision and Tesseract text were positively correlated (Pearson $r = 0.55$; Spearman $\rho = 0.62$), so the trends are not a result of one engine's errors. OCR accuracy was then evaluated against two independent gold standards. The first was an author-verified transcription of 767 characters from ten editions spanning all seven titles and the full time period, drafted with AI assistance from the page scans and checked character by character by the author against the scans. The Vision model recovered 98.4% of the characters and 83.6% of the rare characters (corpus frequency below 1 in 10,000), with 74.7% of characters in contiguous reading order. The second was an independent transcription of 1,278 characters from different editions, produced by a vision-language model (Claude, Anthropic); against it the Vision output performed similarly, recovering 97.7% of characters and 83.1% of rare characters, with 76.0% in contiguous reading order. A character counts as recovered when it is present in the OCR output of the corresponding edition irrespective of position, so the capture rate is an order-agnostic measure and reading-order fidelity is reported separately. The convergence of the two evaluations across different transcriptions and editions indicates that the figures are not an artifact of one sampling frame.

The quality of the scans is not uniform, and the dataset quantifies quality for each edition and title. Eight quality indicators, all recomputable from the raw text, are included in editions.csv. They include the shares of Han, Latin and digit characters (han_ratio, which is low when an edition is dominated by advertising, Latin text or noise; latin_ratio and digit_ratio, proxies for mastheads, advertising matter, rate cards and price tables); the rate of CJK Extension A/B glyphs per 10,000 Han characters (cjk_ext_rate_per_10k, elevated when Vision hallucinates rare glyphs); the share of Han characters that occur fewer than once per 100,000 across the whole corpus (corpus_rare_han_ratio, high where an edition uses characters no other edition uses, indicating misrecognition); mean block length and the share of blocks of two characters or fewer (mean_line_len and short_line_ratio, which mark pages shattered into fragments by the vertical multi-column setting); and the rate of Han-adjacent Latin characters, which are almost always misrecognized punctuation marks (circle_punct_rate_per_10k, discussed below). Four of these metrics are used by a quality_flags column: low_han_ratio (below 0.75), high_latin (above the 95th percentile), heavy_circle_loss (above 10 per 10,000) and fragmented_layout (short_line_ratio above the 90th percentile). The flags are heuristic screening aids used to identify scans with poor quality. Reading order is the weakest aspect of the scans. The pages are largely not shuffled, as page position and file position correlate at Spearman 0.98. Instead, the limitation is that Vision reads the vertically set pages partly across columns, so a long block of text can weld several unrelated articles together. Blocks of 41 characters or more are only about 15% single-article text, while essentially 100% of 15-character sequences come from the same article (Spearman $-0.85$). In essence, the text jumbles together segments of text around 10–20 characters of length. Character-frequency measures are unaffected by this. The merge-risk grade indicates how likely a block of text is to contain text from more than one article based on length. Order fidelity is lowest for Hunan Guomin Ribao (48%). The Yuehua Bao employed punctuation in the form of two adjacent circles: '○○', which was frequently misrecognized as Latin characters. These punctuation markers were misrecognized at a rate of 21.5 runs per 10,000 characters against 0.33–1.35 for other titles. This is simply a misrecognition of a unique punctuation style rather than a scanning failure and does not affect the Han-filtered similarity measures.

## (3) DATASET DESCRIPTION

The deposit contains 847 files in four groups — text, metadata, quality and similarity — with code, figures and a data dictionary defining every column of every table. Everything is keyed on the Internet Archive edition identifier, so any figure reported here traces back to the edition, page and block it came from.

| | |
|---|---|
| **Repository name** | Zenodo |
| **Object name** | chinese-newspaper-similarity-dataset (v2.0) |
| **Format names and versions** | TXT (UTF-8); CSV; XLSX (Office Open XML); PNG; JSON; Python 3 (.py); Markdown; ZIP |
| **Creation dates** | 2025-06-14 to 2026-09-01 |
| **Dataset creators** | Louis L. Merriam (Sidwell Friends School) — data collection, OCR, analysis, curation |
| **Language** | Traditional Chinese (corpus); English (documentation and variable names) |
| **License** | Creative Commons Attribution 4.0 International (CC BY 4.0) |
| **Publication date** | 2026-06-20 (v1.1); 2026-09-03 (v2.0) |

The text comes in three tiers. The raw tier is the unmodified Google Vision output, one UTF-8 file per edition (324 files), for users studying OCR behavior itself. The normalized tier (324 files) fixes the circle-punctuation misrecognition issue and collapses whitespace, and this tier is the default for most reuse. The third tier holds the 136 length-matched yearly corpora of 50,000 characters that are the unit of the similarity analysis.

Editions.csv (324 rows) carries the provenances, identifiers, character counts, eight quality indicators and screening flags (defined under Quality control). Yearly_corpora.csv records which editions compose each corpus and in what order. Pages.csv (2,931 rows) gives the block span of every source page with the method used to place its boundary and an independent validation status; and blocks.csv.gz (672,312 rows) profiles every OCR text block by page, content type and merge-risk grade, so that advertising, layout fragments and merges can be filtered without discarding whole editions.

Both gold-standard transcription sets are deposited in full for reference. All code used to compile the corpus is also deposited, so that the corpus is fully reproducible.

## (4) REUSE POTENTIAL

The OCR text corpus is 11.6 million characters from seven regional Republican-era newspapers, digitized with a Traditional-character pipeline and stratified by dialect region. The dataset supports a diachronic and dialectal study of early twentieth-century written Chinese. Because the yearly corpora are length-matched at 50,000 characters, vocabulary and character-frequency measures can be compared across years and across titles without renormalizing for sample size. Additionally, the varied dialect sampling allows researchers to study the pace of vernacularization across regions and dialects. Because the deposit documents every source edition and the full processing chain, others can extend the corpus to further titles or years, apply alternative metrics, or re-run the analysis with improved OCR, while the similarity tables and figures allow a direct replication of the reported findings.

The analyses for which the corpus was built illustrate what it can support (Merriam, 2025; expanded in Merriam, under review). The study demonstrates that mean pairwise similarity between the newspapers rose irregularly into the early 1920s, declined to a trough in the mid-1930s and rose again through the 1940s. On a segmented regression both the pre-1935 decline and the post-1935 recovery are significant (Vision: $-0.016$ and $+0.020$ per year; Tesseract: $-0.004$ and $+0.022$ per year), and both survive when the set of compared titles is held constant. Restricting the vectors to 84 grammatical characters showed that the late convergence, but not the earlier divergence, is present in grammatical vocabulary. A vernacularization index built from ten Classical and ten vernacular grammatical markers rose from 0.06 in 1910–1914 to 0.60 by 1945, taking off in the early 1920s immediately after the May Fourth Movement. The older cosmopolitan titles (Shenbao, Beijing Ribao) began almost fully Classical and converted gradually, whereas later regional titles entered the corpus already largely vernacular. Hierarchical clustering identified a Southwestern cluster (Xinxin Xinwen, Shangwu Ribao, Hunan Guomin Ribao), a Shenbao–Beijing Ribao pair, and the Yuehua Bao and Jiangsheng Bao as outliers, with the Southwestern cluster and the two outliers stable under both OCR engines. Each of these analyses can be reproduced from the deposited yearly corpora and similarity tables, and extended: the per-year corpora can be compared against

contemporaneous vernacular literature to date the diffusion of particular lexical or grammatical features, cross-referenced with regional dialect data, or used to study how quickly individual newspapers adopted vernacular forms.

The dataset can also aid with improving OCR transcription. The dataset pairs raw text with per-edition quality metrics and two gold-standard transcription sets, supporting error-pattern analysis and evaluation of recognition or post-correction models.

However, there are some barriers to reuse. While the text is in page order and page spans are supplied, as discussed above, the OCR merges parallel columns inside long blocks, so article boundaries cannot be recovered. The text is also difficult to read as the text segments are jumbled together. The corpus is not segmented by genre. Advertisements were not removed, so the yearly corpora mix news and advertising matter. Users who need news text only should filter the block-level content type. Temporal coverage is also uneven and weighted toward the January months, three newspaper-years fall below the length target, and the page images remain hosted by the Internet Archive. Regardless, the dataset provides a useful tool in conducting computational analyses on Chinese during this pivotal period.

## AI DECLARATION

Generative AI tools (Anthropic's Claude) were used to write utility scripts that extracted the OCR text, compiled the edition-level manifest, and tallied character counts. AI assistance was also used in the OCR evaluation: the 767-character gold-standard transcription was drafted with AI assistance from the page scans and then verified character by character by the author against the scans, and the second, 1,278-character gold standard was produced by a vision-language model (Claude, Anthropic) as an independent check. The underlying data were produced by the author's original OCR and analysis pipeline; generative AI was not used to generate, alter, or fabricate the corpus, the similarity results, or any reported figure, and all counts were verified by the author.

## ACKNOWLEDGEMENTS

The author thanks the Internet Archive for hosting the digitized newspaper collections that made this work possible.

## COMPETING INTERESTS

The author has no competing interests to declare.

## AUTHOR CONTRIBUTIONS

**Louis L. Merriam:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing.

## AUTHOR AFFILIATIONS


**Louis L. Merriam** orcid.org/0009-0004-1734-8019
Sidwell Friends School, Washington, D.C., USA

## REFERENCES

**Blouin, B., Huang, H.-H., Henriot, C.,** & **Armand, C.** (2023). Unlocking transitional Chinese: Word segmentation in modern historical texts. In *Proceedings of NLP4DH 2023* (pp. 92–101). Association for Computational Linguistics. https://aclanthology.org/2023.nlp4dh-1.11/

**Buntinx, V., Bornet, C.,** & **Kaplan, F.** (2017). Studying linguistic changes over 200 years of newspapers through resilient words analysis. *Frontiers in Digital Humanities*, *4*, 2. https://doi.org/10.3389/fdigh.2017.00002

**Chen, P.** (1999). *Modern Chinese: History and sociolinguistics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139164375

**Hengchen, S., Ros, R., Marjanen, J.,** & **Tolonen, M.** (2021). A data-driven approach to studying changing vocabularies in historical newspaper collections. *Digital Scholarship in the Humanities*, *36*(Suppl. 2), ii109–ii126. https://doi.org/10.1093/llc/fqab032

**Henke, K.,** & **Arnold, M.** (2022). Building an OCR pipeline for a Republican Chinese entertainment newspaper. In *Digital Humanities 2022: Conference abstracts*. Tokyo.

**Hill, M. J.,** & **Hengchen, S.** (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, *34*(4), 825–843. https://doi.org/10.1093/llc/fqz024

**Manning, C. D., Raghavan, P.,** & **Schütze, H.** (2008). *Introduction to information retrieval.* Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

**Merriam, L. L.** (2025). *The evolution of written Chinese (1910–1945): A computational study of newspapers [Preprint]*. Research Square. https://doi.org/10.21203/rs.3.rs-7745967/v1

**Merriam, L. L.** (2026). *Chinese Newspaper Similarity Dataset (1910–1945) [Data set]*. Zenodo. https://doi.org/10.5281/zenodo.20777803

**Merriam, L. L.** (under review). *The evolution of written Chinese (1910–1945): A computational study of newspapers*. Manuscript submitted for publication.

**Ren, T.** (2026). Transforming the Northwest frontier: Development discourse in Republican China through computational analysis of the historical press. *Humanities and Social Sciences Communications*, *13*, 334. https://doi.org/10.1057/s41599-026-06682-6

**Smith, R.** (2007). An overview of the Tesseract OCR engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition* (ICDAR 2007) (pp. 629–633). IEEE. https://doi.org/10.1109/ICDAR.2007.4376991

**TO CITE THIS ARTICLE:**
Merriam, L. L. (2026). The Evolution of Written Chinese (1910–1945): A Computational Newspaper Similarity Dataset. *Journal of Open Humanities Data*, 12: 144, pp. 1–6. DOI: https://doi.org/10.5334/johd.615

**Submitted:** 02 July 2026
**Accepted:** 07 October 2026
**Published:** 04 November 2026

**COPYRIGHT:**
© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See https://creativecommons.org/licenses/by/4.0/.

*Journal of Open Humanities Data* is a peer-reviewed open access journal published by Ubiquity Press.