BA2: Digital Korea — Final Paper[1,2]

Research Report with Replication Repository | Dr. Steven Denney | Korean Studies, Leiden University | Spring 2026

Workshop	Monday, 11 May 2026 (in class)
Due	Friday, 5 June 2026, 23:59 — submit on Brightspace
Length	2,500–6,000 words; 2,500 is a hard minimum, with a $\pm 10\%$ tolerance over the 6,000 maximum
Format	PDF only; the replication-repository URL is embedded as a footnote on the paper's title (see Replication and FAIR)

Overview

The final paper is a short research report. You generate a research question, answer it with computational text analysis methods learned in this course, and publish a public replication repository so any reader can re-run your analysis and confirm your findings. The paper is the headline product; the replication repository is what makes the work count as research.

Pick one corpus from the curated dataset menu (linked at the top of the page; the full menu is in the Appendix). The menu is deliberately different from the corpora you have used in class, so the one you pick will be new to you. Use it to answer one research question of your own.

1 The Week 12 workshop (11 May)

The workshop is for two things. First, I will review your draft research question and dataset choice with you in conversation. This is when to catch scoping or feasibility problems, before you commit. Second, you will workshop your analysis plan with classmates working on different corpora, who will see things you miss.

Come prepared. Arrive with a corpus picked from the menu and a draft research question phrased as a single sentence. The question and the dataset need to be linked: your question should be answerable with the corpus you chose. If you would prefer a corpus from `scdenney/nlp_corpora` that is not on the menu, or another corpus entirely, email me before the workshop. I will consider it, but it needs approval.

2 Required structure

Suggested word allocations are guidance; the 2,500–6,000-word total is what counts.

Introduction and research question (~200–500 words)

Open with the question and the motivation. State your research question clearly in one sentence. Briefly explain why it matters, and what kind of answer would count as informative. Specify the kind of answer you expect (e.g., "three to five distinct topical clusters whose prevalence varies across leader era").

[1] Dataset menu repository: `https://github.com/scdenney/ba2-final-paper-data`.

[2] Citation and formatting: BAKS Thesis Seminar Style Guide (use this for citation style and formatting only — not for length conventions). General guidance on writing a research report: Thesis & Research Supervision.

Brief literature review (~400–1,000 words; about 5 sources)

A modest, *preliminary* literature review — not a comprehensive survey. Cite around five sources that situate your question in scholarly conversation. At least one should be a methods source (e.g., a chapter from Grimmer, Roberts & Stewart, or a journal article using the same method). The remainder should be substantive Korea-area scholarship that your question engages. The aim is to anchor the question.

Data and methods (~500–1,000 words)

Describe *before* you show results: the corpus, the slice you analyze, your preprocessing decisions, and the methods you apply with reasons. A reader of this section should be able to predict roughly what your findings will contain.

Analysis and findings (~1,000–3,000 words; 1–3 figures or tables)

Apply your methods. Each figure or table is labeled clearly and referred to in the text. Figures are exported from Orange or generated in R, embedded inline in the PDF, and also placed in the `figures/` folder of your replication repository.

Conclusion (~400–800 words)

Summarize your headline finding. Reflect on the limitations of your analysis and what a follow-up project might do next. Mention anything that surprised you or that you couldn't fully explain.

3 Replication and FAIR

Create a *new*, *public* GitHub repository, separate from your weekly-coursework repository, that anyone can clone to reproduce your analysis. Structure it per the FAIR principles (Findable, Accessible, Interoperable, Reusable).

Tooling is your choice. You can do the analysis in Orange Data Mining (saved as a `.ows` workflow) *or* in R (one or more scripts). Submit whichever you used — you do not need to provide both. R is an option, not a requirement.

Required contents:

File / folder	Purpose	FAIR
`README.md`	Project description, RQ, headline finding, instructions to reproduce	F, A
`data/` *or* `data/SOURCE.md`	The CSV directly, or a `SOURCE.md` pointing to the menu repo URL + commit hash	F, A
`data/data_dictionary.md`	Column-by-column with types, units, examples	R
`analysis/`	Your Orange `.ows` workflow file *or* R script(s) — whichever you used	R, I
`figures/`	PNG exports of every figure that appears in your paper	F, A
`LICENSE`	MIT for code, CC-BY-4.0 for any data you produced	R
`CITATION.cff`	Author, paper title, course, date	F, R
`requirements.md`	Orange version, R version, non-default packages	I

Spot-checks: I will clone two or three replication repositories per cohort on a fresh machine and re-run the analysis; reproducibility failures count against this component of the grade.

Linking the repository from the paper

The repository URL must appear in the paper itself, as a footnote on the title. This follows the convention used when papers cite a Dataverse or OSF replication package. Use a statement along these lines (substituting your own URL):

> *Replication package — including the data, analysis workflow, exported figures, and a data dictionary — is publicly available at* `https://github.com/[your-username]/[your-repo-name].`

4 Submission

Submit your paper as a **single PDF** on Brightspace by 23:59 on **Friday 5 June 2026**. There is no separate field for the repository URL — the link lives in the title footnote of the paper itself.

5 Assessment

The paper is marked out of 10. Each of five components (research question 15%, methods application and interpretation 40%, replication repository / FAIR 20%, writing quality and structure 15%, brief literature review 10%) is marked on a 10-point scale; the weighted sum is the final paper grade. See the **rubric (PDF)** for per-component criteria at each band. A failing grade qualifies for a re-sit per the standing course policy.

6 Tips

- **Start with a question your dataset can answer.** Methods application is the largest component of the grade, but methods only do their work if your research question can actually be answered with the data you have. Spend the first part of your time making sure the question fits the corpus.
- Keep the literature review preliminary. About five sources is the spec; a long lit review eats the word count you need for findings.
- Build the replication repository as you go, not at the end. Commit the workflow file every time you make progress; commit figures as soon as you export them.
- Read the data first. The Corpus Viewer and Word Cloud widgets in Orange — or `glimpse()` and a quick word-frequency table in R — catch tokenization and encoding problems in 30 seconds.
- You can combine methods — clustering with sentiment, LDA with grouping by metadata, embeddings with clustering. Explain the combination in your data and methods section.
- Pre-modern or Hanmun-mixed corpora (colonial magazines, *Kaebyok*) need a Hanja-aware preprocessing step. Use `hanja_preprocessing_*.py` from the Data & Scripts page — it converts Chinese characters to their Hangul readings before tokenization, so Kiwi handles the text cleanly. Note in your data and methods section that the KNU sentiment dictionary is contemporary, so historical valence may not match perfectly.

Appendix — Dataset Menu

Eleven curated corpora. Pick one. Full per-dataset documentation (provenance, columns, suggested research questions) is in the menu repository.

Dataset	Rows	Best for
Authoritarian-era presidential speeches	600	Political rhetoric under Park, Chun, Roh
Inter-Korean summit coverage	451	Comparative media framing across summits
Colonial magazines (multi-title)	495	Colonial-period intellectual debates across 19 magazines
Kaebyok (single-title)	400	Single-magazine diachronic analysis 1920–1935
Korean newspapers (Twitter)	2,745	Cross-outlet ideology and engagement, 6 outlets in 2017
KPoEM (Korean poems)	615	Sentiment / emotion analysis on a labeled corpus
Immigrant interviews (open-text)	1,006	Citizen voice on immigrant preference
NK migrants interviews (open-text)	6,023	Citizen voice on North Korean migration
Korean newspaper archive (modern slice)	2,000	Diachronic / cross-newspaper analysis of the late-colonial press
Rodong Sinmun (English)	600	Temporal analysis of DPRK propaganda 2018–2021
NIKH high-school textbooks (auth. + dem. era)	21	Textbook narration across authoritarian and democratic curriculum eras (1973–2016)

Rodong Sinmun (#10) is English text, so the KNU sentiment dictionary and KLUE BERT do not apply; pick a method that works on English. Colonial magazines and *Kaebyok* are Hanmun-mixed — use the `hanja_preprocessing_*.py` script (Data & Scripts page) for clean tokenization.