BA2: Digital Korea

Week 2: Foundations of Computational Text Analysis

Steven Denney

Korean Studies
Leiden University

February 9, 2026

Today's Agenda

1. About you
2. Week 1 check-in
3. Markdown & GitHub Desktop
4. Foundations: Social science research & text analysis
5. CTA, CSS & the digital humanities
6. Corpora and data sources
7. Demo: Loading a corpus in Orange Data Mining
8. Looking ahead

Welcome

About You

Let's go around the room.

Week 1 Check-In

Swirl R Programming lessons:

- 1: Basic Building Blocks
- 2: Workspace and Files
- 4: Vectors
- 6: Subsetting Vectors
- 7: Matrices and Data Frames
- 12: Looking at Data

Quick poll

How did it go? Any questions or difficulties?

Markdown & GitHub

What Is Markdown?

Markdown is a lightweight markup language for formatting plain text.

Why use it?

- Simple, readable syntax
- Renders beautifully on GitHub
- Standard for documentation
- Used in README files, wikis, notebooks

Where you'll see it

- GitHub repositories (README.md)
- This course website
- R Markdown documents
- Jupyter notebooks
- Research documentation

Markdown Basics

What you type:

```
# Heading 1
## Heading 2

**bold** and *italic*

- Bullet point
- Another point
  - Sub-point

1. First item
2. Second item

[Link text](URL)
```

What you get:

Heading 1

Heading 2

bold and *italic*

- Bullet point
- Another point
  – Sub-point

1. First item
2. Second item

Link text

More Markdown

Code blocks:

````
```
print("Hello, world!")
```
````

Tables:

```
| Column A | Column B |
|----------|----------|
| Data 1   | Data 2   |
```

Images:

```
![Alt text](image.png)
```

Blockquotes:

```
> This is a quote.
```

Horizontal rule:

```
---
```

Full reference: `https://www.markdownguide.org/basic-syntax/`

Markdown on GitHub

- GitHub automatically renders `.md` files
- Every repository should have a README.md
- Use Markdown for project documentation, research notes, data descriptions
- GitHub "flavored" Markdown adds extras: task lists, tables, emoji

Up next: GitHub Desktop demo

We'll use GitHub Desktop to create folders in your local repo, write a Markdown file, commit, and push to GitHub.

Foundations of Computational Text Analysis

The Research Process: Two Models

(Grimmer, Roberts & Stewart, Ch. 2)

Deductive (linear)

1. Start with a theory
2. Derive hypotheses
3. Collect data to test them
4. Report results

Assumes theory comes *before* data.
Useful, but not always how research with text works.

Iterative (cyclical)

1. Start with a question or dataset
2. Explore the data
3. Refine your question
4. Develop hypotheses
5. Collect new data to test
6. Report results

Embraces discovery. Especially suited to text analysis.

Three Stages of the Research Process

1. **Discovery**
   - Develop the research question
   - Build conceptualizations
   - Find patterns in data
   - Explore what is there
2. **Measurement**
   - Quantify concepts
   - Assign categories
   - Measure prevalence
   - Validate measures
3. **Inference**
   - Make predictions
   - Estimate causal effects
   - Test hypotheses
   - Generalize findings

These stages don't have to happen in order – and often don't.

Case Study: Chinese Internet Censorship

King, Pan & Roberts (2013) – a story of iterative research:

1. Started with automated text analysis for Chinese social media
2. Discovered posts were disappearing – government censorship
3. New question: What types of posts get censored?
4. Measured sentiment and topic of censored vs. uncensored posts
5. Found censorship targets collective action, not government criticism
6. Validated with a randomized field experiment (KPR, 2014)

Lesson

They didn't start with the censorship question. The data led them there – through exploration, iteration, and substantive knowledge of Chinese politics.

Why Does This Matter for Us?

Think about Korean Studies research:

- Analyzing North Korean state media – what patterns emerge?
- Exploring presidential speeches – how does language shift across administrations?
- Mining online discourse – what topics dominate, and when?
- Reading colonial-era periodicals – how do themes evolve over decades?

The point

You may start with one question and end up somewhere more interesting. That's not a failure – it's how good research with text data works.

Six Principles of Text Analysis

GRS offer six principles (Table 2.1) to guide our work:

1. Substantive knowledge is essential for research design
2. Text analysis augments humans – it doesn't replace them
3. Building and testing theories requires iteration and cumulation
4. Text analysis methods distill generalizations from language
5. The best method depends on the task
6. Validations are essential and depend on theory and task

We will return to these principles throughout the course.

Key Takeaways from Chapter 2

- Research with text data is iterative, not purely deductive
- The process has three stages – discovery, measurement, inference – and they interact
- Computational methods augment human reading, including close reading and archival work
- Your substantive knowledge of Korea is what makes text analysis meaningful
- There is no single "best" method – it depends on the task and the question

CTA, CSS & Digital Humanities

Where Does CTA Fit?

Computational Social Science

- Uses computational tools to study social phenomena
- Large-scale data analysis
- Prediction, causal inference
- Often quantitative

Digital Humanities

- Applies digital tools to humanistic inquiry
- Text mining, digitization, cultural analytics
- Extends traditions of close reading
- Bridges qualitative and quantitative

Computational text analysis sits at the intersection – and is used by both fields.

CTA in Korean Studies

Text analysis is increasingly applied in research on Korea:

- Historical text digitization and analysis (colonial-era periodicals, historical archives)
- Political discourse analysis (presidential speeches, legislative records)
- Media studies (news coverage, North Korean state media)
- Online discourse and public opinion (social media, forums)
- Literary and cultural analysis

This course

We focus on methods that apply across all of these – using Korean-language primary sources as our material.

Corpora & Data Sources

What Is a Corpus? (Revisited)

Corpus (pl. *corpora*): A structured collection of texts assembled for analysis. Now, with Chapter 2 in mind:

- A corpus is not just a pile of texts – it reflects choices
- What you include (and exclude) shapes what you can discover and measure
- Metadata matters: who, when, where, what type?
- The corpus is the foundation of everything we do in this course

Think about

What kind of corpus would *you* want to build for a research question about Korea?

Our Course Corpora

Repository

`https://github.com/scdenney/nlp_corpora`

Available collections include:

- Presidential speeches (이승만 – 문재인)
- Newspaper articles (Korean press)
- Periodicals (개벽 / *Kaebyŏk*)
- Moon Jae-in Twitter posts
- Interview data (immigrants, North Korean migrants)
- And more to come

We will work with truncated versions in class to keep things manageable.

Today's Corpus: Presidential Speeches (Democratic Era)

For the demo today, we use a sample of Korean presidential speeches:

- Democratic-era presidents only (6th Republic onward)
- 749 speeches, proportionally sampled from 5,840
- Variables: president, title, date, location, speech type, full text

President	Speeches
Roh Tae-woo (노태우)	77
Kim Young-sam (김영삼)	93
Kim Dae-jung (김대중)	106
Roh Moo-hyun (노무현)	100
Lee Myung-bak (이명박)	132
Park Geun-hye (박근혜)	63
Moon Jae-in (문재인)	178

Orange Data Mining Demo

Orange Data Mining: Quick Recap

- Widget-based, drag-and-drop analysis platform
- You should have watched Getting Started tutorials 01–04
- Today: We load a real corpus and explore it

Tutorials to review:

1. Getting Started 01: Welcome to Orange
2. Getting Started 02: Data Workflows
3. Getting Started 03: Widgets and Channels
4. Getting Started 04: Loading Your Data

Playlist: `https://www.youtube.com/playlist?list=PLmNPvQr9Tf-ZSDLwOzxpvY-HrE0yv-8Fy`

Demo: Loading and Exploring the Corpus

Step 1: Get the data

1. Go to the course website – Data page
2. Download `president_speeches_democratic_era.csv`
3. Add it to a subfolder in your repo (e.g., `/data/president_speeches/`)
4. Commit and push via GitHub Desktop

Step 2: Load in Orange

1. Open Orange Data Mining
2. Add a Corpus widget to the canvas
3. Load the CSV from your `/data` folder
4. Explore the corpus – browse speeches, filter, search

Follow along

Open Orange on your computer and follow each step.

Looking Ahead

For Next Week

Week 3: Text Preprocessing Basics

- Tokenization – breaking text into units
- Part-of-speech (POS) tagging
- The preprocessing pipeline

Recommended reading:

- Grimmer, Roberts & Stewart – Chapter 5: Bag of Words
- Denny & Spirling (2018) – Text preprocessing for unsupervised learning

Orange tutorial:

- Getting Started 16: Text Preprocessing

For Next Week

R Programming:

- DataCamp: Introduction to Text Analysis in R – Chapter 1: Wrangling Text

Recommended (optional) reading:

- Wilkerson & Casas (2017) – Large-scale text analysis in political science
- Macanovic (2022) – Text mining for social science

Remember: Widget Catalog is your reference:

- `https://orangedatamining.com/widget-catalog/`