

Text Preprocessing

BA3 Text as Data · Session 2 · 25 September 2026

Steven Denney

Korean Studies, Leiden University

Class objective

By the end of class, you can apply and document a small set of preprocessing choices and explain how those choices change the text available for analysis.

Attendance and review questions

Scan the QR code with your phone and take the survey for attendance and review.

No (smart) phone? Ask me for a piece of paper.



Push A commit stays on your laptop until you click Push origin.

Pull Fetch origin, then Pull origin, brings changes from github.com.

Orange video Tokenization splits a text into units.

Before/after Save the settings and compare the processed text with the original.

Today you push your first preprocessing workflow.

Roadmap for today

When	What	Minutes
09:15	Opening survey and review	10
09:25	What preprocessing does	45
10:10	Orange demonstration, raw speeches to word cloud	35
10:45	Checklist and next week	15

Screens away

No computers or phones

Sample sentence from Moon Jae-in

Moon Jae-in, first-anniversary message

“국민이 문재인정부를 세웠다는 사실을 결코 잊지 않습니다.”

May 2018 · 100-speech course corpus

Sample sentence from Moon Jae-in

Moon Jae-in, first-anniversary message

“국민이 문재인정부를 세웠다는 사실을 결코 잊지 않겠습니다.”

May 2018 · 100-speech course corpus

I will never forget that the people established the Moon Jae-in government.

What is preprocessing?

Preprocessing · 전처리

The steps that turn raw text into “clean” *units* for analysis.

Preprocessing addresses three main things:

- | | |
|--|---------------|
| 1. What gets counted as one unit? | tokenization |
| 2. Which forms of a word count as the same unit? | normalization |
| 3. Which units are deliberately left out? | filtering |

Seven terms for the preprocessing steps

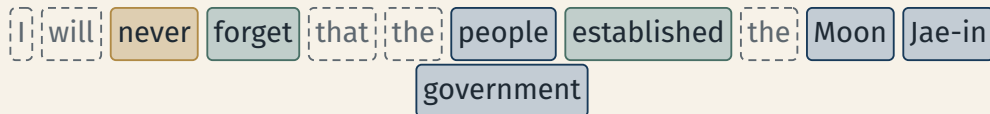
Term	Meaning	Example
Token	One unit the software counts	people
Tokenization	Splitting text into tokens	I · will · never · forget
Normalization	Harmonizing words/units	established → establish
Lemma	The base form of a word	establish
Stopword	A “useless” word to remove	I, the, that
Part of speech (POS)	The word class of a token	noun, verb, adverb
Filtering	Dropping tokens as per rules	stopwords, POS, length

A sample pipeline

Step	Result
Raw text	I will never forget that the people established the Moon Jae-in government.
Clean	I will never forget that the people established the Moon Jae-in government
Tokenize	I · will · never · forget · that · the · people · established · the · Moon · Jae-in · government
Normalize	i · will · never · forget · that · the · people · establish · the · moon · jae-in · government
Filter	never · forget · people · establish · moon · jae-in · government

Normalizing lowercased the tokens and reduced *established* to its lemma. Filtering removed the stopwords.

A token and a type



12 tokens

Every word occurrence, counted separately.

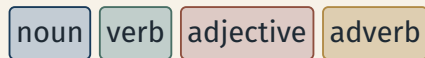
11 types

the appears twice but is one type.

The set of types in a corpus is its **vocabulary**.

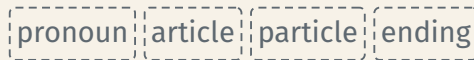
Parts of speech (POS) sort tokens into content and grammar

Content words



What a text is about

Grammar



Tense, case, and politeness.

A **POS tag** labels each token with its class. Tags let us keep only the classes a question needs.

In English and Dutch

English I will never forget that the people established the Moon
Jae-in government

Dutch Ik zal nooit vergeten dat het volk de regering-Moon tot
stand bracht

noun verb adjective adverb grammar

Splitting at spaces gives usable tokens in both languages.

Korean text can be split at five levels

Level	Example	Units
Sentence · 문장	국민이 문재인정부를 세웠다는...	1
Eojeol · 어절	국민이	7 in the sentence
Morpheme · 형태소	국민 + 이	16
Syllable · 음절	국 · 민 · 이	25
Letter · 자모	ㄱ ㅌ ㄱ · ㅁ ㄴ · ㅇ	

An **eojeol** is a space-separated unit. A **morpheme** is the smallest unit with meaning or grammatical function. Text analysis of Korean counts morphemes.

Each Korean eojeol has several morphemes

7 eojeol 국민이 문재인정부를 세웠다는 사실을 결코 잊지 않겠습니다

16 morphemes 국민 이 문재인 정부 를 세우 었 다는 사실 을 결코
잊 지 않 겠 습니다

noun verb adjective adverb grammar

Only 4 of the 16 morphemes are nouns. The rest are verbs, an adverb, and grammar.

Korean POS tags name each morpheme

Kiwi, the analyzer we use, labels morphemes with the Sejong tag set.

Content · usually kept

NNG	general noun	국민
NNP	proper noun	문재인
VV	verb	세우
VA	adjective	강하
MAG	adverb	결코

Grammar · usually dropped

JKS	subject particle	이
JKO	object particle	를, 을
EP	tense ending	었, 겠
EF	final ending	습니다
VX	auxiliary verb	않

The course script keeps NNG and NNP. You can add VV and VA.

Three features of Korean make it hard to use the same splitting tools

Particles attach 국민이 → 국민이 사실을 → 사실

Endings stack 세웠다는 → 세우었다는 않겠습니다 → 않겠습니다

No space, two words 문재인정부를 → 문재인 정부

English and Dutch tools assume that a space ends a word. In Korean it often does not.

Oh nee! Orange does not preprocess Korean

Problem Orange's text tools are built for languages such as English and Dutch. They do not split Korean into morphemes or tag Korean POS.

Solution A short script runs Kiwi, a Korean morphological analyzer, inside Orange's Python Script widget.

Your job Copy the script I wrote for you, run it. No programming is expected. You can tinker!

Analyzers still make mistakes, especially with names and new words. Check the output against the text.

Kiwi splits Korean into morphemes and tags each one

Kiwi

Korean Intelligent Word Identifier, a free, open-source morphological analyzer by bab2min. It installs as `kiwipiepy`.

국민이 문재인정부를 → 국민/NNG 이/JKS 문재인/NNP 정부/NNG 를/JKO

- It returns each morpheme's base form and its POS tag.
- It splits words written without a space, such as 문재인정부.
- The course script installs it on the first run, on Windows, macOS, and Linux.

The script keeps nouns, and you can add verbs

국민 이 문재인 정부 를 세우 었 다는 사실 을 결코 잊 지 않 겠 습니다

Nouns only (course default)

국민 · 문재인 · 정부 · 사실

Who and what the sentence mentions.

The script also drops one-syllable tokens and numbers. You'll also use **stopwords**.

Nouns, verbs, and adjectives

국민 · 문재인 · 정부 · 세우 · 사실

Adds the action, 세우 (to establish).

Stopwords are frequent words that a list removes

English I will that the

Dutch ik zal dat het de

Korean 것 우리 있다 이번 모두 동안 시간

The **course stopwords file** (let op!) lists 679 Korean words. Orange loads it in Preprocess Text, after the script.

A list is a choice. 시간 (time) is on it, but a question about time needs it back.

Different filters

Filter	Rule	Removes	Set in
Part of speech	keep NNG, NNP	이, 를, 었, 세우	script
Minimum length	at least 2 characters	남, 일	script
Maximum length	at most a set number	long strings, links	not used
Numbers	digits only	2018	script
Stopword list	words in a file	이번, 동안, 이상	Preprocess Text
Document frequency	minimum and maximum share of documents	words in almost every speech	Session 3

Every filter you use goes in your README.

The app runs these steps on any sentence from the speeches

1 Raw text

2 Clean

3 Tokenize

4 Tag parts of speech (POS)

5 Filter

6 Result

“열집은 남의 것을 얻어서 자식을 먹였고, 우리 부모는 자림심을 키워 주고 희망을 주었습니다.”

이명박, 미국 NPR 라디오 인터뷰

Keep POS

NNG, NNP

Stopwords

있다 없다 되다 하다 그 저 이 것 등 및 수 때 년 월 일 더 또 속 통해 위해

Minimum length

2 characters

Numbers

dropped

☐ Also keep verbs and adjectives (VV, VA)

열집

은

남

의

것

을

얻

어서

자식

을

먹

었

고

우리

부모

는

자림심

을

키우

주

고

희망

을

주

었

습

니

다

Previous

Filtering keeps nouns and drops stopwords, one-character tokens, and numbers.

Next

noun

verb

adjective

adverb

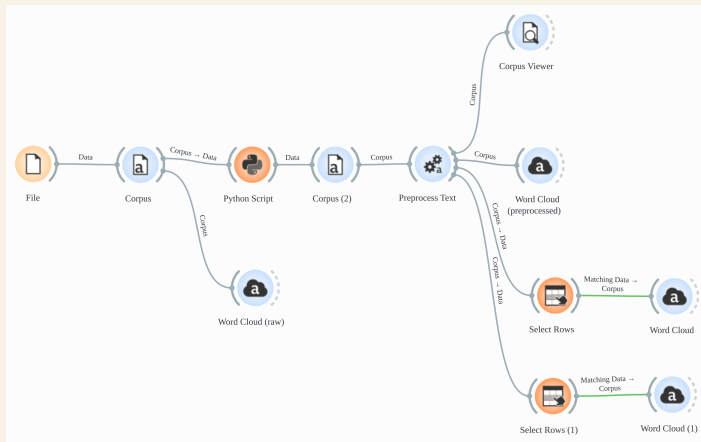
grammar

Random sentences, six steps, and the verb toggle. <https://scdenney.github.io/ba3tad-26/interactive/preprocessing-pipeline/>

Laptops open

Orange first. We build it together.

Your Session 2 workflow in Orange



The snapshot is also on the Slides page.

Before you leave

What must be finished, submitted, and prepared next

Demo 2 checklist

- ☐ The workflow or script opens from the repository.
- ☐ The README lists the corpus and the preprocessing settings.
- ☐ The word cloud of the preprocessed speeches is saved as `output/wordcloud.png`.
- ☐ Everything is committed and pushed.

Demo 2 is due Oct. 2, 09:15

Files go in `demos/02-preprocessing/`. Graded for checklist completion.

Next week counts the words you kept

Demo 2

Finish the checklist and push.

Due Oct. 2 at 09:15.

Also watch the Orange Bag of Words tutorial. Details are on the Session 3 page.

Programming exercises

DataCamp Introduction to R, chapter 1, checked online.

Appendix

Key terms and definitions

Appendix · Units of text

Token	One unit the software counts after preprocessing.
Type	One distinct form. <i>the</i> twice is two tokens but one type.
Vocabulary	The set of types in a corpus.
Tokenization	Splitting a text into tokens.
Eojeol · 어절	A space-separated unit in Korean, often a word plus particles or endings.
Morpheme · 형태소	The smallest unit with meaning or grammatical function, such as 국민 or 이.
Syllable · 음절	A written Hangul block such as 국. Its letters, ㄱ ㅊ ㄱ, are jamo (자모).

Appendix · Preprocessing steps

Preprocessing · 전처리	The steps that turn raw text into units a computer can count and compare.
Pipeline	The steps applied in a fixed order. Clean, tokenize, normalize, then filter.
Cleaning	Removing punctuation, symbols, and links.
Normalization	Making different forms count as one, such as established and establish.
Lemma	The base form of a word, such as establish.
Stopword	A frequent word removed before counting. A stopwords list is a choice.
Filtering	Dropping tokens by a rule, such as part of speech, length, number, list, or document frequency.

Appendix · Grammar and tags

Part of speech (POS)	The word class of a token, such as noun, verb, or adverb.
POS tag	A label for a token's part of speech. Kiwi uses Sejong tags such as NNG, NNP, VV, VA, and MAG.
Content word	A noun, verb, adjective, or adverb. It carries what a text is about.
Grammar	Particles, endings, and other forms that connect content words.
Particle · 조사	A form attached to a noun, such as 이 or 를.
Ending · 어미	A form attached to a verb stem, such as 었 or 습니다.
Morphological analyzer	Software that splits Korean into morphemes and tags them. We use Kiwi.

Appendix · Orange widgets

File	Opens the CSV file.
Corpus	Turns the table into a corpus. Set the text feature, the column Orange analyzes.
Python Script	Runs the course script, which adds the processed_text column.
Preprocess Text	Tokenizes and filters. The course stopwords file is loaded here.
Corpus Viewer	Shows raw and processed text side by side.
Word Cloud	Shows the most frequent tokens, sized by count.