scientific data

OPEN

DATA DESCRIPTOR

South Korean Election Campaign Booklet and Party Statements Corpora

Tae Hyun Lim

This descriptor presents two comprehensive text datasets from South Korean politics: the South Korean Election Campaign Booklet Corpus and the South Korean Party Statements Corpus. The Election Campaign Booklet Corpus comprises manifesto booklets from 49,678 individual candidates who ran for offices in single-member or multi-member districts during six National Assembly elections, six local elections, and five presidential elections in South Korea between 2000 and 2022. The Party Statements Corpus contains 82,723 official statements released by the two major political parties in South Korea from 2003 to 2022. As primary sources of political communication—the campaign materials used by individual candidates (Election Campaign Booklet Corpus) and the official communications from party leadership (Party Statements Corpus)—these datasets provide a novel opportunity for researchers to analyze inter-party and intra-party variations in political communication, political agendas, and policy positions in South Korea during the twenty-first century. Furthermore, these datasets can contribute to comparative studies in political science and communication studies by offering insights into the dynamics of election campaigning and agenda setting in a democratic context.

Background & Summary

This descriptor introduces two comprehensive text datasets from South Korean politics: the South Korean Election Campaign Booklet Corpus and the South Korean Party Statements Corpus. These datasets offer valuable resources for analyzing political communication, policy positions, and agenda setting in South Korea during the twenty-first century.

The South Korean Election Campaign Booklet Corpus is a collection of manifesto booklets from 49,678 individual candidates who ran for offices in single-member or multi-member districts during six National Assembly elections, six local elections, and five Presidential elections in South Korea between 2000 and 2022. Due to stringent restrictions on media campaigns—such as prohibitions on television and radio advertisements for candidates—election campaign booklets serve as the primary medium for candidates to reach mass voters. With no alternative mass media channels available, these booklets are crucial for candidates to communicate their policies and messages to the electorate.

The South Korean Party Statements Corpus comprises 82,723 official statements from party spokespersons and minutes from daily leadership meetings of the two major political parties in South Korea, covering the period from 2003 to 2022. Party statements are the primary means by which party leadership communicates with the public on a daily basis. These statements are posted on the parties' official websites and are frequently reported by the media, making them influential in shaping public discourse and party image.

The text data for the Election Campaign Booklet Corpus were collected from the National Election Commission's (NEC) website, while the Party Statements Corpus was compiled from the respective parties' official websites. The texts from the Election Campaign Booklets were extracted from PDF files using an automated text extraction method—Optical Character Recognition (OCR). Both datasets were parsed using the khaiii Python library.

The Election Campaign Booklet Corpus contributes to the literature in three ways. First, most existing data sets or articles focused on social media postings[1–3], campaign ads[4–7], or official speeches[8]. These communication methods allow effectively unlimited spaces for the politicians to convey their messages[9]. Unlike these datasets, the corpus derives from a highly regulated communication medium with strict constraints on booklet size, page count, and distribution. Such page constraints create a forced trade-off that reveals candidates' strategic

Hamilton College, Clinton, USA. e-mail: tlim@hamilton.edu

Branch of the Offices	National Government	Local Government	
		Metro	Basic
Executive	President	Metropolitan Mayor/Governor	District Head/Mayor
		Education Superintendent	
Legislature	National Assembly	Metro Assembly	Basic Assembly

Table 1. Types of Elections in South Korea.

priorities. Candidates must carefully allocate limited space, making these booklets ideal for analyzing their decision-making processes in campaign communications.

Second, most data sets or articles focused on higher level offices, such as Presidents, Senatorial, Congressional, or Gubernatorial races. This corpus provides campaign messages from the candidates spanning from presidential races to basic assembly elections. This captures messaging across all electoral levels.

Lastly, whereas almost all existing datasets, with the notable exception of the Wesleyan Media Project[6], concentrate on one or two election cycles, the Election Campaign Booklet dataset spans more than two decades of elections. This extended timeframe, combined with the diversity of offices represented, enables researchers to track genuine issue attention and strategic calculations across different office levels, office types, and party affiliations over time.

The companion South Korean Party Statements Corpus contains both official party leadership remarks and starting remarks of daily leadership meetings, distinguishing it from datasets featuring only official statements[10]. Unlike the Manifesto Project Database[11], which provides coded results of party manifestos, the Party Statement Corpus offers raw text for researcher-directed analysis. It captures less refined remarks made during party leadership meetings, contrasting with datasets of parliamentary speeches[12,13]. This allows researchers to observe position formation, internal decision-making processes, and intra-party differences outside formal parliamentary settings.

As the primary campaign materials in South Korean elections (Election Campaign Booklet Corpus) and the primary sources of communication from party leadership (Party Statements Corpus), these datasets provide a novel opportunity for researchers to analyze inter-party and intra-party variations in strategic political communication, political agendas, and policy positions across electoral levels, time periods, and institutional contexts in South Korea in the twenty-first century. Moreover, as rare datasets containing both individual candidates' campaign materials and party leadership's statements, these datasets can contribute to comparative studies in political science and communication studies by offering insights into the dynamics of election campaigning and agenda setting in a democratic context. These datasets can facilitate a wide range of research topics, including comparative policy analysis, political communication strategies, agenda setting and issue salience, and intra-party dynamics.

Part of this dataset was initially collected and analyzed for the author's dissertation project[14], which focused on data created from 2000 to 2012. The current dataset expands upon the original by including data created after 2013, which have not been used in any other research to date. To the best of the author's knowledge, aside from the aforementioned dissertation project, there is currently no dataset or study using the full text of Election Campaign Booklets. While recent advancements in web crawling have enabled some researchers to analyze portions of party statements—for example, Park and Lee analyzed party spokespersons' remarks[15]—there has been no dataset or research examining the full extent of party statements either. The dataset introduced in this descriptor fill this gap and offer valuable resources for scholars interested in South Korean politics and comparative political communication. Example use cases are provided in Supplementary Information Appendix A.

Methods

South Korean election campaign booklet corpus. I collected Election Campaign Booklets of individual candidates running for single- and multi-member districts in local and national offices in regular elections held between 2000 and 2022 in South Korea from the National Election Commission (NEC) of South Korea website (www.nec.go.kr) using Python scripts. This comprehensive dataset encompasses campaign materials from various types of elections, providing an extensive overview of candidates' messages. In this section, I outline the types of elections in South Korea, describe the specifics of the Election Campaign Booklets, detail the data collection and processing methods, and discuss the construction of the final dataset.

Types of elections in South Korea. South Korea holds three main types of elections: presidential elections, National Assembly elections, and local elections. Table 1 presents the types of elections and the offices involved.

- Presidential Elections: Held every five years, where voters elect the president, the head of the national executive branch.
- National Assembly Elections: Occur every four years, with voters electing members of the National Assembly, which constitutes the national legislative branch.
- Local Elections: Also held every four years, involving the election of officials for seven offices in both higher-level (metropolitan) and lower-level (basic) local governments.

The third and fourth columns of Table 1 detail the offices within local governments. Metropolitan cities like Seoul elect a metropolitan mayor, an education superintendent, metro assembly members, district heads, and basic assembly members. Provinces like Gyeonggi elect a governor, an education superintendent, metro

assembly members, mayors, and basic assembly members. I collected Election Campaign Booklets from all these offices for regularly held elections between 2000 and 2022. I did not collect the manifestos from by-elections. The Election Campaign Booklets for candidates running for Education Superintendent are collected from 2010 because the 2010 local elections were the first time the election for these offices were held nationally. The Election Campaign Booklets for candidates running for metro assembly and basic assembly are collected from 2006 because the NEC website does not provide the booklets for candidates running for these offices in 2002.

Election campaign booklets. Candidates are permitted to submit an official document called an Election Campaign Booklet to the NEC, which is then mailed to voters approximately two weeks before the election date.

- Before 2005: Candidates for the National Assembly or heads of local governments could submit up to eight pages, known as a "small election booklet (소형인쇄물)," containing their name, party affiliation, policy pledges, and personal information.
- After 2006: The document was renamed the "Election Campaign Booklet (공보물)," including additional details such as tax history, criminal records, and agendas. The page limit was set to eight pages for metro and basic assembly candidates, twelve pages for National Assembly and the heads of local governments, and sixteen pages for presidential candidates.

These booklets are the primary source of wide-reaching campaign material due to strict campaign regulations in South Korea. Radio and TV advertisements by individual candidates are permitted only on limited occasions, mostly for presidential elections. For example, in the 2020 National Assembly Election, the average number of voters per district was over 200,000, ranging from approximately 139,000 to 278,000 voters. With limited alternative methods to reach such a large electorate, candidates rely heavily on these booklets, crafting concise and focused messages about the issues and policies they prioritize. The booklets rarely mention competitors or policies the candidates oppose, making them valuable resources for analyzing candidates' core pledges and campaign strategies.

Data acquisition. I developed custom Python scripts to scrape individual booklets from the NEC website for:

- Presidential Elections: 2002, 2007, 2012, 2017, and 2022
- National Assembly Elections: 2000, 2004, 2008, 2012, 2016, and 2020
- Local Elections: 2002, 2006, 2010, 2014, 2018, and 2022

Most candidates' booklets were provided as individual PDF files, one per candidate. However, for the elections held in 2000, 2002, and 2004, the booklets were available only as collections of JPEG images encompassing all candidates' materials by region, rather than as individual consolidated PDFs for each candidate. In the 2006 election, the format varied depending on the candidate; some booklets were available as individual PDFs, while others were provided as collections of JPEG images. For the booklets provided in JPEG format, I downloaded each page individually, then concatenated the pages to create a single PDF file for each candidate using the page numbers provided by the NEC. These reconstructed PDFs were then processed similarly to the PDFs from later elections.

Text extraction and processing. I employed Optical Character Recognition (OCR) to extract text from the booklets, as the content was embedded as images rather than selectable text. Three OCR methods were tested using Python:

1. Full-color, whole PDFs using Google Drive API
2. Black-and-white, page-by-page using Tesseract
3. Black-and-white, page-by-page using Google Drive API

The final dataset contains the OCR results from the third method—black-and-white, page-by-page using Google Drive API—as it yielded the highest accuracy in text extraction. Details of the methods and a comparison of their results are presented in Technical Validation section.

After OCR processing, I parsed the text using the Python library khaiii for morphological analysis. This step involved tokenization and part-of-speech tagging, which are crucial for subsequent text analysis.

The khaiii (Kakao Hangul Analyzer III) is a Korean lexical analyzer Python library using deep learning developed by Daum Kakao, the leading internet texting company in South Korea. The khaiii library is designed for morphological analysis of Korean text. It performs tokenization and part-of-speech tagging, which are essential for analyzing the Korean language due to its agglutinative structure. Moreover, based on the Sejong Corpus, a Korean text corpus built by the National Institute of Korean Language, khaiii provides ways to include new words in its dictionary, which helps to improve parsing.

Biographical data collection and integration. I collected candidates' biographical information provided by the NEC from South Korea's public data portal API (www.data.go.kr). This data included: election date, office identifier, district, metropolitan region, party, name, gender, birthday, age, job identifier, job, education identifier, education, two careers, and candidate identifier. I merged the biographical data with the booklet data in the final dataset.

Despite the NEC providing both biographical and booklet data, inconsistencies and errors existed in the formatting of regional and district names, as well as in candidate identifier numbers across the two datasets. These

Year	Office	Total Entries	Missing Booklets		Missing Text		Missing Bio	
			Frequency	%	Frequency	%	Frequency	%
2000	National Assembly	1,050	108	10.3	108	10.3	10	1.0
2002	Basic Head	756	64	8.5	64	8.5	6	0.8
	Metro Head	58	11	19.0	11	19.0	3	5.2
	President	7	1	14.3	1	14.3	0	0.0
2004	National Assembly	1,186	60	5.1	60	5.1	11	0.9
2006	Basic Assembly	7,994	200	2.5	200	2.5	0	0.0
	Basic Head	848	33	3.9	33	3.9	0	0.0
	Metro Assembly	2,176	176	8.1	176	8.1	108	5.0
	Metro Head	69	4	5.8	4	5.8	3	4.3
2007	President	12	2	16.7	2	16.7	0	0.0
2008	National Assembly	1,119	11	1.0	16	1.4	0	0.0
2010	Basic Assembly	5,867	732	12.5	743	12.7	3	0.1
	Basic Head	780	50	6.4	53	6.8	0	0.0
	Education Superintendent	81	6	7.4	6	7.4	0	0.0
	Metro Assembly	1,778	105	5.9	106	6.0	0	0.0
	Metro Head	58	2	3.4	2	3.4	0	0.0
2012	National Assembly	930	35	3.8	35	3.8	2	0.2
	President	7	0	0.0	0	0.0	0	0.0
2014	Basic Assembly	5,413	50	0.9	50	0.9	1	0.0
	Basic Head	729	36	4.9	36	4.9	2	0.3
	Education Superintendent	72	1	1.4	1	1.4	0	0.0
	Metro Assembly	1,736	20	1.2	20	1.2	0	0.0
	Metro Head	61	4	6.6	4	6.6	0	0.0
2016	National Assembly	944	8	0.8	8	0.8	0	0.0
2017	President	15	2	13.3	2	13.3	0	0.0
2018	Basic Assembly	5,337	49	0.9	50	0.9	1	0.0
	Basic Head	757	9	1.2	9	1.2	0	0.0
	Education Superintendent	61	2	3.3	2	3.3	0	0.0
	Metro Assembly	1,890	28	1.5	29	1.5	1	0.1
	Metro Head	71	0	0.0	0	0.0	0	0.0
2020	National Assembly	1,118	20	1.8	20	1.8	0	0.0
2022	President	14	0	0.0	0	0.0	0	0.0
	Basic Assembly	4,445	320	7.2	320	7.2	0	0.0
	Basic Head	580	18	3.1	18	3.1	0	0.0
	Education Superintendent	61	4	6.6	4	6.6	0	0.0
	Metro Assembly	1,543	112	7.3	112	7.3	0	0.0
	Metro Head	55	1	1.8	1	1.8	0	0.0
Total		49,678	2,283	4.6	2,306	4.6	151	0.3

Table 2. Number and Proportion of Missing Manifestos, Missing Text, and Candidate Bios by Year and Office.

discrepancies posed challenges when attempting to merge the biographical data with the text data. Additionally, some candidates had missing information; Table 2 presents the number and proportion of missing entries.

To merge the biographical data with the booklet data, I first matched records based on the following key attributes: election date, metropolitan region, office, party, and candidate name. Then, I deleted the entries where the first two letters of district names in booklet did not match those in bio data. This process let me identify 291 duplicate entries—instances where candidates running in the same region, from the same party, for the same office, had identical names, and had the same first two letters for their district names. To resolve these duplicates, I manually reviewed each case by checking district names and biographical information. This examination led to the removal of 55 erroneous entries from the dataset.

Even after these steps, 2,283 candidates lacked a booklet and 151 were missing biographical information. Among those with a booklet, twenty-three had faulty files that would not open. Additionally, one booklet opened but contained no content. As a result, we obtained a total of 47,372 filtered text entries from campaign booklets. During the data cleaning process, I encountered several special cases that required careful handling:

- Candidates with Identical Names: There were two instances (a total of four candidates) where two candidates with the same names ran as independents for the same office in the same district, all contesting basic assembly seats. Despite sharing identical names and partisan identifiers, these were distinct individuals. All of these candidates have been retained in the dataset.

Year	Conservative Party	Progressive Party
2022	People Power Party	Democratic Party of Korea
2021		
2020	Unified Future Party	
2019	Liberty Korea Party	
2018		
2017		
2016	Saenuri Party	
2015		
2014		New Politics Alliance for Democracy
2013		Democratic Party
2012		Democratic United Party
2011	Grand National Party	
2010		United Democratic Party
2009		
2008		
2007		Uri Party
2006		
2005		
2004		
2003		

Table 3. Names of the Two Major Parties by Year.

- Incorrect Booklet Assignments: Three candidates had booklets incorrectly assigned to them—these booklets belonged to other candidates. The erroneous booklet information for these candidates has been removed from the dataset to ensure accuracy. These three candidates are included in the total of 2,283 candidates with missing booklets.
- Shared Booklet Among Candidates: Three candidates from the same party, running for basic assembly seats in a multi-member district, shared a single booklet throughout their campaigns. Each of these candidates is included in the dataset, with the shared booklet associated with all three.
- Unprocessable Booklets: Twenty-three booklets contained errors that prevented processing, such as corrupted files or unreadable content. The booklet codes for these entries are retained in the dataset, but no text data is available.
- Empty Booklet: One booklet file opened successfully but contained no content. Its booklet code remains in the dataset, with an empty text column and a missing value in the filtered column.

South Korean party statements corpus. I collected official remarks released by South Korea's two major political parties—the center-left Progressive Party (www.theminjoo.kr) and the center-right Conservative Party (www.peoplepowerparty.kr)—from their respective official websites using a custom Python script. Minor parties exist in South Korea. But they are not included in the dataset because they have a small number of seats, making them less consequential in legislative processes, most of them end up merging with the two major parties included in the dataset, and their websites no longer exist after getting merged. Throughout this section, I refer to these parties by their political inclinations, the Conservative Party and the Progressive Party, rather than their names because both parties have undergone frequent name changes, divisions, and mergers over the years (see Table 3). Table 3 outlines the name changes of the two parties from 2003 to 2022. The Conservative Party has been known as the People Power Party since late 2020, while the Progressive Party has been known as the Democratic Party of Korea since 2015.

Both parties regularly post minutes from portions of their party leadership meetings and statements from their spokespersons on their official websites. Using a custom Python script, I scraped these web pages and collected all available postings from 2003 to 2022. I began with 2003 because September 29, 2003, is the earliest date for which postings exist for the Progressive Party. For the Conservative Party, postings are available starting from March 24, 2004. After collecting the data, I parsed the text using the Python library khaiii for morphological analysis.

Data Records

The datasets are publicly available on the Open Science Framework (OSF) repository under the title "South Korean Election Campaign Booklet Corpus and Party Statements Corpus"[16]. Users can access the data in CSV format:

- Election Campaign Booklet Corpus:
  - Files: sk_election_campaign_booklet_v2022.csv
- Party Statements Corpus:
  - Files: sk_party_statements_v2022.csv

Variable Name	Measurements
date	date of the election
name	name of the candidate
region	metro-level region where the district is located
district	name of the district
office_id	office identifier (see Table 5)
office	office where the candidate is running in English
giho	candidate identifier per NEC
party	name of the party
party_eng	name of the party in English
result	election result
result_code	not elected = 0, elected = 1
sex	candidate's sex
sex_code	female = 0, and male = 1
birthday	candidate's birthday
age	candidate's age
job_id	identifier for candidate's job per NEC
job	candidate's job
job_name	candidate's job category per NEC
job_name_eng	English translation of job_name
job_code	standardized identifier for job_name (see Table 6)
edu_id	identifier for candidate's education level per NEC
edu	candidate's education
edu_name	candidate's education category per NEC
edu_name_eng	English translation of edu_name
edu_code	standardized identifier for edu_name (see Table 7)
career1	candidate's career
career2	candidate's career
pages	number of pages in the booklet
code	booklet identifier per NEC
text	full text from the booklet
filtered_text	khaiii parsed and filtered text from the booklet

Table 4. Variable Description.

office_id	office
1	president
2	national_assembly
3	metro_head
4	basic_head
5	metro_assembly
6	basic_assembly
11	education_superintendent

Table 5. Mapping of office_id and office.

Additionally, a PDF file containing the detailed data descriptor and a PDF file with supplementary information is included to facilitate understanding and utilization of the datasets.

Repository Access:

- Repository Name: Open Science Framework (OSF)
- Project Title: South Korean Election Campaign Booklet Corpus and Party Statements Corpus
- DOI/URL: https://doi.org/10.17605/OSF.IO/RCT9Y

The *South Korean Election Campaign Booklet Corpus* dataset comprises 49,678 observations with thirty-one variables, providing comprehensive information on candidates' campaign messages and biographical details. This rich dataset enables detailed analysis of electoral strategies over time. Table 4 describes the variable names and their measurements. Notably, the filtered text refers to the parsed text after removing numerical values, words written in non-Korean letters (such as Roman alphabets or Chinese characters), and other symbols

job_code	job_name	job_name_eng
1	국회의원	national_assembly
2	정치인	politician
3	농축산업	agriculture
4	상업	commerce
5	광공업	mining
6	운수업	transportation
7	수산업	fishery
8	건설업	construction
9	언론인	journalism
10	금융업	finance
11	약사의사	medical
12	변호사	lawyer
13	종교인	religion
14	회사원	corporate_employee
15	교육자	education
16	정보통신업	ict*
17	출판업	publication
18	공무원	public
19	무직	unemployed
20	기타	other
21	지방의원	regional_assembly
22	교육감	education_superintendent
23	교육위원	education_committee
24	교육의원	education_assembly

Table 6. Mapping of job_code, job_name, and job_name_eng. *ict stands for Information and Communication Technology.

edu_code	edu_name	edu_name_eng
1	미기재	missing
2	독학	none
3	초퇴	elementary_withdrew
4	초졸	elementary_graduated
5	중재	middle_enrolled
6	중퇴	middle_withdrew
7	중졸	middle_graduated
8	고재	high_enrolled
9	고퇴	high_withdrew
10	고졸	high_graduated
11	전문대재	juniorcollege_enrolled
12	전문대퇴	juniorcollege_withdrew
13	전문대졸	juniorcollege_graduated
14	대재	undergrad_enrolled
15	대퇴	undergrad_withdrew
16	대학교수료	undergrad_completed
17	대졸	undergrad_graduated
18	대학원재	graduate_enrolled
19	대학원퇴	graduate_withdrew
20	대학원수료	graduate_completed
21	대학원졸	graduate_graduate

Table 7. Mapping of edu_code, edu_name, and edu_name_eng.

or special characters like punctuation marks. party_eng contains the English name of the party. If the official English name is unavailable, I used transliteration, representing the pronunciation of the Korean party name using the Latin alphabet (See Supplementary Information Appendix A for the mapping of date, party, and party_eng). sex_code is coded as female = 0, and male = 1. result_code is coded as elected = 1, not elected = 0.

Year	Office	Average Pages	Average Words from Full Text	Average Words from Filtered Text
2000	National Assembly	7.8	849.2	1616.7
2002	Basic Head	7.9	787.0	1480.9
	Metro Head	7.3	775.9	1444.7
	President	9.8	591.8	1148.7
2004	National Assembly	7.7	833.8	1602.2
2006	Basic Assembly	6.7	611.1	1147.8
	Basic Head	10.2	1038.3	1913.8
	Metro Assembly	6.8	657.5	1228.6
	Metro Head	9.0	922.7	1712.9
2007	President	11.1	1018.5	1982.9
2008	National Assembly	9.1	864.0	1620.4
2010	Basic Assembly	6.6	554.1	1036.6
	Basic Head	10.1	996.1	1790.4
	Education Superintendent	8.8	845.3	1529.0
	Metro Assembly	6.7	589.5	1091.3
	Metro Head	8.6	808.1	1473.7
2012	National Assembly	9.6	1091.6	1973.8
	President	5.9	528.3	1011.1
2014	Basic Assembly	6.4	618.9	1124.7
	Basic Head	9.7	1115.5	1946.3
	Education Superintendent	9.5	1051.3	1869.1
	Metro Assembly	6.4	698.3	1258.6
	Metro Head	8.3	962.4	1708.0
2016	National Assembly	9.2	1029.3	1826.8
2017	President	5.9	593.5	1124.3
2018	Basic Assembly	6.5	631.3	1137.3
	Basic Head	10.0	1179.9	2018.0
	Education Superintendent	10.0	1070.4	1897.3
	Metro Assembly	6.6	693.9	1236.3
	Metro Head	8.1	984.4	1734.1
2020	National Assembly	7.7	943.7	1619.7
2022	President	4.6	308.9	600.5
	Basic Assembly	6.3	563.2	1009.2
	Basic Head	9.9	1053.3	1776.5
	Education Superintendent	10.0	979.9	1684.4
	Metro Assembly	6.7	641.8	1133.8
	Metro Head	8.0	895.1	1525.2
Total		7.1	695.0	1267.0

Table 8. Average Number of Pages, Words from Full Text, and Words from Filtered Text by Year and Office.

Table 5 presents the mapping of office_id and office.

Table 6 presents the mapping of job_code, job_name, and job_name_eng. The NEC categorizes candidates' jobs into twenty-four categories and provides a job_name and a job_id. However, while job_name remains the same, job_id varies across election years. To ensure consistency, I standardized the identifiers across election years and created job_code. Additionally, I translated job_name into English, resulting in job_name_eng. The abbreviation ICT in job_code 16 represents Information and Communication Technology.

Table 7 presents the mapping of edu_code, edu_name, and edu_name_eng. The NEC categorizes candidates' education level into twenty-one categories and provides an edu_name and an edu_id. However, as with job_name and job_id, while edu_name remains the same, edu_id varies across election years. To ensure consistency, I standardized the identifiers across election years and created edu_code. Additionally, I translated edu_name into English, resulting in edu_name_eng.

Table 8 reports the average number of pages, the number of words in the full text, and the number of words in the filtered text, categorized by year and office.

A note of caution in analyzing the booklets published by candidates running for Basic Assembly is necessary. As shown in the table, Basic Assembly election booklets exhibit notably lower average total and filtered text word counts compared to other election categories, even after factoring in their typically fewer pages. Specifically, the average number of words per page for basic assembly is 170.6, whereas other categories feature higher averages: basic head (192.1), education superintendent (182.6), metro assembly (182.3), metro head (196.7), and National Assembly (203.1). The only exception with a lower average is president (159.7). The lower average words may be due to genuinely fewer

Year	Conservative Party	Progressive Party
2003		478
2004	1,632	2,324
2005	1,536	1,903
2006	1,267	2,063
2007	2,431	1,806
2008	1,683	2,595
2009	1,572	2,596
2010	1,420	3,380
2011	1,654	3,355
2012	2,700	5,159
2013	1,197	2,568
2014	1,457	2,395
2015	1,120	2,071
2016	1,008	1,894
2017	1,896	2,223
2018	1,761	1,858
2019	2,395	1,873
2020	2,154	1,795
2021	2,956	2,394
2022	3,276	3,356
Total	35,115	47,608

Table 9. Number of Entries in the South Korean Party Statements Dataset by Year for Each Party.

words in basic assembly pamphlets or suboptimal PDF quality resulting in more missing words in the OCR process than in other elections. As mentioned in the Technical Validations, the author excluded Basic Assembly booklets from our validation because of their visibly lower quality, particularly in earlier years. However, even the more recent Basic Assembly booklets (which appear to be in better condition) continue to show fewer word counts, suggesting that additional factors, such as the different electoral system, typically shorter careers of the candidates, or the constituency's size, may also play a role. Since the author cannot definitively conclude that poor PDF quality alone caused the lower word count, the author suggests researchers to exercise caution when analyzing Basic Assembly booklets.

The *South Korean Party Statements Corpus* dataset comprises a total of 35,115 entries from the Conservative Party and 42,335 entries from the Progressive Party. Table 9 presents the number of entries by party for each year. The dataset includes nine variables:

1. no: A unique identifier for each entry within each party.
2. year: The year the entry was posted.
3. ymd: The full date (year-month-day) the entry was posted.
4. title: The title of the entry.
5. text: The full text of the entry.
6. filtered: The text after morphological parsing, which includes only Korean text and excludes numbers, foreign language text, and grammatical components.
7. partisan: The political party (Progressive or Conservative).
8. conservative party indicator: A binary variable indicating whether the entry is from the Conservative Party (1) or not (0).
9. id: A concatenation of the party identifier and the entry number to create a unique ID for each entry.

Technical Validation

To ensure the technical quality and reliability of the Election Campaign Booklet Corpus, I conducted validation tests by comparing the text extracted using three different Optical Character Recognition (OCR) methods with manually transcribed text. This approach allowed us to assess the accuracy of the OCR methods and validate our data collection procedures.

I evaluated three OCR methods for extracting text from the PDF files of the election campaign booklets:

1. Full-Color, Whole PDFs using Google Drive API:
   Process: The original PDF files were used without any preprocessing. I uploaded each PDF to Google Drive using the Google Drive API and opened it with Google Docs, which automatically extracted the text, and saved the result to a text file.
   Rationale: Using the full-color PDFs without modification aimed to preserve the original quality and color information.
2. Black-and-White, Page-by-Page using Tesseract:
   Process: The PDF files were converted to black-and-white images and split into individual pages using

Method	Full Text		Korean Only		Filtered Text	
	Mean	Standard Error	Mean	Standard Error	Mean	Standard Error
Full-color, whole PDFs using Google Drive API	72.4%	1.35%	76.9%	1.35%	86.2%	1.13%
Black-and-white, page-by-page using Tesseract	29.0%	1.33%	33.0%	1.50%	41.8%	1.73%
Black-and-white, page-by-page using Google Drive API	77.8%	0.72%	86.1%	0.71%	94.6%	0.54%

Table 10. Comparison of Proportion of Extracted Text.

the pdf2image and Pillow Python libraries. I then applied the Tesseract OCR engine (via the pytesseract Python library) to each page individually. The extracted text from all pages was concatenated to create a single text file for each booklet.
Rationale: Converting to black-and-white can improve OCR performance by providing uniform gradients and enhancing text clarity and contrast, and processing page-by-page can isolate errors and prevent a single problematic from affecting the entire document. Tesseract OCR engine is one of the leading OCR engines by time of writing.

3. Black-and-White, Page-by-Page using Google Drive API:
Process: Similar to the second method, PDFs were converted to black-and-white and divided into individual pages. Similar to the first method, each page was then uploaded to Google Drive via the API and opened with Google Docs to extract the text. Extracted texts from all pages were concatenated for each booklet. The examples for each step can be found in Supplementary Information Appendix B.
Rationale: Combining the benefits of black-and-white conversion and Google Docs' OCR capabilities, this method aimed to maximize extraction accuracy.

I randomly selected 197 entries from the dataset for validation. This sample size was determined by selecting approximately 1% of a subset of 16,948 entries where the full-color, whole PDFs using Google Drive API method had successfully extracted text. This subset excluded candidates running in basic assembly elections to focus on higher-level offices which generally had better quality booklets. To ensure the sample was representative, I stratified entries based on the following criteria: election date, office type, metropolitan region, and party affiliation. For party affiliation, candidates were categorized into five groups: Conservative Party, Progressive Party, Leftist Party, United Liberal Democrats, and others (parties appearing in fewer than three elections). This stratification ensured a balanced representation across different political contexts and election types.

I employed human coders to manually transcribe the text from the selected 197 campaign booklets. The coders were provided with detailed instructions to ensure consistency. Coders were instructed to type all readable Korean text. For booklets containing newspaper article clippings with overlapping or partially obscured text, coders were to transcribe only the readable headlines and subheadings, not the detailed article text. For photographs featuring banners, signs, or maps, coders were to transcribe any legible Korean text without attempting to infer or supply unreadable portions. Coders were instructed not to transcribe non-Korean text, illegible content, or decorative elements without textual significance. After transcription, the manually typed texts were processed using the khaiii Python library for morphological analysis, parsing, and filtering.

To evaluate the performance of each OCR method, I calculated the proportion of words from the manually typed text that were correctly extracted by each method. The comparison was conducted at three levels:

1. Full Text Comparison:
Method: Comparing the raw text output from each OCR method with the raw manually typed text.
Purpose: Assess overall word recognition accuracy without any preprocessing.
2. Korean Text Only:
Method: Both OCR outputs and manually typed texts were processed to remove all non-Korean characters, including numbers, Roman alphabets, Chinese characters, punctuation marks, and special symbols.
Purpose: Focus on the accurate recognition of Korean words, which are the primary interest in the dataset.
3. Parsed and Filtered Text:
Method: Both texts were processed using the khaiii library, performing morphological analysis to tokenize and normalize the text, and removed all non-Korean characters.
Purpose: Evaluate the OCR accuracy at the level most relevant for linguistic and textual analyses.

Table 10 presents the results of this comparative analysis. Using the Full-Color, Whole PDFs with Google Drive API method, the OCR captured an average of 72.4% of the words from the manually typed text in the full-text comparison. The proportion of detected words increased as I limited the transcribed text to Korean only and reached 86.2% when I compared the filtered texts. For the Black-and-White, Page-by-Page using Tesseract method, the OCR extracted less than a third of the words from the manually typed text in the full-text comparison. Although there was marginal improvement when I restricted the transcribed text, this method had the lowest accuracy among the three methods. The Black-and-White, Page-by-Page using Google Drive API method achieved the highest extraction rate, capturing 77.8% of the words from the manually typed text in the full-text comparison. There was further improvement when considering Korean text only by filtering out non-Korean characters, correctly detecting 86.1 percent of the text. Most notably, in the parsed and filtered text, this method reached an impressive 94.6% accuracy, effectively capturing nearly all the relevant Korean text needed for analysis.

Measure	Korean Only		Filtered Text	
	Mean	Standard Error	Mean	Standard Error
Precision	0.8506	0.0080	0.9064	0.0061
Recall	0.8754	0.0065	0.9433	0.0051
F1 Score	0.8590	0.0070	0.9218	0.0052

Table 11. Precision, Recall, and F1 Scores of Extracted Text.

These findings indicate that the black-and-white, page-by-page method using the Google Drive API is the most effective OCR approach for our dataset. Consequently, I have included the results obtained from this method in the final dataset. The average number of words in the manually transcribed text is 788.6, whereas the average number of words extracted using the Black-and-White, Page-by-Page using Google Drive API method is 807.7. The higher word count in the OCR-extracted text likely results from its ability to capture additional Korean text found in newspaper article clippings, banners within photographs, and text on maps.

Table 11 presents additional statistics evaluating the accuracy of text extracted using the Black-and-White, Page-by-Page method with the Google Drive API, focusing on both the Korean-only text and the parsed and filtered text. In the context of F1 scores, values between 0.9 and 1 are generally considered to indicate excellent performance, while scores between 0.8 and 0.9 signify good performance[17,18]. The results show an average F1 score of 0.9218 for the filtered text with high scores on both precision and recall, providing further evidence of the high performance and success of this text extraction method.

The technical validation for Party Statements corpus was done as part of my assessment of the performance of the khaiii Python library. I conducted a random sampling of parsed and filtered text from both the Party Statements Corpus and the Election Campaign Booklet Corpus. While not all text was parsed and tagged perfectly to my expectations, I was able to confirm that khaiii performed impressively overall, accurately parsing and tagging the majority of the text and that the parsed data in the corpus is reliable for subsequent linguistic and textual analyses.

A note about the potential use of Large Language Models (LLMs) to improve text extraction—and my decision not to use them—would be helpful. In an effort to improve text extraction accuracy, I considered employing LLMs to detect and correct misformed or fragmented words in the extracted text. Due to the way the booklets were formatted, some words that spanned across pages were not properly detected during the OCR process. Additionally, the similarity of certain Korean characters led to misrecognition by the OCR software. I experimented with using ChatGPT-4 to restore broken or missing text. However, I ultimately decided against incorporating the LLM-corrected text into the final dataset. The primary reason was that ChatGPT-4, utilizing its generative algorithms, tended to produce text that was plausible or likely to be in the booklet but was not actually present in the original material. This generative aspect resulted in the creation of content that could misrepresent the candidates by introducing information not found in their official campaign materials. The potential for generating non-authentic text differs fundamentally from the OCR errors encountered with the Google Drive API or parsing issues with the khaiii library. While OCR and parsing errors may omit or misread existing text, they do not introduce new content. In contrast, using an LLM like ChatGPT-4 could amplify and consolidate errors by adding fabricated text, thereby compromising the integrity of the dataset. To maintain the accuracy and reliability of the dataset, I chose not to include the LLM-generated corrections.

Code availability

The code used to create the dataset presented in this descriptor are available on the OSF repository[16]. The code is developed using Python.

The workflow for generating the Election Campaign Booklet corpus involved the following scripts. I followed the order in which the files are listed in running the files. Note that the biographical information, booklets, and their parsed text are organized in separate CSV files by year and election types.

1. booklet_1_get_booklet_jpg_to_pdf.ipynb: Downloaded booklets provided in JPEG format and consolidated them into PDF files.
2. booklet_2_get_booklet_pdf.ipynb: Downloaded booklets provided directly in PDF format.
3. booklet_3_extract_text.ipynb: Extracted text from PDF files via the Google Drive API and saved the output in.txt format.
4. booklet_4_collect_text.ipynb: Compiled text files to create a CSV file.
5. booklet_5_count_pages.ipynb: Counted the number of pages in each PDF file.
6. booklet_6_create_booklet_meta.ipynb: Retrieved meta-information for election campaign booklets, including booklet IDs and URLs to the booklet files.
7. booklet_7_create_candidate_bio.ipynb: Collected candidates' biographical information, such as sex and birthdate.
8. booklet_8_create_election_campaign_booklet.ipynb: Consolidate all text files and parse their content. Merge biographical information, meta-information, and page numbers with the extracted text. Organize and structure the data to generate the final datasets.

The final dataset produced is:
- sk_election_campaign_booklet_v2022.csv

For the Party Statements corpus, the following scripts were used:
1. get_parties_text.ipynb: Collected entries from the two parties' websites. Note that both parties have restructured their web pages since the last data collection; the code may require revisions to function correctly as of October 2023.
2. extract_text.ipynb: Parsed the extracted text from the websites.
3. create_parties.ipynb: Created the final datasets from the parsed text.

The final dataset produced is:
- sk_party_statements_v2022.csv

Received: 11 October 2024; Accepted: 16 May 2025;
Published online: 19 June 2025

References

1. Bast, J., Oschatz, C. & Renner, A.-M. Successfully overcoming the "Double Bind"? A mixed-method analysis of the self-presentation of female right-wing populists on Instagram and the impact on voter attitudes. *Political Communication* **39**(3), 358–382, https://doi.org/10.1080/10584609.2021.2007190 (2022).
2. Gervais, B. Congressional Tweet Archive. *Harvard Dataverse.* https://doi.org/10.7910/DVN/BOK1CF (2019).
3. Gerard, P. Truth Social Dataset. *arXiv*, 20 Mar. https://doi.org/10.48550/arXiv.2303.11240 (2023).
4. Panagopoulos, C. Boy talk/girl talk: Gender differences in campaign communications strategies. *Women & Politics* **26**(3-4), 131–155, https://doi.org/10.1300/J014v26n03_06 (2004).
5. Licenji, L. & Hoxha, J. Contrasting strategies and messages: an in-depth comparative study of Albania's national and municipal election advertisements. *Humanit Soc Sci Commun* **11**, 1558, https://doi.org/10.1057/s41599-024-04118-7 (2024).
6. Wesleyan University. Wesleyan Media Project. *United States, Web Archive*. https://www.loc.gov/item/lcwaN0018466/ (2012).
7. Yoshikawa E., & Roesner, F. Exploring Political Ads on News and Media Websites During the 2024 U.S. Elections. *arXiv preprint*. https://doi.org/10.48550/arXiv.2503.02886 (2025).
8. Chalkiadakis, I., Anglès d'Auriac, L., Peters, G. & Frau-Meigs, D. A text dataset of campaign speeches of the main tickets in the 2020 US presidential election [Data set]. *Zenodo.* https://doi.org/10.5281/zenodo.14785782 (2025).
9. Dolan, K. *When does gender matter?: Women candidates and gender stereotypes in American elections*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199968275.001.0001 (2014).
10. Erfort, C., Stoetzer, L. F., & Klüver, H. The PARTYPRESS Database: A new comparative database of parties' press releases. *Research & Politics*, 10(3). https://doi.org/10.1177/20531680231183512 (Original work published 2023) (2023).
11. Lehmann, P. *et al.* The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2024a. *Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB)/Göttingen: Institut für Demokratieforschung (IfDem)*. https://doi.org/10.25522/manifesto.mpds.2024a (2024).
12. Mügge, L. & Runderkamp, Z. Political Narratives in Representation: Maiden Speeches of Ethnic Minority Members of Parliament. *European Journal of Political Research* **62**(2), 1–24, https://doi.org/10.1111/1475-6765.12632 (2023).
13. Schumacher, G., Hansen, D., van der Velden, M. A. C. G., & Kunst, S. A new dataset of Dutch and Danish party congress speeches. Research & Politics, 6(2). https://doi.org/10.1177/2053168019838352 (Original work published 2019) (2019).
14. Lim, T-H. Party Competition in Swing Districts and Childcare Policy Development in South Korea. *Dissertations - ALL*. **1718**. https://surface.syr.edu/etd/1718 (2023).
15. Park, J-H, & Lee, K. Analysis of Offensive Language Use by South Korea's Two Major Political Parties: Using Party Press Releases from 2007 to 2023. *Korean Political Science Review* **58**:2, 73–104. (in Korean) (2024).
16. Lim, T-H. South Korean Election Campaign Booklet and Party Statements Corpora. *OSF*. https://doi.org/10.17605/OSF.IO/RCT9Y (2025, March 20).
17. Encord. F1 score in machine learning. *Encord*. (n.d.). Retrieved March 10, from https://encord.com/blog/f1-score-in-machine-learning/ (2025).
18. Jurafsky, D. & Martin. J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released August 20, https://web.stanford.edu/~jurafsky/slp3 (2024).

Acknowledgements

The author expresses gratitude to Dr. Margarita Estevez-Abe, Dr. Seth Jolly, Dr. Christopher Faricy, Dr. Simon Weschle, Dr. Yotam Schmargad, Dr. Jong Hee Park, and the participants of the 2024 MPSA New Data Source panel for their valuable feedback and encouragement in the construction of the dataset. Additionally, the author sincerely appreciates Ms. Yu Jeong Hwang, Dr. Hyojeong Hwang, and Mr. Hanbit Hwang for their contributions to the technical validation of the dataset.

Author contributions

Lim is the sole author of this work. He was responsible for all aspects of the project, including the collection and processing of data, as well as drafting the data descriptor.

Competing interests

The author declares that there are no competing interests associated with this work. All research was conducted in the absence of any financial or personal relationships that could be perceived as influencing the results or interpretation of the data.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05220-4.

Correspondence and requests for materials should be addressed to T.H.L.