



KNoTE: A TEI Encoded XML Dataset of Modern Korean Novels

DATA PAPER

]u[ubiquity press

GAYEON KIM**

SEONYEONG PARK**

HAEIN JI

HAGYEONG LEE

BYEONGJOO LEE

CHAEYEON JEONG

BYUNGJUN KIM

*Author affiliations can be found in the back matter of this article

**Gayeon Kim and Seonyeong Park made equal contributions.

ABSTRACT

KNoTE (Korean Novel TEI Encoded) is an open XML dataset of 33 modern Korean novels (1906–1954) by 22 authors, encoded according to the TEI P5 Guidelines. Source texts were collected from Korean Wikisource and semi-automatically encoded using a large language model via prompt engineering, followed by multi-layered manual validation by seven annotators. The markup captures named entities, character ontologies, direct speech, Hanja annotations, and period-specific metadata including political ideologies, colonial-era place names, and social class markers. As the first TEI-conformant corpus of modern Korean literature, the dataset supports computational literary analysis, NLP model fine-tuning, and cross-corpus interoperability.

CORRESPONDING AUTHOR:

Byungjun Kim

Cultural Informatics, The
Academy of Korean Studies,
Seongnam, Republic of Korea
bjkim@byungjunkim.com

KEYWORDS:

XML; TEI; modern Korean
novels; large language models;
semi-automatic encoding

TO CITE THIS ARTICLE:

Kim, G., Park, S., Ji, H., Lee, H., Lee, B., Jeong, C., & Kim, B. (2026). KNoTE: A TEI Encoded XML Dataset of Modern Korean Novels. *Journal of Open Humanities Data*, 12: 102. DOI: <https://doi.org/10.5334/johd.590>

1 OVERVIEW

REPOSITORY LOCATION

The KNoTE (Korean Novel TEI Encoded) dataset is publicly accessible through the Zenodo repository: <https://doi.org/10.5281/zenodo.18679321>.

CONTEXT

The dataset was produced as part of a structured training programme in humanities data design at the Academy of Korean Studies (AKS). While TEI-based corpora have been developed for various European literary traditions, notably eighteenth-century French novels (Röttgermann, 2024) and early Holocaust testimony texts and camp literature (Entrup et al., 2013), and while recent data papers have begun to address non-Western modernist literatures in Asian languages (Wong, 2026), modern Korean fiction of 1906–1954 has so far lacked a structured TEI representation. Written across the transitional Hangul–Hanja mixed script and shaped by the colonial-modern experience, these texts present distinctive encoding challenges that existing frameworks do not address. Meanwhile, natural language processing research on Korean novels has advanced through character-centric annotated datasets (Kim et al., 2024), yet no TEI-encoded corpus has been made available for computational analysis of the broader literary structure. KNoTE addresses this gap by providing the first open, TEI-conformant XML dataset of modern Korean novels.

2 METHOD

The construction of the KNoTE dataset followed a structured workflow divided into data acquisition, semi-automatic encoding, and validation.

STEPS

The dataset production involved two main procedures. First, in the data collection phase, source texts were collected from Korean Wikisource using a hybrid approach¹ combining the Wikisource XML dump file with the Wikisource and Wikidata APIs. A custom Python-based scraping and parsing script was developed to take a pre-compiled CSV of target works as input and retrieve each work’s wikitext and metadata. For works split across multiple chapters, the script parsed the full dump file, collected and sorted all related chapters, and reassembled them into a single complete text. All wiki markup was automatically stripped, with outputs generated in XML format. Second, in the semi-automatic encoding phase, initial TEI markup was generated using a web application (“Conversion of Korean Modern Literature xml/tei”²) built on Google AI Studio (Gemini 2.5 Flash, Figure 1). This methodology builds on the capacity of large language models (LLMs) to handle TEI annotation tasks; as recent evaluations suggest, these models can now automate labour-intensive markup steps such as structural tagging and entity disambiguation (Strutz, 2026).

Utilising prompt engineering tailored to modern Korean fiction, the application performed automated construction of the TEI header, generation of a character ontology (<listPerson>), named entity tagging (<persName>, <placeName>, <orgName>), distinction between direct speech and interior monologue (<said>), Hanja annotation (<foreign xml:lang="zh">), and Korean postpositional particle separation in entity tagging. Encoding was based on the TEI ALL schema of the TEI P5 Guidelines. Tailored prompt engineering is necessary because, without strict guidance, LLMs frequently produce XML validation errors or bypass external reference data during complex tasks (De Cristofaro & Zilio, 2025). Structuring the prompts in this way reduced such errors and improved markup accuracy.

Table 2 lists representative custom values from this process, including political ideologies, social classes, modern Korean spaces, and a project-specific virtual speaker `edcmt` for narratorial commentary. Many named entities also carry an external authority link: Korean places, organisations, events, and concepts to the Encyclopedia of Korean Culture (encykorea.aks.ac.kr),

¹ <https://github.com/AKS-DHILAB/KNoTE>.

² <https://ai.studio/apps/drive/1gca8QHQLFCb79saG23wcFyidHFKjgUAb>.

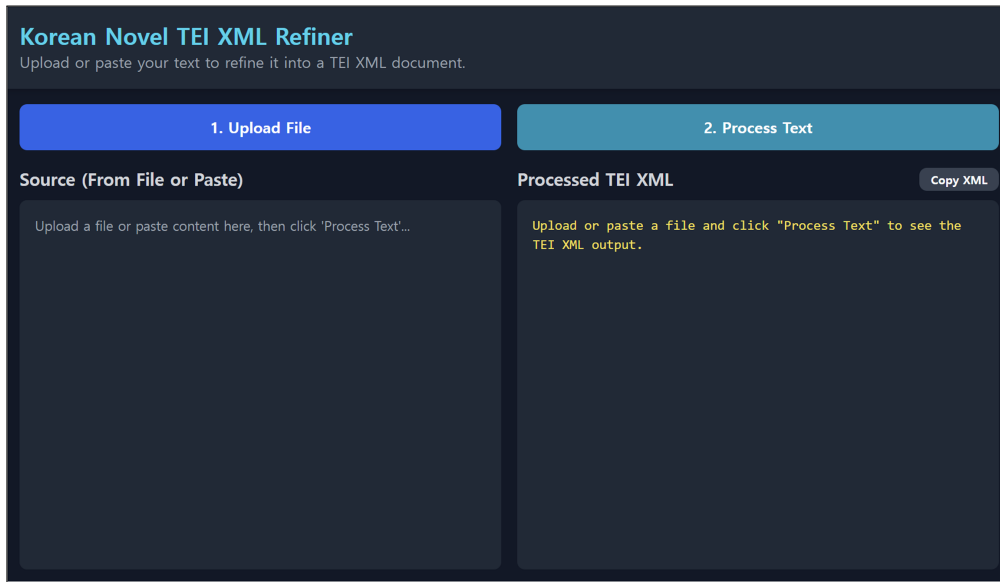


Figure 1 Conversion of Korean Modern Literature xml/tei App Screen.

CATEGORY	ELEMENT	KEY ATTRIBUTES	FUNCTION
TEI Header	titleStmt	—	Title, author, encoder credits
	respStmt	—	Encoder/reviewer attribution
	sourceDesc/bibl	type	Source provenance
	encodingDesc	—	Encoding standard (TEI ALL)
	revisionDesc/change	when, who	Encoding/review dates
Characters	listPerson	—	Character registry
	person	xml:id	Individual character
	personGrp	xml:id	Collective character group
	persName	xml:lang	Name in Korean or Hanja
Entities	persName	ref	Named character mention
	rs	ref, type	Pronoun/epithet reference
	placeName	—	Place name
	orgName	—	Organisation name
	date	when, type	Date reference
	time	when, type	Time-of-day reference
Speech	said	who, aloud, direct, mode	Direct speech, thought, or monologue
Language	foreign	xml:lang="zh", xml:lang="ja"	Hanja preservation and Japanese in Hangul
Structure	div	type, n	Chapter/section division
	p	—	Paragraph
	head	—	Chapter title

Table 1 TEI elements and attributes used in the KNoTE dataset.

and persons to Wikipedia (ko.wikipedia.org) or Wiktionary (ko.wiktionary.org). A complete inventory of all TEI elements and attributes in the corpus is provided in the repository (KNoTE_TEI_tags_full.md) and the Zenodo record.

SAMPLING STRATEGY

The KNoTE dataset aims to present a reproducible TEI encoding workflow and provide a seed dataset for evaluating automated TEI tagging systems. Texts were therefore deliberately sourced from Korean Wikisource—a freely accessible, public-domain repository—to ensure that the dataset can be independently verified and extended without access restrictions. The dataset comprises 33 works of modern Korean fiction by 22 authors, spanning from Yi In-jik's *Tears of Blood* (Hyeol-ui Nu, 1906) to Kang So-cheon's *The Photo Studio that Takes Pictures*

Table 2 Period-revealing markup values in the KNoTE dataset; schema-level inventory in Table 1.

CATEGORY	ELEMENT	KEY ATTRIBUTES	SAMPLE VALUES
Political ideologies	term	type="ideology"	socialism, militarism, humanitarianism, feminism
	orgName	ref="#COMMUNIST_PARTY"	Communist Party
	rs	type="person_group", subtype="communist"	communists, fighters, party members
	rs	type="anarchism"	anarchists, seditious thought
Modernization concepts	term	type="concept"	New Woman, bourgeois, feudal aristocracy, cultural-style housing, Four-Pillars fortune
Social classes	rs	type="class"	yangban (gentry), sangnom (commoners)
	rs	type="laborer", "farmer"	farmers, labourers, first-rank workmen
	rs	type="intelligentsia"	"high-collar" (haikara, modern dandy)
	rs	type="prisoner", "official"	detainees, detectives, guards
	rs	type="immigrant"	migrant parties bound for Manchuria
Modern Korean Spaces	placeName	type="city"	Gyeongseong (colonial Seoul), Tokyo, Hsinking, Sinuiju
	placeName	type="region", "border"	Manchuria, Gando, the colonial border
	placeName	type="facility"	detention cells, Special Higher Police interrogation rooms
	placeName	type="hotel", "cafe", "bar"	Chosen Hotel, tabang (tearoom), Bar Lily
	orgName	type="police"	Special Higher Police, Gyeongbangdan (Defence Corps)
Currency & measures	measure	type="currency"	1,500 nyang (Joseon); 1 won, 30 jeon, 50 jeon (colonial)
	measure	type="distance", "age"	15 ri, 14 years old
Formal architecture	div	type="frame", "embedded"	Dr. K's Research, The Man Met in the Detention Center
	div	type="preface", "afterword"	Blood of Flowers (Hwa-ui Hyeol)
	div, q	type="letter"	Nostalgia (Hyangsu), The Sorrows of the Weak, Ready-made Life, and four other works
Project extension	said	who="#edcmt"	"Who could have foreseen..." (The Girl of Mystery (Uisim-ui Sonyeo)); "But how could she escape those deep waves!" (Adada the Idiot (Baekchi Adada))

of Dreams (1954), covering the formative decades of modern Korean literature. Following the corpus-balancing framework of Röttgermann (2024), works were selected by four explicit criteria rather than canonical reputation alone: (i) *literary-historical canonicity*, operationalised through inclusion in standard literary histories and reference works (e.g., the *Encyclopedia of Korean Culture*) and the availability of Wikidata/Wikisource records; (ii) *representative or "first-of-its-kind" status* (e.g., the first *sinsoseol*, the first original Korean science fiction, and the debut work of the first modern woman novelist); (iii) *balance* across publication period, narrative form, and literary movement; and (iv) a *verifiable public-domain source text* suitable for annotation. Table 3 lists each work with its publication year, narrative form, and selection rationale; the resulting corpus ranges from *sinsoseol* (新小説) and short fiction to early modern science fiction. Women authors remain a minority (6 of 33 works), reflecting the historical scarcity of published women writers; we accordingly included all

NO.	AUTHOR	TITLE (ROM.)	YEAR	FORM	SELECTION RATIONALE
1	Yi In-jik	Tears of Blood (<i>Hyeol-ui Nu</i>)	1906	Sinsoseol	Korea's first <i>sinsoseol</i> (new novel).
2	Yi Hae-jo	The Iron World (<i>Cheol-segye</i>)	1908	Sinsoseol (transl. SF)	First Korean adapted novel and first SF-genre work (after Jules Verne).
3	Yi Kwang-su	The Heartless (<i>Mujeong</i> – short story)	1910	Short fiction	Short-story version preceding the first modern novel, <i>Mujeong</i> .
4	Yi Hae-jo	Blood of Flowers (<i>Hwa-ui Hyeol</i>)	1911.04	Sinsoseol	<i>Sinsoseol</i> written after the 1910 annexation; a “new old-style” novel.
5	Kim Myeong-sun	The Girl of Mystery (<i>Uisim-ui Sonyeo</i>)	1917.11	Short fiction	Debut work of Korea's first modern woman novelist.
6	Na Hye-seok	Kyung-hee	1918.03	Short fiction	Representative work on the colonial New Woman intellectual before marriage.
7	Na Hye-seok	To the Revived Granddaughter	1918.09	Short fiction	Bears witness to the 1910s from a woman's perspective (Kwon Bodurae).
8	Kim Dong-in	The Sorrows of the Weak	1919.02–03	Short fiction	Kim Dong-in's first published fiction.
9	Yi Ik-sang	The Straggler (<i>Nagoja</i>)	1919.07.14	Short fiction	Debut work of Yi Ik-sang, a leading writer of proletarian and neo-tendency (<i>sin-gyeonghyangpa</i>) literature.
10	Hyun Jin-geon	A Poor Wife (<i>Bincheo</i>)	1921.01	Short fiction	Early work of Hyun's self-reflective realism.
11	Na Hye-seok	Gyu-won	1921.07	Short fiction	Post-marriage work on “the grievances of the inner chamber.”
12	Hyun Jin-geon	A Society That Drives You to Drink	1921.11	Short fiction	Early work of Hyun's self-reflective realism.
13	Choi Seo-hae	Nostalgia (<i>Hyangsu</i>)	1924.04	Short fiction	A neo-tendency (<i>sin-gyeonghyangpa</i>) story depicting the lives of lower-class people who emigrated beyond Korea—to Siberia and northern Manchuria.
14	Hyun Jin-geon	A Lucky Day (<i>Unsu Joeun Nal</i>)	1924.06	Short fiction	Representative work of Hyun's naturalist gaze on the common people.
15	Kim Dong-in	Potato (<i>Gamja</i>)	1925.01	Short fiction	One of Kim Dong-in's 1920s naturalist stories.
16	Hyun Jin-geon	Director B and the Love Letters	1925.02	Short fiction	Satire on the repression of women's education under colonial rule.
17	Na Do-hyang	The Watermill (<i>Mullebang-a</i>)	1925.09	Short fiction	A later work marking Na Do-hyang's shift from his early romanticism to realism.
18	Bang Jeong-hwan	For Our Friends	1927.02	Children's fiction	By a leading children's-literature writer of the youth-liberation movement.

(Contd.)

Table 3 Selection rationale for the 33 works in the KNoTE dataset.

NO.	AUTHOR	TITLE (ROM.)	YEAR	FORM	SELECTION RATIONALE
19	Bang Jeong-hwan	The Eternal Shirt (<i>Mannyeon Shirt</i>)	1927.03	Children's fiction	Representative children's short fiction by the same author.
20	Bang Jeong-hwan	The Gold Watch	1929.01–02	Children's fiction	Representative children's short fiction by the same author.
21	Kim Dong-in	Dr. K's Research	1929.12	Short fiction (SF)	Regarded as Korea's first original (creative) SF story.
22	Kim Nam-cheon	Water (<i>Mul</i>)	1933.06	Short fiction	Short story by a leading writer of KAPF (proletarian federation).
23	Chae Man-sik	Ready-made Life	1934.05–07	Short fiction	Chae Man-sik's early self-satirical story.
24	Kang Kyeong-ae	Salt (<i>Sogeum</i>)	1934.05–10	Novella	Destitute life of a colonial subaltern woman, set in Manchuria.
25	Gye Yong-mook	Adada the Idiot (<i>Baekchi Adada</i>)	1935	Short fiction	Pure-art-period work; also read in disability-discourse studies.
26	Kim Yu-jeong	The Camellias (<i>Dongbaek-kkot</i>)	1936.05	Short fiction	Representative native, rural-set story by Kim Yu-jeong.
27	Yi Sang	The Wings (<i>Nalgae</i>)	1936.09	Short fiction	Representative work of Yi Sang, leading 1930s modernist (Guinhoe).
28	Yi Hyo-seok	When Buckwheat Flowers Bloom	1936.10	Short fiction	Representative lyrical work by a Guinhoe member.
29	Chae Man-sik	Uncle Chi-suk	1938	Short fiction	Intellectual-satire story from the year of <i>Tangnyu</i> and <i>Taepyeong-cheonha</i> .
30	Jeong In-taek	Melancholy (<i>Uuljeung</i>)	1940.09	Short fiction	A modernist story depicting an intellectual's inner life in the early 1940s, by Jeong In-taek (later a novelist in the DPRK).
31	Kim Sa-ryang	The Man Met in the Detention Center	1941	Short fiction	By a Korean-Japanese writer grappling with decolonization in the late colonial era.
32	Ji Ha-ryeon	The Journey (<i>Dojeong</i>)	1946.07	Short fiction	By a leftist woman novelist who later defected to the North.
33	Kang So-cheon	The Photo Studio that Takes Pictures of Dreams	1954.03	Children's fiction	Representative post-Liberation Korean children's story.

major available women writers of the period (Kim Myeong-sun, Na Hye-seok, Kang Kyeong-ae, and Ji Ha-ryeon).

QUALITY CONTROL

To ensure encoding consistency and accuracy, multiple layers of manual validation were implemented. Following the initial artificial intelligence (AI)-assisted markup, seven annotators reviewed, corrected, and supplemented the tags based on internally developed encoding guidelines. Subsequently, the curator and data manager conducted a full-coverage cross-review of all files, checking `<persName>`, `<ref>` consistency, `<said>` attribute integrity, and `<foreign xml:lang>` accuracy. Finally, the project supervisor performed a final scholarly review. Each file's `<revisionDesc>` records the dates of encoding completion and verification. This encoding and quality control process was carried out as part of a structured training programme in humanities data design at the Academy of Korean Studies (AKS). A preliminary

3 DATASET DESCRIPTION

REPOSITORY NAME

Zenodo

OBJECT NAME

KNoTE dataset (on Zenodo: <https://doi.org/10.5281/zenodo.18679321>)

FORMAT NAMES AND VERSIONS

XML, conformant to TEI P5 (TEI ALL schema)

CREATION DATES

2025-09-01–2026-04-27

DATASET CREATORS

Kim, Gayeon (Data manager, Academy of Korean Studies); Park, Seonyeong (Data curator, Academy of Korean Studies); Ji, Haein (Data collector, Academy of Korean Studies); Lee, Hagyeong (Annotator, Sogang University); Lee, Byeongjoo (Annotator, Republic of Korea Air Force Academy); Jeong, Chaeyeon (Annotator, Korea University); Lee, Jae-yeol (Annotator, Academy of Korean Studies); Jo, Gyungmin (Annotator, Academy of Korean Studies); Lim, Iro (Annotator, Academy of Korean Studies); Ismayilov, Orkhan (Annotator, Academy of Korean Studies); Kim, Byungjun (Supervisor, Academy of Korean Studies).

LANGUAGE

Korean, with Hanja (Chinese character) annotations marked via `xml:lang="zh"`.

LICENSE

Creative Commons Attribution 4.0 International (CC-BY 4.0).

PUBLICATION DATE

2026-04-27

4 REUSE POTENTIAL

The KNoTE dataset encodes characters, locations, and temporal markers in TEI XML, complementing recent character-centric dataset work on Korean novels (Kim et al., 2024). This standardised markup makes the texts amenable to quantitative literary analysis. Crucially, the encoding distinguishes spoken dialogue from interior monologue and reported thought, so that the representation of interiority—a defining concern in accounts of the modern Korean novel—becomes a separable object of analysis rather than collapsing into undifferentiated “speech.” The corpus can also serve as a fine-tuning resource for language models trained on historical Korean text, where structured TEI input has been shown to improve encoding accuracy (Strutz, 2026). Because each entity carries a unique identifier linked to external authorities (Encyclopedia of Korean Culture, Wikipedia), the dataset can be integrated into broader knowledge graphs. KNoTE is also a reference implementation for projects seeking to add period-specific metadata layers to literary corpora.

However, because TEI permits considerable schema variation, cross-corpus comparison will require tag mapping between projects that use different encoding conventions. In addition, the reliance on public-domain texts limits the temporal and authorial scope. To offset this constraint, the selection covers a range of genres, from *sinsoseol* to early science fiction. The dataset should

therefore be understood as a representative sample of the early modern tradition rather than an exhaustive archive of modern Korean fiction. A project-specific TEI customisation (ODD) for stricter validation is planned for a future release.

AI DECLARATION

Generative AI tools were used at two stages of this work. During the data production phase, Google AI Studio (Gemini 2.5 Flash) was employed for semi-automatic TEI encoding of the source texts, as described in the Method section. During the manuscript preparation phase, Anthropic Claude Opus 4.6 was partially used to assist with drafting and editing portions of the text. All AI-generated markup and manuscript content were reviewed, verified, and finalised by the human authors, who bear full responsibility for the accuracy and integrity of the published work.

ACKNOWLEDGEMENTS

We would like to express our deepest gratitude to Jae-yeol Lee, Gyungmin Jo, Iro Lim, and Orkhan Ismayilov for their contributions to the initial version of the KNoTE dataset, which laid the foundation for the final version. We are also grateful for the constructive comments provided by the editors and the reviewers, which greatly improved the quality of this work.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

The authorial roles and responsibilities for this research are specified as follows.

Gayeon Kim: Software, Methodology, Writing – original draft.

Seonyeong Park: Data curation, Writing – original draft.

Haein Ji: Data collection, Methodology, Writing – original draft.

Hagyeong Lee: Investigation, Validation.

Byeongjoo Lee: Investigation, Validation.

Chaeyeon Jeong: Investigation, Validation.

Byungjun Kim: Supervision, Project administration, Writing – review & editing.

Gayeon Kim and Seonyeong Park made equal contributions.

AUTHOR AFFILIATIONS

Gayeon Kim  orcid.org/0009-0005-1231-6939

Cultural Informatics, The Academy of Korean Studies, Seongnam, Republic of Korea

Seonyeong Park  orcid.org/0009-0001-1340-0455

Cultural Informatics, The Academy of Korean Studies, Seongnam, Republic of Korea

Haein Ji  orcid.org/0000-0001-6864-6789

Cultural Informatics, The Academy of Korean Studies, Seongnam, Republic of Korea

Hagyeong Lee  orcid.org/0009-0007-5309-883X

Department of Korean Language and Literature, Sogang University, Seoul, Republic of Korea

Byeongjoo Lee  orcid.org/0009-0001-0366-9944

Department of Language and Literature, Republic of Korea Air Force Academy, Cheongju, Republic of Korea

Chaeyeon Jeong  orcid.org/0009-0000-7367-9999

Department of Korean Language and Literature, Korea University, Seoul, Republic of Korea

Byungjun Kim  orcid.org/0000-0001-9925-9343

Cultural Informatics, The Academy of Korean Studies, Seongnam, Republic of Korea

REFERENCES

- De Cristofaro, M., & Zilio, D. (2025). Automating XML-TEI encoding of unpublished correspondence: A comparative analysis of two LLM approaches. *Umanistica Digitale*, 531–537.

- Entrup, B., Bärenfänger, M., Binder, F., & Lobin, H. (2013). Introducing GeoBib: An annotated and geo-referenced online bibliography of early German and Polish Holocaust and camp literature (1933-1949). In *Digital Humanities 2013, DH 2013, Conference Abstracts, University of Nebraska-Lincoln, Lincoln, NE, USA, July 16–19, 2013* (pp. 489–490). Alliance of Digital Humanities Organizations (ADHO). <https://dh2013.unl.edu/abstracts/ab-229.html>
- Kim, K., Lee, S., & Lee, S. (2024). KoCoNovel: Annotated dataset of character coreference in Korean novels. arXiv preprint, arXiv:2404.01140. <https://doi.org/10.48550/arXiv.2404.01140>
- Röttgermann, J. (2024). The Collection of Eighteenth-Century French Novels 1751–1800. *Journal of Open Humanities Data*, 10(1), 31. <https://doi.org/10.5334/johd.201>
- Strutz, S. (2026). A Multi-Dimensional evaluation framework for assessing LLM performance in TEI encoding. *Journal of Open Humanities Data*, 12(1), 39. <https://doi.org/10.5334/johd.484>
- Wong, N. Y. H. (2026). Modernism's vector: Dataset for a Chinese-Language journal in late 1950s Malaya. *Journal of Open Humanities Data*, 12(1), 40. <https://doi.org/10.5334/johd.485>

TO CITE THIS ARTICLE:

Kim, G., Park, S., Ji, H., Lee, H., Lee, B., Jeong, C., & Kim, B. (2026). KNoTE: A TEI Encoded XML Dataset of Modern Korean Novels. *Journal of Open Humanities Data*, 12: 102. DOI: <https://doi.org/10.5334/johd.590>

Submitted: 11 May 2026

Accepted: 07 July 2026

Published: 30 July 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.