JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 5, pp. 165-178, May 2026
https://doi.org/10.9708/jksci.2026.31.05.165

A Study on Rule-Based Text Mining and Named Entity Recognition Approaches for Data Formalization of "The Memoirs of Casanova"

Sunghoon Jeong*, Yujin Noh**, Jeongeun Hwang***, Jinsun Kim***, Hajin Kim***, Hyoji Ha****

*Student, Dept. of History, Ajou University, Suwon, Korea
**Student, Dept. of Cultural and Contents, Ajou University, Suwon, Korea
***Student, Dept. of French Language and Literature, Ajou University, Suwon, Korea
****Research Assistant Professor, Humanities Research Institute, Ajou University, Suwon, Korea

[Abstract]

This study constructs and analyzes structured data by applying digital humanities methodologies to Casanova's memoirs, considered the most extensive autobiographical record of the 18th century. Based on the text from the memoir, we designed a data refinement pipeline that integrates NLP technologies—such as Stanza, spaCy, and NRCLex—with generative AI. Specifically, to resolve the complex naming conventions and title issues, a rule-based algorithm was introduced to verify data accuracy. Through this process, we constructed structured data for a total of 1,924 individuals, encompassing seven attributes including gender, mention frequency, and associated emotion words. The analysis of the data revealed a distinct alternation between sections peaking with large-scale influxes of new characters and sections where a select few individuals repeatedly appeared, creating dense relationship networks. Notably, in sections centered around public and institutional events, as well as in the latter half of the narrative, the proportion of female characters plummeted to below half, demonstrating a pattern where the narrative converges upon a core group of male figures. To validate the efficacy of this methodology, the identical system was applied to Benjamin Franklin's autobiography. The results demonstrated stable operation and achieved higher accuracy in areas such as regional classification compared to conventional methods, thereby proving its accessibility and scalability.

▸ **Key words**: Rule-based text mining, Casanova memoirs, NER, Digital humanities, Tabular data

• First Author: Sunghoon Jeong, Corresponding Author: Hyoji Ha
*Sunghoon Jeong (jsh010219@ajou.ac.kr), Dept. of History, Ajou University
**Yujin Noh (no0405@ajou.ac.kr), Dept. of Cultural and Contents, Ajou University
***Jeongeun Hwang (zeong911@ajou.ac.kr), Dept. of French Language and Literature, Ajou University
***Jinsun Kim (seonkj247@ajou.ac.kr), Dept. of French Language and Literature, Ajou University
***Hajin Kim (gimh09925@ajou.ac.kr), Dept. of French Language and Literature, Ajou University
****Hyoji Ha (hjha0508@ajou.ac.kr), Humanities Research Institute, Ajou University
• Received: 2026. 03. 27, Revised: 2026. 04. 30, Accepted: 2026. 05. 05.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

본 연구는 18세기 최대의 자전적 기록물인 카사노바의 회고록을 대상으로 디지털 인문학적 방법론을 적용하여 정형 데이터를 구축하고 분석하였다. 회고록의 텍스트를 기반으로 Stanza, spaCy, NRCLex 등 자연어 처리 기술과 생성형 AI를 결합한 데이터 정제 파이프라인을 설계하였다. 특히 복합적 호칭 문제를 해결하기 위해 규칙 기반 알고리즘을 도입하여 데이터의 정확도를 검증하였다. 이를 통해 성별, 언급 횟수, 감정어 등 7개 속성을 포함한 총 1,924명의 정형 데이터를 구축하였다. 데이터 분석 결과, 대규모 신규 유입으로 최대치를 기록하는 구간과 소수의 인물이 반복 등장하며 관계가 밀집되는 구간이 뚜렷하게 교차했다. 특히 공적·제도적 사건이 중심이 되는 구간과 서사 후반부일수록 여성 비율이 절반 이하로 급감하며 핵심 남성 인물군으로 서사가 수렴하는 패턴을 확인하였다. 본 방법론의 유효성을 검증하기 위해 벤자민 프랭클린의 회고록에 동일한 체계를 적용한 결과, 안정적인 작동과 함께 기존 방식 대비 지역 분류 등에서 더 높은 정확도를 나타내어 접근성과 확장성을 입증하였다.

▸ **주제어**: 규칙 기반 텍스트 마이닝, 카사노바 회고록, 개체명 인식, 디지털 인문학, 정형 데이터

I. Introduction

자코모 카사노바(Giacomo Girolamo Casanova, 1725~1798)는 '희대의 유혹자'나 '희대의 방탕아'와 같은 신랄한 평가가 이어지는 평민 출신으로, 신분이 중요한 18세기 유럽 사회에서 교황 및 추기경, 볼테르(Voltaire), 괴테(Goethe), 모차르트(Mozart) 등 다양한 신분 계층과 교류하며 당시 유럽 사회 전반에 영향을 미친 복합적인 인물로 알려져 있다. 그의 자전적 회고록 『The Memoirs of Jacques Casanova de Seingalt』(원제: Histoire de ma vie)는 이러한 다층적 교류의 생생한 기록으로, 귀족과 성직자, 상인, 예술가 등 수천 명의 인물이 등장하는 방대한 1차 사료이다[1].

본 연구의 목적은 카사노바를 긍정적으로 재평가하거나 그의 삶을 변호에 머무르는 대신, 그의 회고록에 담긴 이야기를 18세기 유럽 사회와 문화[2] 들여다보는 분석의 도구로 활용하는 데 중점을 둔다. 왕실 인물이 아니면서도 다양한 계층과 관계를 맺은 카사노바의 사례는 기존 역사 서술이 놓쳤던 18세기 유럽 사회에 대한 입체적 관점을 새롭게 제시할 수 있다는 점에서 학술적 의의가 있다.

최근 디지털 인문학(Digital humanities) 분야에서는 방대한 문학 텍스트를 계산적 방법으로 분석하는 연구가 활발히 진행되고 있다. 특히 역사적 인물의 회고록은 당시 사회 구조와 인간 관계망을 연구하는 데 독보적인 1차 사료적 가치를 지닌다. 그러나 회고록은 저자의 기억에 의존하는 주관적 서술이라는 특성으로 인해 왜곡이나 자기 미화, 선택적 정보 전달의 가능성이 있다[3]. 따라서 텍스트에서 나타나는 서사 구조와 인물 관계 파악을 위해 데이터의 정형화가 선행되어야 한다[4]. 또한 기존의 디지털 인문학에서 주로 다루는 문학작품은 하나의 통일된 문체와 언어를 사용하여 AI 학습 과정이 비교적 용이하다. 당대 유럽 지식인 사회의 다국어 사용 관습으로 인해, 단일 저자의 자서전일지라도 텍스트 내의 다양한 언어적 층위와 고유명사 표기법이 나타난다. 따라서 회고록에 걸맞는 규칙 기반 텍스트 마이닝 및 개체명 인식 방법을 고안하여 19세기 이전 유럽 인물들의 회고록 데이터 구축 접근법을 제시해 보고자 한다.

본 연구는 회고록 전체 29개 에피소드에 걸쳐 Stanza[5] 기반 개체명 인식(NER), spaCy[6] 자연어 처리, NRCLex[7] 감정 분석을 결합한 데이터 정제 파이프라인을 구축하고, 1,924명의 고유 인물을 추출하여 카사노바와의 관계성, 그리고 관계의 강도를 나타내는 데이터를 구축하였다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 관련 연구를 검토하고, Ⅲ장에서는 데이터 수집 및 정제 방법론을 상세히 기술한다. IV 장은 구축된 데이터에 대한 분석 결과를 서술하고, V 장은 타 모델과의 비교 검증 및 다른 회고록에 적용을 통한 확장성 검증을 다룬다. 마지막으로 VI 장에서는 본 연구의 결론 및 제언을 기술한다.

II. Related works

1. Data structuring and metadata extraction from historical texts and archives

김바로·강우규[8]는 고전문학 연구에 빅데이터 분석을 접목하여 텍스트 내 인물을 식별하고 그 관계를 정량화하는 방법론을 고찰하였다. 연구진은 형태소 사전과 띄어쓰기 등 '표상체 식별 체계'를 통해 텍스트 요소를 식별하고, 원문을 현대어로 변환하여 HAM(한국어 형태소 분석 라이브러리)으로 주석 말뭉치를 구축하는 전처리 과정을 제시하였다. 또한, 표상체와 해석체를 연결하기 위해 NRC 감정 사전을 활용하여 형태소별 감정 지수를 매칭하고, 연구자가 직접 서사 단락에 내용 유형을 태깅하여 계층적 군집 분석을 수행하는 등 이본 계통 분류의 정교함을 높였다. 다만, 사전 기반 분석이 문맥에 따른 감정 변화를 반영하지 못하고 번역 과정에서 의미 왜곡이 발생하는 한계를 지적하며, 중세 한국어에 특화된 형태소 분석기와 유연한 감정 지수 부여 알고리즘 구축의 필요성을 강조하였다.

임진왜란 시기 데이터 분석 연구[9]는 방대한 역사 기록물을 디지털 분석이 가능한 구조로 변환하기 위해 체계적인 데이터 구축 공정을 수행하였다. 해당 연구는 선조실록 등 임진왜란 시기의 주요 사료를 대상으로 비정형 한문 기록과 국역 텍스트를 분석 단위로 분절하였으며, 연도·월·일 등의 시간 정보와 교전 지역 등의 공간 정보를 메타데이터로 추출하여 정형 데이터 세트를 구축하였다. 또한, 전쟁 중 발생한 주요 사건을 빈도와 유형에 따라 분류하여 시계열 데이터로 변환함으로써 전쟁의 전개 양상을 수치화할 수 있는 토대를 마련하였다. 특히 사료에 등장하는 지명을 현대 좌표계와 매칭하여 공간 데이터로 변환함으로써, 기록 속의 텍스트를 위치 정보 기반의 가시적인 데이터 모델로 고도화하여 정량적 역사 분석의 기초를 제공하였다.

2. Morphological preprocessing and quantitative representation of unstructured text data

심규진[10]은 학술 문헌의 비정형 텍스트를 LDA 토픽 모델링과 네트워크 분석 알고리즘에 적용하기 위해 철저한 노이즈 제거와 수치화 기반의 데이터셋 구축 과정을 보여준다. 먼저 KCI, RISS 등 학술 데이터베이스를 통해 수집된 1,175편의 문헌 중 분석 목적에 부합하는 202편만을 엄선하는 1차 필터링을 수행한다. 이후 텍스트 마이닝 전처리 과정을 통해 분석의 신뢰도를 떨어뜨리는 불용어를 철저히 제거하고, 문맥상 유의미한 형태소(명사 등)만을 추출하는 형태론적 정제를 거친다. 최종적으로 정제된 텍스트로부터 단어 출현 빈도(TF) 및 역문서 빈도(TF-IDF) 가중치를 산출하여, 정성적인 학술 텍스트를 통계적 처리가 가능한 규격화된 수치형 데이터 매트릭스로 변환하여 구축한다.

강지훈[11]은 인문학 사전 편찬에 사용된 정성적이고 비정형적인 텍스트에서 객관적인 글쓰기 패턴을 도출하기 위해, 텍스트 마이닝 기술을 활용한 강도 높은 데이터 정제 및 구축 과정을 제시한다. '지중해 문명 사전'의 표제어 텍스트를 분석 대상으로 삼아, 형태소 분석기 및 텍스트 마이닝 알고리즘을 적용하여 문장 내에 포함된 조사, 어미 등 분석에 무의미한 문법적 요소들을 제거한다. 이와 함께 각종 특수기호나 중복 표현 등의 노이즈를 일괄적으로 소거하는 데이터 클렌징 작업을 수행한다. 이러한 규격화 과정을 통해 순수한 핵심어 단위로 텍스트를 정제하며, 어휘의 출현 빈도와 핵심어 간의 패턴을 수치화하여 인문학 콘텐츠의 질적 검증이 가능한 정량적 데이터셋으로 가공한다.

III. Data collection and preprocessing

1. Raw data

본 연구에서는 구텐베르크 프로젝트(Project Gutenberg)에 공개된 『The Memoirs of Jacques Casanova de Seingalt, 1725-1798. Complete by Casanova』를 분석 대상 텍스트로 사용하였다. 이 텍스트는 한 볼륨이 약 5개의 에피소드로 이루어져 있으며, 각 에피소드는 다시 3~6개의 챕터로 구성되어 있다. 분석 데이터는 총 6개 볼륨, 30개 에피소드, 122개 챕터로 이루어져 있으나, 에피소드 30 이후에 수록된 부록과 작가 생애 소개의 저자는 카사노바가 아니기에 분석에서 제외하여 최종적으로 29개 에피소드를 분석 대상으로 삼았다. 에피소드 단위로 데이터 정제 작업을 수행하였으며, 전체 데이터 처리는 Google Colab 환경에서 진행하였다. 수집된 텍스트 데이터의 구조화 및 메타데이터 설계에 있어서는 이정연[12]이 제안한 구술사 기록물 메타데이터 모델링을 참고하였다. 이 연구는 프로젝트, 관리, 레코드, 관련 레코드의 4개 영역으로 구성된 계층적 메타데이터 스키마를 제시하며, 기록물의 내용·형식·저작 정보를 체계적으로 기술하는 방법론을 제공한다.

2. Data preprocessing pipeline

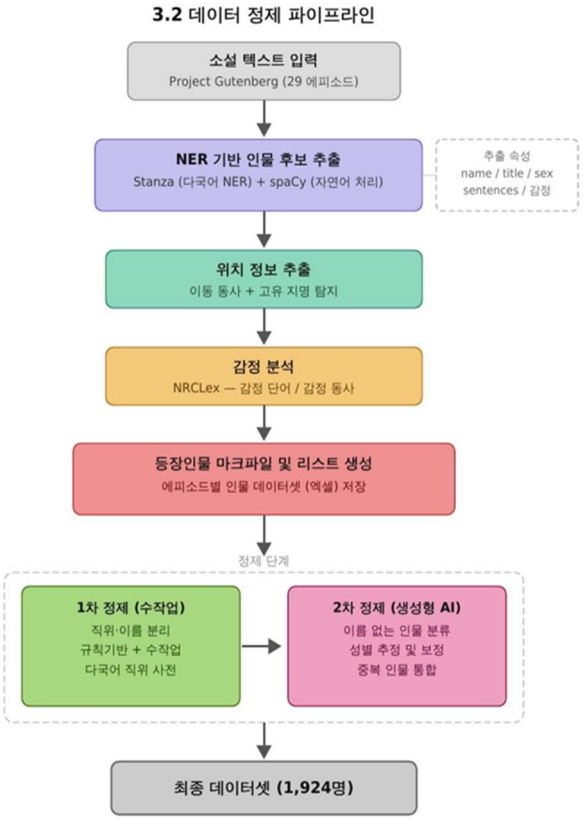


Fig. 1. Data Preprocessing Pipeline

데이터 정제 과정은 생성형 AI 기반 코드만으로는 정확도와 전문성에 한계가 있다고 판단하여, 선행연구를 참고해 Stanza NER 기반 인물 추출 라이브러리와 spaCy 자연어 처리 도구 등 기존 파이썬 라이브러리를 추가 활용하여 정제 과정의 신뢰성과 정확도를 높이고자 하였다. 본 연구에서 Stanza를 핵심 개체명 인식(NER) 도구로 채택한 근거는 다음과 같다. 첫째, Stanza는 Stanford NLP Group에서 개발한 다국어 자연어 처리 파이프라인으로, 66개 이상의 언어에 대해 토큰화, 형태소 분석, 품사 태깅, 의존 구문 분석, 개체명 인식 등 통합적인 처리 기능을 제공한다. 카사노바 회고록은 영어 번역본을 기반으로 하되, 원문의 특성상 프랑스어, 이탈리아어, 라틴어 등 다국어 표현이 등장하므로, 단일 언어에 최적화된 도구보다 다국어 처리에 강점을 가진 Stanza가 적합하다고 판단하였다.

둘째, Stanza는 역사적 텍스트 분석에서의 적용 가능성이 선행연구를 통해 검증된 바 있다. Zilio et al.의 연구[13]는 18세기 포르투갈어 의학 문헌을 대상으로 Stanza의 성능을 평가한 결과, 사용자 지정 토큰화 및 문장 분할 설정을 보다 유연하게 수용할 수 있어 역사적 텍스트의 비표준적 철자와 구문 구조에 대응하기에 적합하다고 하였다. 해당 연구에서는 철자 정규화를 통해 품사 태깅 정밀도가 약 4% 이상 향상됨을 확인하였으며, 이는 근대 유럽어 텍스트에 현대 NLP 도구를 적용할 때 Stanza가 효과적으로 기능할 수 있음을 실증적으로 보여준다. 본 연구의 분석 대상인 카사노바 회고록 역시 동일한 18세기 유럽 문헌이라는 점에서, 해당 선행연구의 방법론적 유효성이 본 연구에도 적용 가능하다.

셋째, Stanza는 사전 훈련된 신경망 기반 모델을 활용하여 문맥적 의미를 반영한 개체명 인식이 가능하며, 특히 PERSON 및 LOC 태그의 추출 정확도가 높아 본 연구의 인물 데이터 구축에 적합하다. 또한 Python 기반의 간결한 API를 제공하여 Google Colab 환경에서의 파이프라인 구축과 반복적 실험에 용이하다는 실용적 장점도 고려되었다. 구체적인 실행 설정으로, 언어 모델은 다국어(Multilingual) 혼재 상황을 고려하여 기본 모델을 적용하였으며, 추출 태그 유형은 인물(PERSON)과 지역(LOC)을 중점적으로 추출하도록 제한하여 노이즈를 최소화하였다. 전체 정제 파이프라인은 Fig 1.과 같이 소설 텍스트 입력 후, NER 기반 인물 후보 추출, 위치 정보 추출, 등장인물 마크파일 및 리스트 생성의 순서로 진행된다.

정제된 데이터는 name(이름), title(직위), sex(성별), mention_count(언급 수), sentences(출현 근거 문장), emotion_words(감정 단어), emotion_verbs(감정 동사)의 총 7개 컬럼으로 구성되어 있다. 출현 근거 문장은 텍스트 내 실제 등장 여부를 직접 확인하기 위해 확보한 것이며, 감정 단어와 감정 동사는 인물 간 관계 태깅을 위한 기초 데이터로 활용된다[14,15].

Table 1. List of Libraries Used for Data Preprocessing

Tools /Libraries	Description	Primary Purpose
spaCy	High-speed Python-based NLP library	Sentence tokenization and natural language processing
Stanza	Highly accurate multilingual NLP tool	Person extraction (NER) - GitHub
NLTK	Comprehensive Python-based NLP library	Text preprocessing support
NRCLex	Based on a sentiment dictionary of 14,200 words	Analysis of emotion words and sentiment verbs

3. First-Stage Semi-Automated preprocessing

NER 기반으로 추출된 인물 데이터는 성별 및 직위 정보가 누락되거나 동일 인물이 중복 등장하는 문제가 존재

하였다. 이에 따라 출현 근거 문장을 기반으로 수작업 검토를 수행하여 데이터의 정확도를 1차적으로 보완하였다.

본 연구에서는 NER 기반으로 추출된 인물 데이터의 정확도를 향상시키기 위해 인물 이름(Name)과 직위(Title)의 분리 및 정제 작업을 수행하였다. 초기 데이터에서는 직위와 이름이 혼합된 형태(예: "Count de X", "Madame Y")로 존재하는 경우가 다수 확인되었으며, 이를 구조적으로 분리하기 위해 자동화된 규칙 기반 처리와 수작업 보정을 병행하였다.

우선 코드 기반 전처리 단계에서는 인물명(Name) 필드에서 특정 직위 키워드를 탐지하여 이를 직위(Title)와 이름(Name)으로 분리하였다. 직위는 성별에 따라 다음과 같이 정의하였다.

남성 직위: M., MM., king, prince, Count/Comte, Earl, abbé/abbe, duc/duke, Marquis, Baron, Don, cardinal, Pope 등

여성 직위: Mdlle, Madame, princess, Countess/Comtesse, abbesse, Duchesse, Donna 등

해당 규칙에 따라 Name 문자열의 앞부분에서 직위가 탐지될 경우, 이를 Extracted_Title과 Cleaned_Name으로 분리하고, 직위가 없는 경우에는 Title을 "없음"으로 처리하였다. 이후 Extracted_Title이 존재하는 경우, 이를 최종 Title 컬럼에 반영하고 정제된 이름을 Name으로 재구성하였다.

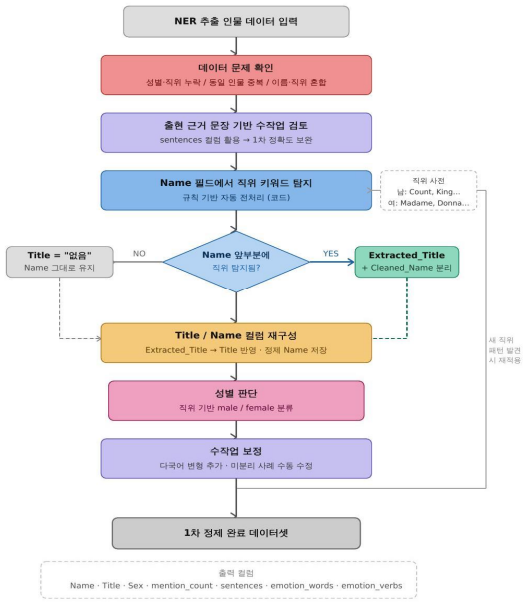


Fig. 2. First-Stage Semi-Automated Preprocessing

그러나 자동화된 규칙만으로는 다양한 언어(프랑스어, 이탈리아어, 영어 등)에서 나타나는 직위 표현과 변형[16]을 모두 포괄하기 어려웠다. 이에 따라 본 연구에서는 각 볼륨별로 추출된 인물 데이터셋을 직접 검토하며 추가적인 직위 패턴을 수작업으로 보완하였다.

특히 다음과 같은 작업을 수행하였다.

- 동일 직위의 다양한 언어 표현 추가 (예: Comte, Conte, Count 등)
- 축약형 및 변형 형태 보정 (예: M. → Monsieur, Mme → Madame)
- 성별 판단에 영향을 주는 직위 확장
- 직위와 이름이 분리되지 않은 사례 수동 수정

이 과정에서 새로운 직위 표현이 발견될 경우, 이를 직위 사전에 추가하고 전체 데이터에 재적용하는 방식으로 반복적인 정제 과정을 수행하였다. 이러한 반자동(automated + manual) 정제 방식은 다국어 텍스트 환경에서 인물 속성 추출의 정확도를 향상시키는 데 중요한 역할을 하였다.

이와 같은 1차 정제 과정을 통해 데이터의 구조적 오류를 최소화하였다.

4. Second-Stage preprocessing using generative AI

2차 정제에서는 생성형 AI를 활용하여 중복 인물 제거 등의 보정 작업을 진행하였다. 또한 인물 간 관계를 우선적으로 파악하고 이를 가족·연인·친구·왕족·비즈니스 등으로 대분류하였다. 표 2는 본 연구에서 정의한 인물 관계 대분류 체계를 나타낸다.

Table 2. Major Category of Interpersonal Relationships

Major Category	English Noun Form
I. Narrator	Narrator, Protagonist
II. Family & Relatives	Father, Mother, Brother, Sister, Grandmother, Aunt, Niece, Cousin, Widow, Canoness
III. Romantic Relations	Mistress, Rival, Waiting Woman
IV. Helper/Friend /Professional	Patron, Benefactor, Friend, Advisor, Broker, Physician, Chemist, Poet, Director
V. Government /Law Enforcement	Pope, Statthalter, Majordomo, Secretary, Councillor, Nobleman, Marquis, Agent
VI. Peripheral /Acquaintances	Servant, Coachman, Nurse, Door-Keeper, Cook, Merchant, Captain, Lady, Wife
VII. Other /Historical Figures	Sorceress, Academicians, Sclavs, Thief, Murderer, Greybeard, Executioner
VIII. Servants & Employees	Servant, Coachman, Nurse, Door-Keeper, Postillion

규칙 기반 및 수작업 정제 이후에도 남아 있는 데이터의 불확실성과 모호성을 해결하기 위해 생성형 AI(ChatGPT, Google Gemini)를 추가적으로 활용하였다. 생성형 AI는 단일 단계에서 결과를 도출하기보다, 반복적인 프롬프트 수정과 결과 검증을 통해 점진적으로 정제 정확도를 향상하는 방식으로 활용되었다.

4.1 Classification of unidentified individuals (p.list)

NER 기반 인물 추출 과정에서 “이름 없는 인물”(예: the priest, a lady, the servant 등)이 다수 포함되었으며, 이를 별도의 P.list로 분리하여 관리하였다. 해당 데이터는 고유 인물인지 일반 명사인지 판단하기 위해 생성형 AI를 활용하였다.

분류 체계는 고유 인물과 일반 명사 2가지로 분류했다. 이를 판단하는 기준은 다음과 같다. 첫째, 이름이 존재하지 않고, 역할만을 나타내면 일반 명사, 둘째, 특정 개인을 지칭하면 고유 인물로 분류. 이 과정을 통해 비인물 데이터 및 일반 명사형 표현을 제거하거나 별도 카테고리로 유지하였다.

4.2 Gender estimation and correction

NER 결과에서 성별이 누락된 인물에 대해 생성형 AI를 활용하여 성별을 추정하였다. 인명의 성비, 직위(예: Madame, Countess 등)를 기준으로 인물의 성별을 분류하였다. AI의 결과는 자동 반영하지 않고, 실제 텍스트 문장과 출현 맥락을 연구자가 직접 검토하여 최종 성별을 확정하였다.

4.3 Merging duplicate individuals

문학 텍스트에서는 동일 인물이 다양한 이름, 축약형, 직위 포함 표현, 또는 언어적 변형으로 등장하는 경우가 많아 중복 인물 통합이 중요한 정제 단계이다[17]. 이러한 이름 변형은 약어, 별칭, 다국어 표기 등 다양한 형태로 나타나며, 단순 문자열 일치만으로는 동일 인물 여부를 판단하기 어렵다. 이를 해결하기 위해 생성형 AI를 활용하여 동일 인물 후보군을 도출하였다.

동일 인물을 지칭하는 가능성이 있는 항목들을 1차로 묶어 그룹화를 한 뒤, 아래 판단 기준을 통해 세부 분류를 진행하였다 . 판단 기준은 첫째, 이름의 일부가 동일한 경우(예: 미들네임), 둘째, 직위만 다른 경우, 셋째, 언어 변형 (Comte/Count 등)이다.

AI가 제시한 결과는 1차 후보군으로 활용되었으며, 이후 문맥 기반 검토를 통해 최종적으로 통합 여부를 결정하였다.

특히 본 연구에서 분석한 특정 볼륨에서는 다음과 같은 유형의 중복 인물 사례가 관찰되었다.

Table 3. Examples of Name Variations and Title Changes for the Same Individual

Episode	13	15	19
Name variant / Title	Cornelis / Cornely / Trenti / Therese Trenti	Theophile Falengue / Merlin / Merlin Coccaeus	Corticelli / Lascaris / Countess Lascaris
Real Name (original name)	Teresa Imer	Teofilo Folengo	Corticelli (pseudonyms), Lascaris (original name)
Notes	Regional use of pseudonyms	Pen names and name variants	Use of pseudonyms and fake noble titles
Source Sentence	"She had exchanged the name of Trenti for that of Cornelis,..."		"the Corticelli, late Lascaris"

Table 3은 동일 인물이 다양한 이름과 호칭으로 등장하는 사례를 정리한 것이다. 이러한 이름 변이는 예명, 필명, 지역별 활동명, 또는 사회적 신분 위장을 위한 가명 등 다양한 이유로 발생하며, 인물 네트워크 구축 과정에서 중복 인물 통합의 필요성을 보여준다[18].

이러한 사례는 다음과 같은 특징을 가진다. 첫째, 예명과 본명을 병행하여 사용한다. 둘째, 활동 지역에 따라 등장인물의 이름이 변화한다. 셋째, 직위 및 신분에 따라 이름이 변화한다. 넷째, 필명 및 가명을 사용한다.

이와 같은 특성은 단순 문자열 유사도 기반 방법만으로는 동일 인물 여부를 정확히 판단하기 어렵다는 점을 보여준다. 따라서 본 연구에서는 생성형 AI를 활용하여 동일 인물 후보군을 도출한 후, 문맥 및 서사적 정보를 기반으로 연구자가 직접 검증하는 방식으로 최종 통합을 수행하였다.

4.4 Iterative prompt refinement process

생성형 AI 활용은 단일 프롬프트가 아닌 반복적 개선 방식으로 이루어졌다. 초기에는 단순 분류 요청 형태로 시작하였으나, 결과의 정확도를 높이기 위해 다음과 같은 방식으로 프롬프트를 점진적으로 개선하였다.

Table 4. Prompt Refinement Rules

Rule	Description
1 Rule	Explicit evaluation criteria
2 Rule	Expansion of output formats [table, grouping]
3 Rule	Structured input information [name, title, context]
4 Rule	Specification of multilingual and variant conditions

이러한 반복적 프롬프트 엔지니어링을 통해 데이터 정제의 일관성과 재현 가능성을 확보하였다.

위 방법을 통한 생성형 AI의 결과는 참고 자료로 활용되었으며 데이터셋에는 연구자의 수작업 검증을 거친 결과만 반영하였다. 모든 방법을 거친 결과물의 예시는 아래와 같다.

episode_1	1	He was se	king	King	Alfonso	Alfonso	1 YES
episode_1	2	He ran aw	pope	Pope	Martin III.	Martin_III.	2 YES
episode_1	2	He ran aw	pope	Pope	Anna	Anna	7 YES
episode_1	2	He ran aw	pope	Pope	Don Jacob	Don_Jacob	34 NO
episode_1	2	He ran aw	don	Don	Martin III.	Martin_III.	20 NO
episode_1	2	He ran aw	don	Don	Anna	Anna	15 NO
episode_1	2	He ran aw	don	Don	Don Jacob	Don_Jacob	12 NO
episode_1	2	He ran aw	don	Don	Martin III.	Martin_III.	32 NO
episode_1	2	He ran aw	don	Don	Anna	Anna	27 NO
episode_1	2	He ran aw	don	Don	Don Jacob	Don_Jacob	0 YES
episode_1	3	All the chi	don	Don	Don Juan	Don_Juan	0 YES
episode_1	3	All the chi	don	Don	Donna Ele	Donna_Ele	9 NO
episode_1	3	All the chi	don	Don	Marco An	Marco_An	20 NO
episode_1	3	All the chi	donna	Donna	Don Juan	Don_Juan	9 NO
episode_1	3	All the chi	donna	Donna	Donna Ele	Donna_Ele	0 YES
episode_1	3	All the chi	donna	Donna	Marco An	Marco_An	11 NO
episode_1	4	In 1481, D	king	King	Don Juan	Don_Juan	9 NO

Fig. 3. Merged Person Data

5. Region extraction per sentence

카사노바의 이동 경로를 추적하기 위해 각 볼륨의 문장을 분석하면서 이동을 의미하는 동사(went, goes, going, travel, traveled, journeyed, arrived, departed, moved, migrated, rode, sailed, crossed, entered, returned, exiled, escaped, fled 등)와 지역을 의미하는 고유명사(Venice, Paris, Tyrol, London 등)가 동시에 등장하면 이를 지역 이동으로 판단해 true로 기록하고, 문장 속에 지역명이나 국가명이 나타날 경우 이를 locations_in_sentence와 current_location에 저장하는 코드를 구현하였다. 만일 이동동사나 지명이 한 문장에 하나의 요소만 등장한다면 앞서 저장한 current_location으로 판단하는 방식으로 모든 문장의 지역데이터를 구축하였다. 해당 알고리즘에 대한 코드는 아래와 같다.

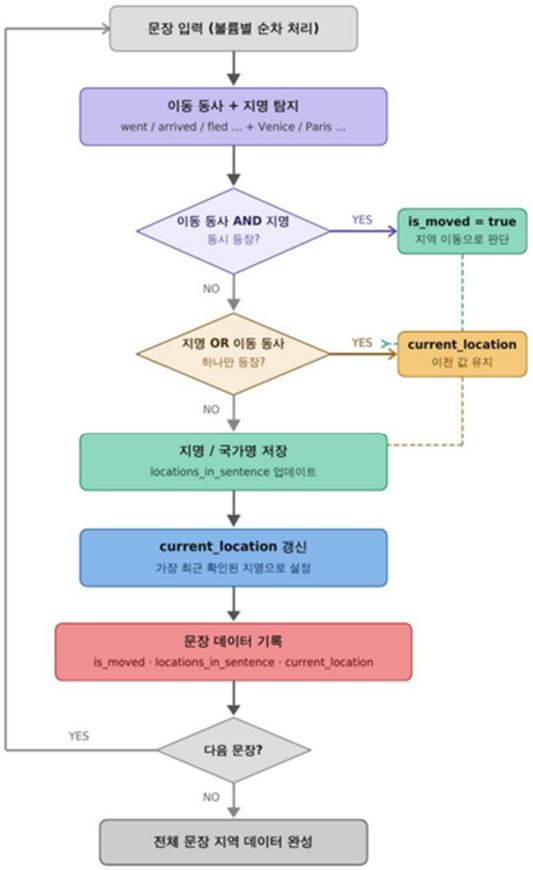


Fig. 4. Region Extraction per Sentence

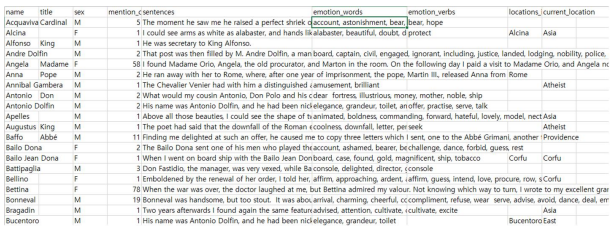
Algorithm 1. Rule-based region classification algorithm

Code

```
FOR EACH sentence IN sentence_spans:
    # 1. Initialize data
    text = Get text from sentence
    doc = Process text using NLP model

    # 2. Extract Movement and Location cues
    movement_detected = CHECK IF text matches movement_patterns
    location_list = EXTRACT clean locations from doc

    # 3. Logic for updating current location
    IF (movement_detected IS TRUE) AND (location_list IS NOT EMPTY):
        # Update global tracker to the first location found in the sentence
        current_location = location_list[0]

    # 4. Store the analysis results
    APPEND to results:
        - index: current sentence index
        - text: original sentence
        - movement_trigger: movement_detected
        - locations_found: all items in location_list
        - current_location: updated or existing location
```

장소 추적 모듈은 단일 패스 규칙 기반 파이프라인으로, 지정된 디렉토리 내의 모든 일반 텍스트 파일을 정렬된 순서로 결합한 뒤 spaCy의 en_core_web_sm 모델을 사용하여 문장 단위로 분할한다. 각 문장에 대해 시스템은 두 가지 조건을 동시에 확인하는데, 첫째는 "went," "arrived," "departed," "fled" 등 사전 정의된 22개의 이동 동사 목록에 해당하는 동사가 문장 내에 존재하는지 여부이며, 둘째는 spaCy의 개체명 인식기가 해당 문장에서 하나 이상의 GPE 또는 LOC 개체를 식별하는지 여부이다. 이 과정에서 토큰 수준, 문장 수준, 문단 수준 등 어떠한 거리 지표도 사용되지 않으며, 단일 문장 경계 내에서 이동 동사와 지명 개체가 동시에 출현하는 것만이 주인공의 추정 위치를 갱신하는 유일한 조건으로 작동한다. 하나의 문장에 복수의 지명 개체가 등장할 경우, 출발지와 도착지 등의 통사적 역할을 구분하지 않고 선형 순서상 첫 번째 개체를 선택하는 방식으로 모호성을 해소한다. 갱신된 현재 위치는 새로운 동시출현 조건이 충족될 때까지 이후 모든 문장에 걸쳐 그대로 유지된다. 아울러 인명과 지명의 동음이의 문제를 처리하기 위한 추가 규칙이 적용되는데, 동일 문장 내에서 spaCy가 특정 토큰을 PERSON과 GPE/LOC로 동시에 레이블링한 경우 PERSON 레이블이 우선시되어 해당 토큰은 지명 후보에서 제외됨으로써 인명이 지리적 참조로 오분류되는 것을 방지한다.

위에서 제시한 규칙 기반의 알고리즘을 통하여 인물, 직위, 성별, 감정, 문장별 지역분류 데이터를 아래의 Fig. 5와 같이 구축하였다.



name	title	sex	mention_c	sentences	emotion_words	emotion_verbs	locations_	current_location
Acquaviva	Cardinal	M	5	The moment he saw me he raised a perfect shriek d	account, astonishment, bear,	bear, hope		
Alcina		F	1	I could see arms as white as alabaster, and hands lik	alabaster, beautiful, doubt, d	protect	Alcina	Asia
Alfonso	King	M	1	He was secretary to King Alfonso.				
Andre Dolfin		M	2	That post was then filled by M. Andre Dolfin, a man	board, captain, civil, engaged,	ignorant, including, justice, landed, lodging, nobility, police,		
Angela	Madame	F	58	I found Madame Orio, Angela, the old procurator, and Marton in the room. On the following day I paid a visit to Madame Orio, and Angela re				
Anna	Pope	M	2	He ran away with her to Rome, where, after one year of imprisonment, the pope, Martin III., released Anna from			Rome	
Annibal Gambera		M	1	The Chevalier Venier had with him a distinguished a	amusement, brilliant			Atheist
Antonio	Don	M	2	What would my cousin Antonio, Don Polo and his c	clear, fortress, illustrious, money, mother, noble, ship			
Antonio Dolfin		M	2	His name was Antonio Dolfin, and he had been nick	elegance, grandeur, toilet, an	offer, practise, serve, talk		
Apelles		M	1	Above all those beauties, I could see the shape of tv	animated, boldness, commanding, forward, hateful, lovely, model, nect		Asia	
Augustus	King	M	1	The poet had said that the downfall of the Roman e	coolness, downfall, letter, per	seek		Atheist
Balfo	Abbé	M	11	Finding me delighted at such an offer, he caused me to copy three letters which I sent, one to the Abbé Grimani, another				Providence
Bailo Dona		F	2	The Bailo Dona sent one of his men who played th	account, ashamed, bearer, be	challenge, dance, forbid, guess, rest		
Bailo Jean Dona		F	1	When I went on board ship with the Bailo Jean Don	board, case, found, gold, magnificent, ship, tobacco		Corfu	Corfu
Battipaglia		M	3	Don Fastidio, the manager, was very vexed, while Ba	console, delighted, director, c	console		
Bellino		F	1	Emboldened by the renewal of her order, I told her, a	affirm, approaching, ardent, c	affirm, guess, intend, love, procure, row, s	Corfu	
Bettina		F	78	When the war was over, the doctor laughed at me, but Bettina admired my valour. Not knowing which way to turn, I wrote to my excellent grar				
Bonneval		M	19	Bonneval was handsome, but too stout. It was abo	arrival, charming, cheerful, cc	compliment, refuse, wear	serve, advise, avoid, dance, deal, en	
Bragadin		M	1	Two years afterwards I found again the same featur	advised, attention, cultivate, c	cultivate, excite		Asia
Bucentoro		M	1	His name was Antonio Dolfin, and he had been nick	elegance, grandeur, toilet		Bucentoro	East

Fig. 5. Processed Data (Casanova)

IV. Data analysis

1. General statistical overview

전체 29개 에피소드에서 추출된 고유 인물 수는 1,924명이며, 에피소드별 평균 등장인물 수는 약 102명이다. 에피소드 6이 260명으로 전체 최대치를 기록하였으며, Vol.2 초입에서 신규 유입(첫 등장 214명)이 증가한 결과이다. 반면 저점은 에피소드 26(34명)으로, Vol.6의 말미에서 인물이 '줄어든다'기보다 핵심 인물 중심으로 압축되는 구간임을 알 수 있다(26개 에피소드에서 34명 중 32명이 2회 이상 언급).

Table 5. Top 5 Individuals by Mention Frequency

Ranking	Person Name	Gender	Total Mentions	Major Featured Episodes
1	Henriette	F	215	Ep.6
2	Esther	F	167	Ep.11-13
3	Bragadin	M	124	Ep.1-13
4	Casanova	M	110	Ep.1-19
5	Lawrence	M	110	Ep.10

2. Gender distribution analysis

에피소드별 성별 분포를 분석한 결과, 여성 비율이 가장 높은 구간 중 하나는 Ep.9(여성 34/71, 47.9%)로 균형잡힌 비율을 나타낸다. 반대로 Ep.10은 여성 비율 15.6%(24/154)로 낮아, 특정 사건 축(제도·권력·공적 관계)의 비중이 증가하는 에피소드에서 성별 구성도 함께 변화하는 양상을 보인다. 볼륨 평균으로는 Vol.4 여성 비율이 가장 높고(평균 35.9%), Vol.6이 가장 낮다(평균 19.1%). 이는 후반부로 갈수록 중심 인물인 핵심 남성 인물군으로 서사가 '수렴'하는 경향이 강함을 보여준다.

V. Evaluation of the proposed method

1. Comparison with other extraction methods

인물 추출 정확도 검증을 위한 데이터는 디지털 인문학 분야 연구자 3명이 독립적으로 구축하였다. 각 주석자는 Episode 1 원문 텍스트에서 인물명에 해당하는 개체를 개별적으로 태깅하고, 해당 인물의 출현 빈도를 각자 집계하였다. 이후 3인 간 교차 검토를 수행하여 인물 식별 여부 및 빈도 값의 최종 레이블을 확정하였다. 주석자 간 불일치가 발생한 경우, 다수결 원칙(3인 중 2인 이상 일치)을 우선 적용하되, 단독 알파벳 표기 인물(D, F, M. 등)과 같이 일반 단어와의 경계가 모호한 사례에 대해서는 출현 문맥을 개별적으로 확인한 뒤 합의 기반으로 최종 판정하였다. 이러한 복수 주석자 검토 체계를 통해 단일 연구자의 주관적 판단에 의한 편향을 최소화하고, 실제 빈도 데이터의 신뢰성을 확보하고자 하였다.

Table 6. Comparison with Other Person Extraction Methods

Evaluation Criteria	Index	Previous research	This Study	Superiority
Basedon Actual Frequency (n=43)	F1	.964	.838	Previous Research
	F1	1.000	.911	Previous Research
	Precision	.976	.912	Previous Research
	Recall	.952	.775	Previous Research
	MAE	.42	5.12	Previous Research
	RMSE	1.14	12.38	Previous Research
	Pearson r	.999	.775	Previous Research
	MAPE	5.9	34.2	Previous Research
Person Entity Corrupted by Abbreviation(n=5)	Contextual Discrimination Accuracy	Overestimation(D=76,F=78,M.=80)	Context Recognition (D=14, F=26, M.=1)	This Study
Total Agreement Rate(n=48)	Number of Exact Matching Person Entities	28/48(58.3%)	28/48(58.3%)	Equal

선행 연구(김바로·강우규)의 전체 텍스트 자동 카운팅과 본 연구의 수동 태깅을 단어 경계 기준 실제 빈도(n=43)와 대조하여 비교한 결과, 두 방법의 성능은 인물 유형에 따라 상이하게 나타났다. 실제 빈도를 기준으로 평가할 때 선행 연구의 자동 카운팅은 F1 = .964(±2 기준; Precision = .976, Recall = .952), MAE = 0.42, Pearson r = .999, MAPE = 5.9%를 기록하여 본 연구의 수동 태깅(F1 = .838, MAE = 5.12, Pearson r = .775, MAPE = 34.2%)을 전 지표에서 상회하였다. 이러한 격차는 중빈도 인물에서 특히 두드러지는데, Nanette의 실제 빈도는 47회임에도 본 연구의 태깅값은 2회(오차 −45)에 불과한 반면 자동 카운팅은 43회(오차 −4)로 실제에 근접하였으며, Malipiero(실제 32, 수동 오차 −31, 자동 오차 0), Marton(실제 23, 수동 오차 −22, 자동 오차 0), Rosa(실제 19, 수동 오차 −17, 자동 오차 −1)에서도 동일한 패턴이 확인되었다. 이는 본 연구의 참조 파일이 전체 빈도를 체계적으로 집계하지 않고 인물별 예시 문장을 선별 저장하는 방식으로 구축된 데 기인하는 것으로, 수동 태깅 범위의 전수화가 향후 과제로 요청된다. 그러나 단독 알파벳으로 표기되는 익명 인물(D, F, M.)에서는 역전이 발생한다. 자동 카운팅은 D(76회)·F(78회)·M.(80회)를 포함한 모든 문장을 집계함으로써 일반 단어 및 지명과의 혼동으로 인한 심각한 과다 계산을 초래한 반면, 본 연구의 수동 태깅은 "M. D——", "Madame F——"와 같은 문맥 패턴을 식별하여 각각 14회·26회·1회로 합리적인 값을 산출하였다. 핵심 고빈도 인물(Bettina: 자동 79 vs. 수동 78, Cordiani: 39 vs. 39, Gozzi: 26 vs. 26)에서는 양 방법이 사실상 동등하게 수렴하였으며, 전체 48명 중 28명(58.3%)에서 완전 일치하였다.

이상의 결과는 두 방법이 상보적 강점을 지님을 시사한다. 선행 연구의 자동 카운팅은 전체 텍스트를 빠짐없이 처리하여 실제 빈도 재현율이 높으나(Recall = .952) 동음이의 오염에 취약하고, 본 연구의 수동 태깅은 문맥 기반 변별이 가능하나 중·저빈도 인물에 대한 체계적 집계가 미흡하다(Recall = .775). 따라서 두 방법을 결합한 혼합 접근—자동 카운팅으로 전체 빈도를 우선 확보한 후 문맥 필터링으로 오염을 제거하는 방식—이 향후 인물 빈도 추출의 정확도를 최적화하는 전략으로 제안된다.

Table 7. Comparison of Movement Detection Performance

Index	Rule-Based	BERT NLP	Gap
TP	17	16	-1
FP	7	7	0
FN	3	4	+1
TN	4	4	0
Precision	0.708	0.696	-0.012
Recall	0.850	0.800	-0.050
F1-score	0.773	0.744	-0.029
Accuracy	0.677	0.645	-0.032

Table 8. Comparison of Location Extraction Performance

Index	Rule-based	BERT NLP	Gap
Destination Accuracy	0.850	0.500	-0.350
Current Location Accuracy	0.806	0.710	-0.096
Total average	0.777	0.682	-0.095

지역명 추출 과정에서 본 연구가 제시한 Rule-based 방식과 BERT NLP 방식[19]의 성능을 비교하기 위한 Ground Truth 데이터를 디지털 인문학 분야 연구자 4명이 독립적으로 수동 레이블링을 수행하여 구축하였다. 각 주석자는 Episode 1의 1,922개 문장을 대상으로 이동 동사와 지명의 동시 출현 여부(이동 TRUE/FALSE), 목적지(destination), 현재 위치(current_location)를 개별적으로 태깅하였다. 이후 3인 간 교차 검토를 통해 최종 31개

Ground Truth 문장(이동 TRUE 20건, FALSE 11건)을 확정하였다. 주석자 간 불일치가 발생한 경우, 다수결 원칙(3인 중 2인 이상 일치)을 우선 적용하되, "The eldest left Parma in 1712"와 같이 중심인물의 직접 이동인지 조상·타인 서술인지 판단이 모호한 사례에 대해서는 원문 서사 맥락과 해당 챕터의 전후 문장을 참조하여 합의 기반으로 최종 판정하였다. 이러한 복수 주석자 검토 체계를 통해 단일 연구자의 주관적 판단에 의한 편향을 최소화하고, Ground Truth 데이터의 신뢰성을 확보하고자 하였다.

평가 지표로는 Precision, Recall, F1-score, 목적지 추출 정확도(destination Accuracy), 현재위치 누적 추적 정확도(current_location Accuracy)를 사용하였다. 평가 결과, Rule-based 방식은 Precision 0.708, Recall 0.850, F1-score 0.773, 목적지 Accuracy 0.850, 현재위치 Accuracy 0.806을 기록하여, 전 지표에서 BERT NLP 방식(각각 0.696, 0.800, 0.744, 0.500, 0.710)을 상회하였다. 이러한 결과는 18세기 자서전이라는 텍스트 도메인이 갖는 세 가지 주요 특성, 즉 고정된 이동 표현 어순, 일관된 지명 표기, 그리고 화자 이동과 조상 서술의 혼재 양상에 기인한다.

첫째, 회고록 내 지역 이동 표현은 이동 의미를 지니는 단어가 선행된 후 구체적인 지명이 뒤따르며, 지역 변화 시에도 이와 동일한 패턴이 반복되는 고정된 어순을 지닌다. 둘째, 지명의 경우 인명이나 직책과 달리 당시에 사용되던 표기(Venice 56회, Padua 26회 등) 외의 다른 표현이 매우 적어 일관되게 출현하는 특성이 있다. 실제로 목적지 추출 정확도에서 두 방식의 차이가 0.350p로 가장 크게 나타났는데, 이는 GT 목적지가 존재하는 12개 문장 중 10건이 "to/at/in [지명]" 구조를 따르는 고정 전치사 패턴에 기인한다. BERT NLP의 다단계 의존구조 패턴은 "took me with her in a gondola as far as Muran"이나 "she summoned me to Venice"와 같은 우회적 이동 표현에서 목적지를 추출하는 데 실패한 반면, Rule-based 방식의 단순 "to [지명]" 매칭은 이러한 일관된 표기와 전치사 구조를 바탕으로 대상을 정확히 포착하였다.

셋째, 자서전에는 중심인물의 직접 이동뿐만 아니라, 중심인물이 타인을 언급하는 과정에서 새로운 지역이 등장하는 역사 및 조상 서술이 혼재되어 있다. 이로 인해 "The eldest left Parma in 1712"(조상 이동)나 "I remained at Pasean"(체류 표현) 등 이동 동사와 지명이 공존하나 실제 중심인물의 이동이 아닌 문장에서 두 방식 모두 7건의 오탐지(FP)가 동일하게 발생하였다. 특히 문맥 기반의 BERT NLP는 이러한 오탐지를 줄이고자 역사 맥락 필터를 추가 적용하였음에도, "In 1481, Don Juan…was compelled to leave Rome"과 같이 역사 연도와 조상 인물명이 공존하는 문장에서 실제 이동까지 과도하게 제거하여 미탐지(FN)가 1건 증가(3건→4건)하고 Recall이 0.050p 추가 하락하였다. 나아가 챕터별 이동 경로 일치율은 전체 7개 챕터 중 4개(57.1%)에서만 두 방식이 동일한 경로를 산출하였는데, 불일치가 집중된 Chapter I~III에서는 BERT NLP의 초기 목적지 오추출이 타인의 위치를 중심인물의 현재 지역으로 잘못 인식하게 만들어 이후 현재위치의 연쇄 오염을 야기하였다. 이는 현재위치 누적 추적 정확도에서 0.096p의 격차가 발생한 주된 원인이었다. 반면, 본 논문에서 제시한 규칙 기반의 알고리즘은 이러한 위치 오염의 가능성을 효과적으로 배제할 수 있었다. 결론적으로, 중심인물의 지역 이동 데이터를 구축하는 과정에서는 18세기 자서전 도메인의 세 가지 특성으로 인하여 복잡한 문맥 기반 의존구조 분석보다 규칙 기반의 알고리즘이 구조적으로 더 유리하게 작용함을 정량적 결과가 뒷받침한다.

2. Applicability to other memoir data

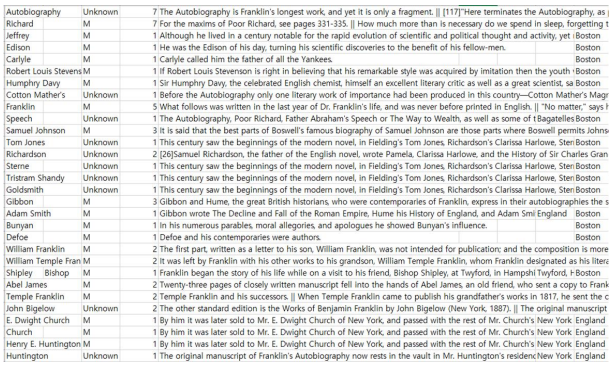
카사노바와 동시대 인물인 벤자민 프랭클린의 회고록 전문[20]에 본 논문에서 제시한 모델을 적용하여 다른 데이터로의 확장성을 검증하고자 하였다. 기본적인 모델은 그대로 사용하되, 중심인물과 직책명에 대한 사전만을 별도로 입력한 뒤 실행하였다.

Fig. 6과 같이 벤자민 프랭클린에 대한 직책, 직위 사전을 만든 뒤, 해당 목록 사전만 교체함으로써 벤자민 프랭클린 회고록에 적합하게 조정하였다. 벤자민 프랭클린의 이름과 이름의 변형들, 직책-직위 사전만 조정하여 Fig. 7과 같은 결과물이 출력되었다.

Fig.7의 결과물과 카사노바 데이터를 비교한 결과, 오류 없이 정상적으로 작동하였다. 인명과 성별, 직위를 분류하였으며, 해당 근거 문장과 문장별 지역을 정상적으로 분류하였다.

```
# 직함/호칭 목록 (프랭클린 자서전 맞춤)
TITLES_SET = {
    # 종교
    "bishop", "reverend", "rev.", "pastor", "minister", "father",
    "brother", "cardinal", "archbishop", "priest", "deacon",
    # 군사
    "general", "colonel", "captain", "major", "lieutenant",
    "sergeant", "commodore", "admiral",
    # 정치/행정
    "governor", "president", "senator", "congressman", "representative",
    "secretary", "commissioner", "judge", "justice", "magistrate",
    "deputy", "speaker", "ambassador", "agent",
    # 학문/전문직
    "doctor", "dr.", "professor", "m.d.", "esquire", "esq.",
    "physician", "chemist", "philosopher",
    # 귀족/사회적 호칭
    "sir", "lord", "lady", "duke", "duchess", "earl", "baron",
    "baroness", "marquis", "count", "prince", "princess",
    "m.", "mr.", "mrs.", "miss", "ms.",
    # 직업
    "printer", "publisher", "merchant", "tradesman", "apprentice",
    "postmaster", "surveyor", "attorney", "lawyer",
}
```

Fig. 6. Dictionary for Application to Other Memoir Data



Autobiography		Unknown	7	The Autobiography is Franklin's longest work, and yet it is only a fragment. || [117]"Here terminates the Autobiography, as					
Richard		M	7	For the maxims of Poor Richard, see pages 331-335. || How much more than is necessary do we spend in sleep, forgetting t					
Jeffrey		M	1	Although he lived in a century notable for the rapid evolution of scientific and political thought and activity, yet					Boston
Edison		M	1	He was the Edison of his day, turning his scientific discoveries to the benefit of his fellow-men.					Boston
Carlyle		M	1	Carlyle called him the father of all the Yankees.					Boston
Robert Louis Stevens		M	1	If Robert Louis Stevenson is right in believing that his remarkable style was acquired by imitation then the youth					Boston
Humphry Davy		M	1	Sir Humphry Davy, the celebrated English chemist, himself an excellent literary critic as well as a great scientist, sa					Boston
Cotton Mather's		Unknown	1	Before the Autobiography only one literary work of importance had been produced in this country—Cotton Mather's Magr					
Franklin		M	5	What follows was written In the last year of Dr. Franklin's life, and was never before printed in English. || "No matter," says F					
Speech		Unknown	1	The Autobiography, Poor Richard, Father Abraham's Speech or The Way to Wealth, as well as some of t	Bagatelles				Boston
Samuel Johnson		M	3	It is said that the best parts of Boswell's famous biography of Samuel Johnson are those parts where Boswell permits Johns					
Tom Jones		Unknown	1	This century saw the beginnings of the modern novel, in Fielding's Tom Jones, Richardson's Clarissa Harlowe, Sten					Boston
Richardson		Unknown	2	[26]Samuel Richardson, the father of the English novel, wrote Pamela, Clarissa Harlowe, and the History of Sir Charles Gran					
Sterne		Unknown	1	This century saw the beginnings of the modern novel, in Fielding's Tom Jones, Richardson's Clarissa Harlowe, Sten					Boston
Tristram Shandy		Unknown	1	This century saw the beginnings of the modern novel, in Fielding's Tom Jones, Richardson's Clarissa Harlowe, Sten					Boston
Goldsmith		Unknown	1	This century saw the beginnings of the modern novel, in Fielding's Tom Jones, Richardson's Clarissa Harlowe, Sten					Boston
Gibbon		M	3	Gibbon and Hume, the great British historians, who were contemporaries of Franklin, express in their autobiographies the s					
Adam Smith		M	1	Gibbon wrote The Decline and Fall of the Roman Empire, Hume his History of England, and Adam Smi	England				Boston
Bunyan		M	1	In his numerous parables, moral allegories, and apologues he showed Bunyan's influence.					Boston
Defoe		M	1	Defoe and his contemporaries were authors.					Boston
William Franklin		M	2	The first part, written as a letter to his son, William Franklin, was not intended for publication; and the composition is more					
William Temple Fran		M	2	It was left by Franklin with his other works to his grandson, William Temple Franklin, whom Franklin designated as his literz					
Shipley	Bishop	M	1	Franklin began the story of his life while on a visit to his friend, Bishop Shipley, at Twyford, in Hampshi	Twyford, H				Boston
Abel James		M	2	Twenty-three pages of closely written manuscript fell into the hands of Abel James, an old friend, who sent a copy to Frank					
Temple Franklin		M	2	Temple Franklin and his successors. || When Temple Franklin came to publish his grandfather's works in 1817, he sent the c					
John Bigelow		Unknown	2	The other standard edition is the Works of Benjamin Franklin by John Bigelow (New York, 1887). || The original manuscript					
E. Dwight Church		M	1	By him it was later sold to Mr. E. Dwight Church of New York, and passed with the rest of Mr. Church's New York	England				
Church		M	1	By him it was later sold to Mr. E. Dwight Church of New York, and passed with the rest of Mr. Church's New York	England				
Henry E. Huntington		M	1	By him it was later sold to Mr. E. Dwight Church of New York, and passed with the rest of Mr. Church's New York	England				
Huntington		Unknown	1	The original manuscript of Franklin's Autobiography now rests in the vault in Mr. Huntington's residenc	New York	England			

Fig. 7. Processed Data (Benjamin Franklin)

위 과정을 통하여 본 논문에서 제시한 규칙 기반의 알고리즘이 카사노바의 회고록이라는 특정 데이터 뿐만 아니라, 회고록이라는 데이터 분류에 대부분 적용할 수 있는 확장 가능성을 검증하였다.

VI. Conclusions

본 연구는 18세기 유럽 사회의 방대한 기록물인 자코모 카사노바의 회고록을 대상으로 디지털 인문학적 방법론을 적용하여 대규모 인물 네트워크 및 지리적 이동 데이터를 구축하는 전 과정을 체계화하였다. 구텐베르크 프로젝트의 텍스트를 기반으로 Stanza, spaCy, NRCLex 등 현대적 자연어 처리 기술과 생성형 AI를 결합한 데이터 정제 파이프라인을 설계하였으며, 이를 통해 총 1,924명의 고유 인물을 정교하게 추출했다. 데이터 분석 결과, 카사노바의 서사는 단순한 에피소드의 나열이 아닌 특정 시점마다 대규모 신규 유입으로 인물이 확장되다가 핵심 인물 중심으로 압축되는 반복적 패턴을 지님을 확인하였다. Henriette, Esther, Bragadin 등 주요 인물과의 관계가 언급 횟수와 감정 데이터를 통해 정량적으로 입증되었으며, 이는 18세기 사교계의 계층 이동과 권력 구조를 입체적으로 재구성하는 기초 자료가 된다. 본 연구의 특성은 다국어와 복합적인 호칭 체계라는 18세기 자전적 텍스트의 한계를 극복하기 위해 '반자동(Automated + Manual) 정제 방식'과 '규칙 기반 알고리즘'을 제시했다는 점이다. 기존의 BERT 기반 NLP 모델이 18세기 특유의 서술 양식에서 보이는 지역 이동에서 한계를 보인 것과 달리, 본 연구에서 제안한 규칙 기반 알고리즘은 지리적 이동 데이터 구축에서 더 높은 정확도를 나타냈다. 또한, 생성형 AI를 중복 인물 통합 및 성별 추정의 보조 수단으로 활용하고 연구자가 이를 최종 검증하는 루프를 형성함으로써 데이터의 신뢰성을 확보하였다. 본 연구에서 제시한 분석 프레임워크는 카사노바 회고록을 넘어 벤자민 프랭클린 등 동시대의 다른 회고록 텍스트로의 확장 가능성을 보여주었고, 지역분류에서 강점을 보이는 것을 검증하였다.

그러나 본 연구는 몇 가지 방법론적 한계점을 지니며 이는 향후 연구를 통해 보완되어야 한다. 첫째, 현재 연구에서 사용한 규칙 기반 방식은 동일 문장 내 공출현(Co-occurrence) 분석에 의존하므로, 문장이 길고 수식어가 많은 역사적 사료의 특성상 단계가 누적될수록 미세한 개체 인식 오류가 발생할 가능성이 있다. 이를 극복하기 위해 가우시안 분포를 활용하여 개체 주변의 핵심 문맥에 높은 가중치를 부여하고 노이즈를 필터링하는 최신 SOTA 모델인 '가우시안 기반 국소 강화 공동 추출 알고리즘(Gaussian-based local reinforcement joint extraction algorithm)'의 도입을 검토할 필요가 있다.

둘째, 감성 분석에 활용된 NRCLex는 사전 기반(Lexicon-based) 모델로, 번역의 한계뿐만 아니라 18세기 문학이 가진 고유한 정서적 뉘앙스와 시대에 따른 단어 의미 변화를 온전히 포착하기 어렵다. 특히 인물 간 관계를 단순 연결성이 아닌 심리적 역동으로 파악하기 위해서는 특정 '인물'을 향한 감정을 명확히 분리해 내야 한다. 따라서 후속 연구에서는 타겟 기반 감성 분석(Target-based Sentiment Analysis, TBSA) 모델을 적용하여, 빈도 위주의 정형 데이터를 심층적인 질적 데이터로 고도화할 계획이다.

나아가 본 연구에서 구축한 데이터셋은 향후 D3.js[21] 기반의 인터랙티브 시각화[22,23]로 확장되어, 카사노바의 유럽 이동 경로와 인물관의 시간적 변화를 직관적으로 탐색할 수 있는 환경을 제공할 것이다.

Appendix. Used Prompts

본 연구는 텍스트 내 인물(entity) 추출 및 정제를 위해 LLM 기반 분석을 수행하였으며, 재현성과 일관성을 확보하기 위해 프롬프트 설계, 파라미터 설정, 재시도 규칙을 명시적으로 구성하였다.

분석 과정은 다음 세 단계로 이루어진다.

1. 이름 없는 인물(P.list) 분류
2. 성별 추정
3. 중복 인물 통합

Table 9. Generative AI Parameters

Parameter	Parameter value
Model	GPT-5.3, Gemini Pro 3.1 Pro
Temperature	0.2
top_p	0.9
max_tokens	≥ 2000
Output format	structured format

Prompt 1. 이름 없는 인물(P.list) 분류
다음 표현들을 두 가지로 분류
1. 고유 인물: 특정 개인을 지칭하는 표현
2. 일반 명사: 직업, 역할, 집단, 불특정 인물을 나타내는 표현
판단 기준:
- 이름이 없고 역할만 나타내면 일반 명사로 분류
- 문맥상 특정 개인을 지칭하면 고유 인물로 분류
- 관계 표현(my father 등)은 기본적으로 일반 명사로 분류하되, 필요시 "문맥상 특정 인물"로 표시
출력 형식:
표현 / 분류 결과 / 판단 이유

Prompt 2. 성별 추정 및 보정
다음 인물의 이름, 직위, 호칭을 기반으로 성별을 추정
판단 기준:
- Madame, Dona, Countess 등 여성호칭 → 여성
- Monsieur, Don, Abbé 등 남성호칭 → 남성
- 이름 기반 문화적 관습을 고려한다.
- 문맥 정보를 보조적으로 활용한다.
- 판단이 어려운 경우 unknown으로 표시.
출력 형식: 인물명 / 직위 또는 호칭 / 추정 성별 / 판단 근거

Prompt 3-1. 중복 인물 통합
다음 이름들이 동일 인물인지 판단하고, 동일 인물일 가능성이 있는 항목들을 그룹화
판단 기준:
- 이름 일부 일치
- 성/이름/미들네임 변형
- 직위 포함 여부 차이
- 언어별 표현 차이
- 예명, 필명, 즉위명
- 철자 및 번역 차이
출력 형식:
동일 인물 후보 그룹 / 대표명 / 통합 근거

Table 10. Criteria for Merging Duplicate Persons

Criteria	Threshold
Same person	≥ 0.85
Candidate pool	0.65 ~ 0.85
Mismatch	< 0.65

위 프롬프트 실행 과정에서 오류와 출력 변동성을 최소화하기 위하여 아래의 규칙을 적용하였다.

재시도 규칙
다음 조건 발생 시 재시도 수행
-출력 형식 오류
-동일 항목 분류 불일치
-unkown 비율 20% 초과
-성별 충돌
-낮은 신뢰도 (confidence < 0.6
-중복 인물 과분리

2. 재시도 단계
2-1) 동일 프롬프트 재실행
2-2) 규칙 강조("이전 결과에서 오류가 발생하여 판단 기준을 엄격하게 적용할 것.")
2-3) 문맥 확장
±2~3 문장 추가
2-4) 규칙 기반 후처리
동일 표현 -> 동일 결과 강제
관계 표현 -> 고유 인물 매핑
직위 제거 후 비교
2-5) 보수적 처리

Table 11. Failure Handling Policy

Case	Handling Method
Gender identification unavailable	Retain 'unknown' status
Unclear if the same person	Keep separated without merging
Unclear if entity is a person	Handle as a generic noun
Insufficient context	Reflect only current information

3. 반복 제한
최대 5회 반복
분석 결과 간 불일치가 발생할 경우 다음의 우선순위 규칙을 적용한다.

Table 12. Classification Mismatch

Priority	Criteria
Primary priority	Inclusion of proper names
Secondary priority	Presence of Role/Occupation
Tertiary priority	Contextual specificity

이름이 없는 경우 일반명사로 통일한다.

Table 13. Gender Mismatch

Priority	Criteria
Primary priority	Titles (e.g., Don, Dona, Abbé)
Secondary priority	Gender inferred from name
Tertiary priority	Contextual information

충돌 시 호칭 기준으로 덮어쓴다

Table 14. Inconsistency in Merging Identical Persons

Priority	Criteria
Primary priority	Presence of aliases or regnal names
Secondary priority	Name similarity ($\geq$ 0.85)
Tertiary priority	Contextual redundancy

불확실한 경우 병합하지 않는다.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2022S1A5C2A02090368)

REFERENCES

[1] G. Casanova, "The Memoirs of Jacques Casanova de Seingalt, 1725-1798," Project Gutenberg, https://www.gutenberg.org/ebooks/2981

[2] T. Nevalainen, "Society and culture in the long 18th century," Patterns of change in 18th-century English, pp. 13-23, 2018.

[3] L. Damrosch, "Adventurer: the life and times of Giacomo Casanova," Yale University Press, 2022.

[4] H. Kim, W. Jang, H. Huh, J. Park, W. Moon, S. Lee, and J. Yoon, "An Efficient Approach for Constructing Intelligent Naval Combat System Dataset based on Combat Structural Data Cleansing Techniques," Journal of the Korea Society of Computer and Information, Vol. 30, No. 12, pp. 87-100, 2025. DOI: 10.9708/jksci.2025.30.12.087

[5] Stanford NLP Group, "Stanza: A Python Natural Language Processing Toolkit," Stanford University, https://stanfordnlp.github.io/stanza/

[6] Explosion AI, "spaCy: Industrial-Strength Natural Language Processing in Python," https://spacy.io/

[7] NRCLex, "NRCLex: Natural Language Processing Meets Emotion Analysis," PyPI, https://pypi.org/project/NRCLex/

[8] B. Kim, and W. Kang, "Big Data and Classical Literature Research Methods," The Journal of Language & Literature, Vol. 78, pp. 7-39, 2019. DOI: 10.15565/jll.2019.06.78.7

[9] H. Cho, "Multi-layer Network Analysis of Imjin War's Historical materials," Yonsei University, 2022.

[10] S. Gyujin, "Using Topic Modeling to Explore Trends in Christian Worldview Research," Faith and Scholarship, Vol. 30, No. 2, pp. 85-107, 2025.

[11] K. Jihoon, "A Study of Atypical Data Analysis Based on Text Mining - Focused on writing pattern analysis," Culture and Convergence, Vol. 42, No. 8, pp. 373-391, 2020.

[12] J. Lee, "A Study on Modeling Metadata and Developing Standard Elements to Establish Oral History Archives." Journal of the Korean Society for Information Management, Vol. 26, No. 1, pp. 163-184, 2009.

[13] L. Zilio, R. R. Lazzari, and M. J. B. Finatto, "NLP for historical Portuguese: Analysing 18th-century medical texts," Proceedings of the 16th International Conference on the Computational Processing of Portuguese (PROPOR 2024), pp. 293-303, Santiago de Compostela, Spain, March 2024.

[14] Y. Kim, and M. Song, "A Study on Analyzing Sentiments on Movie Reviews by Multi-Level Sentiment Classifier," Journal of Intelligence and Information Systems, Vol. 22, No. 3, pp. 71-89, 2016.

[15] H. Ha, H. Han, S. Mun, S. Bae, J. Lee, and K. Lee, "Visualization of movie recommendation system using the sentimental vocabulary distribution map," Journal of the Korea Society of Computer and Information, Vol. 21, No. 5, pp. 19-29, 2016.

[16] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Y. Hawalah, A. Gelbukh, and Q. Zhou, "Multilingual sentiment analysis: state of the art and independent comparison of techniques," Cognitive computation, Vol. 8, No. 4, pp. 757-771, 2016.

[17] M. Lee, and H. Kim, "Construction of Event Networks from Large News Data Using Text Mining Techniques," Journal of Intelligence and Information Systems, Vol. 24, No. 1, pp. 183-203, 2018.

[18] K. H. Lee, J. H. Lee, M. S. Choi, and G. C. Kim, Study on named entity recognition in Korean text, In Annual Conference

on Human and Language Technology. pp. 292-299, Seoul, Korea, October, 2000.

[19] K. Park, S. Na, J. Shin, and Y. Kim, BERT for Korean Natural Language Processing: Named Entity Tagging, Sentiment Analysis, Dependency Parsing and Semantic Role Labeling, Proceedings of the Korea Computer Congress, pp. 2019. Park, K., Na, S., Shin, J., & Kim, Y. (2019-06-26). 84-586, Jeju, Korea, June, 2019.

[20] B. Franklin, "Autobiography of Benjamin Franklin," Project Gutenberg, https://www.gutenberg.org/files/20203/20203-h/20203-h.html

[21] M. Bostock, V. Ogievetsky, and J. Heer, "D3: Data-Driven Documents," IEEE Trans. Visualization & Computer Graphics, Vol. 17, No. 12, pp. 2301-2309, 2011.

[22] D. Edelstein, P. Findlen, G. Ceserani, C. Winterer, and N. Coleman, "Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project," American Historical Review, Vol. 122, No. 2, pp. 400–424, 2017.

[23] Project Maps, "Visualizations of 19th Century Travel Data," HIST 1952.

Authors

Sunghoon Jeong is currently pursuing a B.A. in Department of History and a minor in Data Humanities at Ajou University. Sunghoon Jeong joined Ajou University, Korea, in 2020.

Sunghoon Jeong is interested in systematizing unstructured historical records, data analysis for primary sources, and computational frameworks for historiography.

Yujin Noh is currently pursuing a B.A. in Department of Culture and Contents and a double major in Data Humanities at Ajou University. Yujin Noh joined Ajou University, Korea, in 2021.

Yujin Noh is interested in AI-based media data analysis, digital communication, and the interaction between agentic AI and digital media.

Jeongeun Hwang is currently pursuing a B.A. in Department of French Language and Culture and a double major in Data Humanities at Ajou University. Jeongeun Hwang joined Ajou University, Korea, in 2023.

Jeongeun Hwang is interested in data-driven approaches to language, media, and digital humanities.

Jinsun Kim is currently pursuing a B.A. in Department French Language and Literature and a minor in Data Humanities at Ajou University. Jinsun Kim joined Ajou University, Korea, in 2024.

Jinsun Kim is interested in digital humanities, entity extraction from literary texts, and relationship mapping.

Hajin Kim is currently pursuing a B.A. in Department of French Language and Literature and a double major in International Trade at Ajou University. Hajin Kim joined Ajou University, Korea, in 2025.

Hajin Kim is interested in digital humanities, text analysis, and the application of data-driven methodologies to cultural and literary studies.

Hyoji Ha received the B.S. and Ph.D. degrees in Media from Ajou University, Korea, in 2013 and 2023, respectively. Dr. Ha joined the Humanities Research Institute at Ajou University, Korea, in 2025.

Dr. Ha is currently a Research Assistant Professor at the Humanities Research Institute, Ajou University. Dr. Ha is interested in AI-based digital humanities, data visualization, and media contents.