Data Paper https://doi.org/10.23287/KJDH.2024.1.2.4

이상 단편소설 기초·감각 데이터셋 *

Yi Sang Short Story Basic·Sense Dataset

지해인(Ji, Haein)**
0000-0001-6864-6789

목차

초록

본 논문은 이상(李箱) 단편소설 기초 데이터셋과 이상 단편소설 감각 데이터셋의 설계·구축 과정을 소상히 소개하는 데에 그 목적이 있다. 이상 단편소설 13 편을 대상으로 구축한 '이상 단편소설 기초 문학 데이터셋'과 연구자 주도로 텍스트 내 감각 정보를 레이블링한 '이상 단편소설 감각 데이터셋'을 중심으로, 데이터셋의 구조와 설계 의도 및 활용에 대해 서술하였다. 이상 단편소설 기초 데이터셋은 민음사와 소명출판 판본을 바탕으로 메타 데이터를 문장 단위로 레이블링한 기계 가독형 데이터로 구축되었다. 이상 단편소설 감각 데이터셋은 연구자가 설계한 감각 분류 모델에 기초하여 이상 단편소설에 나타난 감각 정보를 크게 신체 감각과 심리 감각으로 대별하고, 감각을 도합 4 계층으로 세분화하여 문장 단위로 레이블링하였다. 구축한 데이터셋은 이상 단편소설 내 감각 양상에 대한 기계적 분석, 감정 분석 등 여타 분석 방법론을 수행하기 위한 실질적 기반이 되며, 나아가 멀리서 읽기의 가능성을 제공한다.

주제어: 이상(李箱), 이상 단편소설 데이터셋, 문학 데이터, 데이터 리뷰, 디지털인문학

*본 논문은 "디지털 감각·감정 분석을 통한 이상 문학의 에피파니 연구"(한국학중앙연구원 한국학대학원 석사학위논문, 2024)에서 구축한 일련의 데이터셋을 대상으로, 내용 일부를 데이터 논문의 목적과 내용에 적합하게 수정·보완하여 작성된 것이다. 또한 "디지털 감각·감정 분석을 통한 이상 문학의 에피파니 연구"(한국학중앙연구원 한국학대학원 석사학위논문, 2024)를 요약·수정한 논문이 '추천석사논문'으로 다음과 같이 게재되었음을 분명히 밝힌다. 지해인(2024). "디지털 감각·감정 분석을 통한 이상 문학의 에피파니 연구". <상허학보>. 753-827. https://doi.org/10.22936/sh.72..202410.021

**단독저자: 한국학중앙연구원 디지털인문학연구소, 인문정보학·인문지리학(인문정보학), 조연구원, cihayin@gmail.com

Abstract

This paper aims to thoroughly introduce the design and construction processes of two datasets related to Yi Sang's short stories: the Yi Sang Short Story Basic Dataset and the Yi Sang Short Story Sense Dataset. Centered on 13 selected short stories, the Yi Sang Short Story Basic Dataset presents a machine-readable structure created through the annotation of meta-data at the sentence level, based on editions from Mineumsa and Somyeong Publishing. The Yi Sang Short Story Sense Dataset, constructed by the researcher, labels sensory information found within the texts, using a sensory classification model that categorizes perceptions broadly into physical and psychological senses. This model further subdivides sensory details into four hierarchical levels, enabling nuanced, sentence-level labeling. The constructed datasets serve as practical foundations for conducting computational analyses of sensory patterns in Yi Sang's short stories, as well as for other analytical methodologies such as emotion analysis, and further provide the potential for distant reading.
Keywords: Yi Sang, Yi Sang Short Stories Dataset, Literary Data, Data Review, Digital Humanities

1. 서론

본 논문은 연구자가 이상(李箱) 단편소설 13편을 대상 텍스트로 설계·구축한 기초 문학 데이터셋(Literary dataset)과 이상 단편소설 텍스트 내 감각(Sense) 요소를 기계적으로 식별·처리하기 위해 설계·구축한 감각 레이블 데이터(Labeled data)를 소개하는 데 그 목적이 있다. 먼저 데이터셋의 구조와 설계 의도 및 전략 등 설계·구축상의 제 내용을 개괄하고, 그 한계를 논하고자 한다. 데이터 논문 특성상, 본 논문에서 소개한 데이터셋의 구체적인 활용 방안은 해당 데이터셋을 활용한 지해인(2024)[1]의 사례로 갈음하도록 한다.

2. 이상 단편소설 기초·감각 데이터셋

2.1 데이터셋 설명

제공 플랫폼 : OSF(Open Science Framework)
프로젝트 명칭 : 디지털 감각·감정 분석을 통한 이상 문학의 에피파니 연구
프로젝트 링크 : https://osf.io/964nc
프로젝트 파일 위계 :

OSF Storage (United States)
- ├──data
  - │ ├──Rawdata_이상단편소설_기초데이터.tsv : https://osf.io/a69ec
  - │ ├──Rawdata_이상단편소설_기초데이터+ 감각데이터.xlsx : https://osf.io/sjxtn
  - │ ├──Rawdata_타연구자에피파니_데이터.tsv : https://osf.io/3xmga
  - │ └──Rawdata_이상단편소설_감각데이터.tsv : https://osf.io/5st4r
- ├──model
  - │ └──kote_pytorch_lightning.bin : https://osf.io/rdz2j
- ├──process
  - │ ├──Analysis_이상단편소설_감각분석.ipynb : https://osf.io/p6y7c
  - │ ├──Analysis_이상단편소설_감정분석.ipynb : https://osf.io/ca64z
  - │ ├──Analysis_이상단편소설_상관성분석.ipynb : https://osf.io/7f9uw

```
|   └──Analysis_타연구자에피파니_감정분석.ipynb : https://osf.io/4vpe5
└──results
    ├──Results_이상단편소설_감정분석.tsv : https://osf.io/p3bt7
    ├──Results_이상단편소설_감정분석.xlsx : https://osf.io/pq938
    ├──Results_타연구자에피파니_감정분석.tsv : https://osf.io/nsq56
    └──Results_타연구자에피파니_감정분석.xlsx : https://osf.io/eyw98
```

생성일 : 2024-06-28

데이터셋 제작자 : 지해인(Ji, Haein)

이상 단편소설 기초·감각 데이터 파일은 프로젝트 저장소의 'data' 디렉토리에서 확인할 수 있다. 기본적으로 tsv 형식으로 제공하고 있으나, Excel 분석 도구를 활용하는 경우의 편의성 증진을 위해 xlsx 형식으로도 제공하고 있다.

그 외 분석 모델 학습에 필요한 가중치(Weight) 및 파라미터(Parameter) 정보를 담고 있는 바이너리 파일(Binary File)이 'model' 디렉토리에, 분석에 활용한 소스 코드(Source code)가 저장된 주피터 노트북(Jupyter notebook) 파일이 'process' 디렉토리에, 분석을 수행한 결과 파일이 'results' 디렉토리에 저장되어 있으나, 본 논문은 설계·구축된 '이상 단편소설 기초 데이터셋', '이상 단편소설 감각 데이터셋'에 주안점을 두고 있으므로, 여기서는 그에 대해 심도 있게 논하지 않기로 한다. 대신 '이상 단편소설 기초 데이터셋'과 '이상 단편소설 감각 데이터셋'의 설계와 구조를 논하면서 'Analysis_이상단편소설_감정분석.ipynb'과 'Analysis_이상단편소설_감각분석'에 포함된, 데이터셋을 불러오는 일부 코드를 각주를 통해 소개하는 것으로 데이터 논문으로서 본 논문의 활용성 증진을 도모하고자 한다.

2.2 이상 단편소설 기초 데이터셋 설계와 구조

이상 단편소설 기초 데이터셋은 OSF를 통해 'Rawdata_이상단편소설_기초데이터.tsv'라는 명칭으로 제공되며, 이상 단편소설과 관련된 기초적인 데이터와 메타 데이터를 '제1정규형'(First normal form; 1NF)을 만족하는 기계 가독형 데이터(Machine-readable data)로 구축한 데이터셋이다. 데이터셋에 포함된 이상 단편소설은 총 13편으로, 다음과 같다.

표 1. 분석 대상 작품의 기초 정보

작품	발표연도	잡지	문단 수	글자 수
「地圖의暗室」	1932.3	『朝鮮』	43	10,198
「休業과事情」	1932.4	『朝鮮』	24	10,764
「집팽이轢死」	1934.8	『月刊每申』	7	6,707
「鼈龜會豕」	1936.6	『中央』	26	14,028
「날개」	1936.9	『朝光』	136	19,861

「逢別記」	1936.12	『女性』	90	4,573
「童骸」	1937.2	『朝光』	271	15,713
「恐怖의記錄」	1937.4	『每日申報』	109	9,074
「終生記」	1937.5	『朝光』	201	16,016
「幻視記」	1938.6	『青色紙』	74	5,155
「失花」	1939.3	『文章』	142	8,365
「斷髮」	1939.4	『朝鮮文學』	69	6,147
「金裕貞」	1939.5	『青色紙』	51	4,339

이상의 첫 소설이자 유일한 중편소설인 「十二月十二日」(1930)은 단편소설이 아니므로 구축 대상에서 우선 제외하였다. 「황소와독개비」(1937)는 장르상 동화일뿐더러, 이상의 저작인지에 대한 이견[2]이 있어 구축 대상에서 제외하였다. 「恐怖의記錄」(1937)의 경우 수필로 분류되는 경우도 있으나 이상의 소설과 수필은 그 장르적 구분이 명확하지 않은 경우가 빈번하고, 본 논문에서 소개한 데이터셋이 저본으로 삼은 <정본 이상 문학전집>에서는 「恐怖의記錄」을 소설로 분류한 점을 토대로 단편소설로 간주, 구축 대상으로 포괄하였다.

다음은 표 1에서 선정한 이상 단편소설 13편을 대상으로 구축한 이상 단편소설 기초 데이터셋의 샘플이다.

표 2. 이상 단편소설 기초 데이터셋 샘플

ID	작품명	저자	필명	발표지	발표년월	발표시기	장ID	문단ID	문장ID	시점	서술방식	텍스트_민음사	텍스트_소명출판
1	날개	이상	李箱	『朝光』	1936.9	생전		1	1	1인칭	서술	박제가 되어 버린 천재'를 아시오?	『剝製가되어버린 天才』를 아시오?
639	休業과事情	이상	甫山	『朝鮮』	1932.4	생전		1	1	1인칭	서술	삼 년 전이 보산과 SS와 두 사람 사이에 끼어 들어앉아 있었다.	삼년전이보산과SS와 두사람사이에 끼워들어안저잇섯다
829	집팽이轢死	이상	李箱	『月刊毎申』	1934.8	생전		2	7	1인칭	서술	반찬이 열 가지나 되는데 풋고추로 만든 것이 다섯 가지—내 마음에 꼭 들었습니다.	반찬이 열가지나되는데 풋고초로만든것이 다섯가지—내마음에 꼭들엇슴니다
1455	龜龜會豕	이상	李箱	『中央』	1936.6	생전	2	25	589	1인칭	서술	거미내음새는 — 그러나이십원을요 모조모주무르던그 새금한지폐내음새 가참그윽할뿐이었다.	거미내음새는—그러나二十원을요모조 모금물르든그새금 한지페내음새가참 그윽할뿐이었다
1477	逢別記	이상	李箱	『女性』	1936.12	생전	1	1	1	1인칭	서술	스물세 살이오.—	스믈세살이오—
1659	終生記	이상	李箱	『朝光』	1937.5	사후		1	1	1인칭	서술	극유산호(郤遺珊瑚) — 요 다섯 자 동안에 나는 두 자 이상의 오자를 범했는가 싶다.	郤遺珊瑚—요 다섯字동안에 나는 두字以上의 誤字를 犯했는가싶다.
2118	斷髮	이상	李箱	『朝鮮文學』	1939.4	사후		7	16	1인칭	대화	"세상 사람들이 모두 연(衍) 씨를 욕허니까 어디 제가 고쳐 디리지요.	「세상사람들이 모두 衍氏를 욕허니까 어디 제가 고처디리지요.

이상 단편소설 기초 데이터셋은 표 2의 샘플과 같이, ‘ID’, ‘작품명’, ‘저자’, ‘필명’, ‘발표지’, ‘발표년월’, ‘발표시기’, ‘장ID’, ‘문단ID’, ‘문장ID’, ‘시점’, ‘서술방식’, ‘텍스트_민음사’, ‘텍스트_소명출판’의 도합 14열로 구성되어 있다. 행을 문장 단위로 입력하되, 딥러닝 감정 분석 적용을 위해 부득이 한 문장을 여러 행으로 나눠서 입력한 경우가 3건[3] 존재한다.[4]

각각의 열(Column)에 대한 설명은 다음과 같다.

- ID : 전체 단편소설 13편을 문장 단위로 나누어 개별 문장에 부여한 고윳값을 기입한 열이다.
- 작품명 : 작품의 제목으로, <정본 이상문학전집>에 실린 표기를 따랐다.
- 저자 : 이상 단편소설 데이터셋이므로, 모두 이상으로 기입되었다.
- 필명 : 발표 당시 사용한 필명을 기입하였다.
- 발표지 : 수록한 잡지의 명칭을 기입한 열이다.
- 발표년월 : 해당 작품이 발표된 연월을 기입한 열이다.
- 발표시기 : 해당 작품이 생전에 발표되었는지, 사후에 발표되었는지를 기입한 열로, ‘생전’과 ‘사후’ 중 하나의 값을 갖는다.
- 장ID : 작품이 장 단위로 구분되어 있는 경우, 장 정보를 기입한 열이다.
- 문단ID : 개별 작품 속 문단에 대해 순차적으로 고윳값을 부여하여 기입한 열이다. ‘ID’ 열과 달리, 작품별로 초기화된다. 즉, 모든 작품의 ‘문단ID’는 1로 시작한다.
- 문장ID : 개별 작품 속 문장에 대해 순차적으로 고윳값을 부여하여 기입한 열이다. ‘문단ID’와 마찬가지로, 개별 작품의 ‘문장ID’는 모두 1로 시작한다.
- 시점 : 시점 정보를 입력한 열로, 시점에 따라 ‘1인칭’, ‘3인칭’의 값을 갖는다.
- 서술방식 : 서술 방식을 입력한 열로, ‘서술’, ‘대화’
- 텍스트_민음사 : 민음사의 권영민 편, <이상 소설 전집>을 데이터셋 제작자가 직접 읽어 나가면서 문장 단위로 입력한 열이다.
- 텍스트_소명출판 : 소명출판의 김주현 편, <정본 이상문학전집>을 제작자가 직접 읽어 나가면서 문장 단위로 입력한 열이다.

특별히 이론이 없는 ‘작품명’, ‘저자’, ‘필명’, ‘발표지’, ‘발표년월’, ‘발표시기’ 등의 기본 정보와 채택 판본에 근거해 순차적으로 고윳값을 부여한 ‘장ID’, ‘문단ID’와 달리, 문장 부호의 생략 등으로 인해 그 종결 및 구획이 불분명한 경우가 비일비재한 이상 텍스트 특성상 ‘문장ID’, ‘시점’과 같은 값은 연구자의 주관이 일부 반영되었다. 또한 ‘서술 방식’의 경우 최대한 명시적인 기준을 설정하고자 큰따옴표와 같은 문장 부호의 유무를 중요하게 보았으므로 이에 대해 다소간 이론의 여지가 있을 수 있다.

민음사 판본의 텍스트와 소명출판 판본의 텍스트를 함께 기입한 이유는 다음과 같다. 먼저 민음사 판본은 어휘나 정서법 등에 있어 비교적 현대 한국어에 가깝게 다듬어져 있으므로, 현대 한국어를 기반으로 학습시킨 딥러닝 모델 등을 적용시키는 데 이점이 있다. 이는 현대 독자의 이상 텍스트 수용과 관련한 연구에도 유용할 가능성이 있다. 그 구체적인 실례로, 한국어 온라인 댓글 감정 데이터셋인 KOTE(Korean Online That-gul Emotions) 데이터셋[5]과 BERT(Bidirectional Encoder Representations from Transformers)[6] 계열의 언어 모델인 KcELECTRA(Korean comments ELECTRA)[7]를 활용하여 이상 단편소설에 대한 감정 분석을 수행하고, 그 함의를 해석한

지해인(2024)[8]을 참고할 수 있다.

소명출판 판본의 텍스트는 편찬자가 최대한 원본 텍스트에 가깝게 다듬고 상세한 주해를 단 것으로, 근대 한국어 특유의 통일되지 않은 표기가 두드러지는 텍스트이다. 이러한 텍스트는 기계적 처리에 어려움이 있으나 '최대한 원본 텍스트에 가까운' 것이 이점이다. 해당 데이터를 활용한 연구는 아직 부재한 상태이나, 일단 데이터로 구축해 두었으므로 추후 서브워드 분리(Subword segmenation) 등의 방법론을 활용한 추가적인 양적 분석이 가능할 것으로 전망한다.

그 외, 본 데이터셋을 활용하는 연구자는 본인의 연구 질문에 따라 필요한 열을 취사하여 데이터를 재조직하고 적절하게 활용할 수 있다.[9]

2.3 이상 단편소설 감각 데이터셋 설계와 구조

구축된 이상 단편소설 기초 데이터를 바탕으로 연구자가 이상의 단편소설 13편을 '직접' 문장 단위로 톺아보며, 문장 내 감각어를 식별하고 그 품사와 감각 정보를 레이블링하였다. 이때 문장에서 감각 정보가 나타나는 양상을 적절히 분류하기 위해 최현배(1937)[10]에서 기술한 성상 형용사(性狀形容詞)의 감각 분류를 참고하여 분류 기준으로서 감각 분류 모델을 설계하였다. 다음은 최현배(1937)의 성상 형용사 감각 분류를 그대로 발췌한 것이다.

표 3. 최현배(1937), 성상 형용사 감각 분류

감각적	시각적	빛갈(色)		검다, 히다, 푸르다, 누르다, 붉다.
		빛(光)		밝다, 어둡다.
	미각적			달다, 쓰다, 시다, 떫다, 짜다, 맵다.
	청각적	소리(音)		시끄럽다, 고요하다.
		가락(調)		높다, 낮다.
	후각적			지리다, 비리다, 냅다.
	촉각적 (외피감각적)	누름(壓覺)		미끄럽다, 맨지럽다, 까끄럽다, 거칠다, 날카롭다, 둔하다, 무디다, 단단하다, 연하다, 굳다, 무르다, 무겁다.
		따뜻하기(溫度)		차다, 덥다, 뜨겁다, 춥다, 시원하다, 따뜻하다, 선선하다, 미지근하다.
		아픔(痛覺)		아프다, 따갑다, 쓰리다.
		그 나머지(其他)		가렵다, 간지럽다.
	평형감각			어지럽다.
	유기감각			답답하다, 아니꼽다, 뻐근하다, (오줌) 마렵다, 고프다, 부르다.
	시간공간감각 (複合感覺)	시간		빠르다, 더디다, 지리하다, 급하다, 눅다, 이르다, 늦다.
		공간	거리(距離)	멀다, 가깝다.
			물형(物形)	크다, 작다, 길다, 짜르다, 넓다, 좁다, 둥글다, 모나다, 바르다, 삐뚤다, 삐뚜룸하다, 곧다, 굽다.

			상하(上下)	높다, 낮다, 깊다, 얕다, 돋다, 뾰족하다, 움푹하다.

이상 단편소설 감각 데이터셋 구축을 위해 상기 표 3에 정리된 최현배(1937)의 성상 형용사 감각 분류 도식을 일차로 채택하고, 일부 준거에 대한 수정·보완을 거쳐 다음 표 4와 같이 도합 4개 계층으로 구조화된 감각 분류 모델을 정립하였다.

표 4. 감각 분류 모델 도식

1차 감각	2차 감각	3차 감각	4차 감각
신체/심리	외부	시각	
		청각	
		후각	
		미각	
		피부 감각	촉각
			온도각
			통각
	내부	평형	
		유기	

수정·보완한 부분은 다음과 같다. 먼저 최상위 계층으로 1차 감각을 설정하여 신체 감각과 심리 감각을 구분하였다. 전자는 육체적으로 지각하는, 일반적인 감각의 정의와 상통하며 후자는 신체적 감각이 아닌 심리적 감각을 지시한다. 다음 계층으로 외부/내부로 대별되는 2차 감각을 설정하였다. 이는 감각의 방향성, 즉 외부 세계로 향하는지 신체 내부로 향하는지에 따라 감각 정보를 구분해 준 것이다. 다음 계층인 3차 감각은 기본적으로 오감 등 최현배(1937)에서 제시한 개별 감각을 구분한 것이나, '피부 감각'에 한해 4차 감각의 계층을 도입해 보다 세밀하게 분별하였다. 이는 개인이 피부로 느끼는 '촉각', '온도각', '통각'과 같은 감각들이 질적으로 상이한 경험으로 체감될 수 있다고 판단하였기 때문이다.

최현배(1937)에서 외피 감각과 촉각을 동의로 기술한 반면, 이상 단편소설 감각 데이터셋에서 채택한 감각 분류 모델에서는 3차 감각으로 '피부 감각'을 설정하였으므로, 최현배(1937)에서 '누름(壓覺)'으로 설정한 하위 범주는 '촉각'으로 표기하였다. 이때 '압각'의 경우, 중력으로 인해 신체 내부에서 발생하는 감각이므로 촉각 형용사의 갈래로 포괄하기 어렵다는 임윤정(2019)[11]의 논의가 있으나, 감각 분류 모델 설계자는 그 또한 외부 세계에 대한 감각의 한 양상이라고 파악하여 피부 감각의 하위 범주로 포함하였다. 또한 신체적 불편의 한 양상으로 '간지러움'을 파악한 임윤정(2019)[12]의 논의에 따라 '그 나머지(其他)'에 해당하는 '간지러움'의 감각을 통각으로 망라하였다.

또한 '내부 감각'의 경우, 최현배(1937)의 분류를 그대로 준용하였다. '내부 감각' 중 '유기 감각'의 경우, '어지러움'처럼 평형 기관을 통해 지각되는 평형 감각을 제외한, 신체 내부 감각 전반을 포괄하는 감각이다. 이는 '허기', '갈증', '욕지기', '변의'와 같은 소화계 및 비뇨계와 관련된 감

각뿐 아니라 심장의 박동 및 숨 가쁨 등 오장육부 전반으로 느낄 수 있는 감각과 '피로감'과 같이 신체 전체적으로 느끼는 감각을 포함한다.

한편 최현배(1937)에서 '빠르다', '멀다', '높다', '크다' 등의 어휘로 예시한 '시간공간감각'은 그 지각에 시각이 많이 동원된다는 점에서 '시각'으로 분류할 수도 있으나, 김창섭(1985)[13]에서 지적한바, "빛은 오직 시각에 의해서만 지각될 수 있고, 모양·크기·위치·속도는 시각 외의 감각기관에 의해서도 지각될 수 있"다는 점에서 보다 세심한 접근이 필요하다. 가령 최현배(1937) 또한 '시간공간감각'이 단일 감각 기관을 통해 지각되는 하나의 감각이 아닌, 복합적 감각의 결과물로 지각됨을 고려하여 '시간공간감각'을 '복합감각'(複合感覺)으로 분류한 바 있다. 이처럼 시간적·공간적으로 대상을 지각하는 데 활용되는 감각적 창구는 '시각' 외에도 다양하여 어느 하나의 감각으로 예속할 수 없다는 문제가 있다.

'시공간감각'을 '감각'으로 분류하는 것이 적절한지에 대한 이견 역시 성립할 수 있다. 최현배(1937)에서 제시한 공간 감각 중 '멀다', '가깝다' 등으로 예시되는 '거리' 범주, '크다', '작다' 등으로 예시되는 '물형' 범주, '높다', '낮다' 등으로 예시되는 '상하' 범주는 여타 감각, 그중에서도 '시각'을 매개로 감각한 대상에 대한 인지적 범주에 가까워 보인다. 이러한 관점에서 '시간공간감각'은 그 자체로 감각이라고 하기보다는, 개별 감각의 종합에 대한 인지적 해석의 산물로 보는 것이 적절하다고 할 수 있다. 상기 감각 분류 모델에서 '시간공간감각'을 별도의 독립된 감각으로 다루지 않은 것은, 이와 같이 시간적·공간적 개념을 감각 그 자체보다는 감각을 질료로 취합하여 해석해 낸 정신적 소산으로 판단하였기 때문이다.

이렇게 설계한 표 4의 감각 분류 모델을 바탕으로, 이상 단편소설 13편의 개별 문장에 대해 '감각어', '품사', '감각 정보'를 레이블링하였다. 이때 감각어로 지각 동사와 감각 명사, 감각 형용사를 식별하였다. 지각 동사(Verbs of Perception)란 감각 기관을 통해 자극을 수용하는 행위 동사로, '보다', '듣다', '맡다', '맛보다', '느끼다' 등의 동사가 이에 해당하며, 보다 폭넓은 의미의 '느끼다'의 경우 연구자가 문장을 읽고 직접 판단하여 감각 정보를 입력하였다. 감각어에 나타난 감각 정보를 식별하기 위해서는 표준국어대사전의 첫 번째 의미를 참고하였는데, 이는 전은진(2011)[14]과 표준국어대사전 편찬 지침에서 어휘의 첫 번째 의미를 중심 의미로 본 것을 따른 것이다. 다음은 참고한 표준국어대사전의 편찬 지침의 일부이다.

(1) 표준국어대사전(편찬 지침 다의어 뜻풀이 배열 4번)

다의어는 기본 의미를 우선 제시하고 그것에서 번진 뜻의 순으로 배열한다. 기본 의미와 맺는 관련 정도가 명확하지 않을 때에는 널리 사용되는 것을 우선 제시한다.[15]

또한 어휘의 첫 번째 의미가 "볶은 깨, 참기름 따위에서 나는 맛이나 냄새와 같다."인 '고소하다'처럼 둘 이상의 복합 감각을 명시한 경우, 연구자가 직접 문맥을 살펴 판단하였다. 이렇게 구축한 이상 단편소설 감각 데이터셋은 이상 단편소설 개별 문장에 나타난 감각의 양상을 연구자가 사전에 구축한 감각 분류 모델에 기반해 직접 태깅한 것으로, 제1정규형을 만족하는 기계 가독형 데이터라고 할 수 있다. 구축한 데이터셋은 OSF에 'Rawdata_이상단편소설_감각데이터.tsv' 파일로 공개하였다.

표 5. 이상 단편소설 감각 데이터셋 샘플

ID	텍스트_민음사	텍스트_소명출판	감각어	품사	감각
11	그런 생활 속에 한 발만 들여놓고 흡사 두 개의 태양처럼 마주 쳐다보면서 낄낄거리는 것이오.	그런生活속에 한발만 드려놓고 恰似두개의太陽처럼 마조처다보면서 낄낄거리는 것이오.	쳐다보다	동사	신체_외부_시각
46	33번지 18가구의 낮은 참 조용하다.	三十三번지 十八가구의 낮은 참 조용하다.	조용하다	형용사	신체_외부_청각
52	여러 가지 내음새가 나기 시작한다.	여러가지내음새가 나기시작한다.	냄새	명사	신체_외부_후각
192	밥은 너무 맛이 없었다.	밥은 너무 맛이없었다.	맛	명사	신체_외부_미각
198	하룻밤 사이에도 수십 차를 돌쳐 눕지 않고는 여기저기가 배겨서 나는 배겨 낼 수가 없었다.	하룻밤 사이에도 수십차를돌처눕지않고는 여기저기가백여서나는 백여내일수가없었다.	배기다	동사	신체_외부_피부_촉각
347	한 숟갈을 입에 떠 넣었을 때 그 촉감은 참 너무도 냉회와 같이 써늘하였다.	한수깔을 입에떠넣었을때 그 촉감은 참 너무도 냉회와같이 써늘하였다.	써늘하다	형용사	신체_외부_피부_온도각
1158	아프다.	아프다.	아프다	형용사	신체_외부_피부_통각
1214	취하기도취하였거니와이것은배가좀너무부르다.	취하기도취하었거니와이것은배가좀너무부르다.	취하다	동사	신체_내부_평형
1214	취하기도취하였거니와이것은배가좀너무부르다.	취하기도취하었거니와이것은배가좀너무부르다.	부르다	동사	신체_내부_유기
1317	앞이다캄캄하여지기전에 사부로가씨근씨근왔다.	앞이다캄캄하야지기전에 사부로가씨근씨근왔다.	캄캄하다	형용사	심리_외부_시각
1667	나는 일 개 교활한 옵서버의 자격으로 그런 우매한 성인(聖人)들의 생애를 방청하여 있으니 내가 그런 따위 실수를 알고도 재범(再犯)할 리가 없는 것이다.	나는 一個 狡猾한 옵서버— 의자격으로 그런 愚昧한 聖人들의 生涯를 傍聽하야있으니 내가 그런따위 실수를 알고도 再犯할리가 없는것이다.	방청하다	동사	심리_외부_청각
2128	그의 체취처럼 그의 몸뚱이에 붙어다니는 염세주의라는 것은 어디까지든지 게으른 성격이요 게다가 남의 염세주의는 어느 때나 우습게 알려드는 참 고약한 아리아욕(我利我慾)의 염세주의였다.	그의 體臭처럼 그의몸뚱이에 부터다니는 염세주의라는것은 어디까지든지 게을른性格이요 게다가 남의염세주의는 어느때가 우습게알러드는 참 고약한 我利我慾의 염세주의였다.	고약하다	형용사	심리_외부_후각
2193	두 사람은 서로 그리 부드럽지도 않은 피부를 느끼고 공기와 입술과의 따끈한 맛은 이렇게 다	두사람은 서로 그리 부드럽지도않은 피부를느끼고 공기와 입술과의 딱근한맛은 이렇게 다르	맛	명사	심리_외부_미각

	르고나를 시험한 데 지나지 않았다.	고 나를 시험한데 지나지 않었다.			
2194	이 방 소녀는 그의 거친 행동이 몹시 기다려졌다.	이방 少女는 그의 거츠른 행동이 몹시 기다려젔다.	거칠다	형용사	심리_외부_피부_촉각
2272	그는 또 한 번 가슴이 뜨끔했다.	그는 또한번 가슴이뜨끔했다.	뜨끔하다	형용사	심리_외부_피부_온도각
2345	그 때문에 그는 몹시 고민한다.	그때문에그는 몹씨고민한다.	고민하다	형용사	심리_외부_피부_통각
2399	시계를 들고 송 군의 어지러운 손목을 잡아 맥박을 계산하면서 한밤을 새라는 의사의 명령이다.	시게를들고 宋군의어즈러운손목을잡아 맥박을 게산하면서 한밤을새라는 의사의명령이었다.	어지럽다	형용사	심리_내부_평형
2439	그러나 그 남국적인 정열이 애타게 목말라서 벌들과 몇 사람의 환자가 화단 속을 초조히 거니는 것이었다.	그렇나 그남국적인정렬이 애타게목말라서벌들과몇사람의환자가화단속을초조히 거니는것이었다.	목마르다	형용사	심리_내부_유기

상기 표 5에 제시한 샘플에서도 알 수 있듯이, 이상 단편소설 감각 데이터셋은 'ID', '텍스트_민음사', '텍스트_소명출판', '감각어', '품사', '감각'의 도합 6열로 구조화되어 있다. 각각의 열(Column)에 대한 설명은 다음과 같다.

- ID : 전체 단편소설 13편을 문장 단위로 나누어 개별 문장에 부여한 고윳값을 기입한 열로, '이상 단편소설 기초 데이터셋'에서 설정한 ID 값과 동일하다. 이는 같은 ID 값을 기준으로, 보다 손쉬운 데이터 재구조화를 수행할 수 있도록 한 것이다.
- 텍스트_민음사 : 민음사 판본의 텍스트를 연구자가 직접 읽으며 문장 단위로 입력한 것으로, '이상 단편소설 기초 데이터셋'에 기입한 것과 같다.
- 텍스트_소명출판 : 소명출판 판본의 텍스트를 연구자가 직접 읽으며 문장 단위로 입력한 것으로, '이상 단편소설 기초 데이터셋'에 기입한 것과 같다.
- 감각어 : 특정 ID의 문장에서 감각 정보를 담은 감각어를 식별하여 입력한 것이다.
- 품사 : 입력한 감각어의 품사 정보를 입력한 것이다.
- 감각 : 입력한 감각어가 어떠한 감각 정보를 담고 있는지를 입력한 것이다.

이렇게 구축한 데이터셋을 바탕으로 이상 단편소설 속 감각의 양상에 대한 기계적 분석이 가능하다.[16] 가령 다음 표 6은 표 4의 감각 분류 모델에 따라 이상 단편소설 13편에 나타난 감각어를 식별하여 그 목록을 빈도와 함께 정리한 것이다. 개별 감각에 따라 시각어 205개, 청각어 52개, 미각어 20개, 후각어 19개, 피부감각어 171개(촉각어 57개, 온도각어 69개, 통각어 53개), 평형감각어 18개, 유기감각어 38개의 감각어가 도출되었으며, 감각 계층에 따른 별도의 접근을 수행할 수 있다.

표 6. 이상 단편소설 속 감각어 목록

감각				어휘 목록
1차	2차	3차	4차	
신체/심리	외부	시각		보다(191), 보이다(66), 들여다보다(15), 쳐다보다(15), 밝다(10), 내려다보다(10), 빛(9), 창백하다(7), 화려하다(6), 희다(6), 찬란하다(5), 검다(5), 내다보다(5), 발견하다(5), 느끼다(5), 구경(5), 바라다보다(5), 업신여기다(5), 어둡다(5), 뵙다(5), 백지(4), 침침하다(4), 시야(4), 부시다(4), 눈초리(4), 환하다(4), 붉은빛(4), 살피다(4), 퍼렇다(4), 붉다(3), 비치다(4), 파르스레하다(4), 혼도(4), 어둑어둑하다(3), 구경하다(3), 안색(3), 뜨이다(3), 빛깔(3), 초췌하다(3), 얼굴빛(3), 노려보다(3), 새빨갛다(3), 들여다보이다(3), 내다보이다(3), 시선(3), 홍발(3), 돌아다보다(3), 내보이다(3), 혼도하다(3), 시각(3), 둘러보다(3), 혼란하다(2), 불빛(2), 바라보다(2), 푸르다(2), 고동색(2), 색(2), 파랗다(2), 빨갛다(2), 알아보다(2), 돌아보다(2), 황홀하다(2), 캄캄하다(2), 살펴보다(2), 청춘(2), 뵈다(2), 파래지다(2), 백일(2), 희멀겋다(2), 황혼(2), 우매(2), 다홍(2), 홍조(2), 간파하다(2), 표백(2), 사팔뜨기(2), 근시(2), 당홍(2), 투시하다(2), 검은빛(2), 월광(2), 훤하다(2), 청년(2), 적빈(2), 백국(2), 쾌활하다(2), 주관적(2), 무시하다(2), 발갛다(2), 암시(2), 안목(1), 느낌(1), 눈총(1), 엿보다(1), 야맹증(1), 발견(1), 노랗다(1), 발견되다(1), 명월(1), 쳐다보이다(1), 까맣다(1), 백통색(1), 두리번두리번하다(1), 칙칙하다(1), 퀭하다(1), 홍차(1), 볼품(1), 굽어보다(1), 파리하다(1), 눈자위(1), 시뻘겋다(1), 새파랗다(1), 우매하다(1), 형안(1), 홍안(1), 목도하다(1), 청산(1), 혼사(1), 허옇다(1), 현란하다(1), 백면서생(1), 이색(1), 손색(1), 노랑돈(1), 도색(1), 당목하다(1), 맹목적(1), 백구(1), 백사(1), 벽(1), 간과하다(1), 창산(1), 건너다보다(1), 해반죽룩하다(1), 창연하다(1), 투시(1), 출색(1), 난시(1), 색맹(1), 녹엽(1), 혼수(1), 창공(1), 한눈팔다(1), 빛나다(1), 결백하다(1), 달빛(1), 암중모색(1), 엿보이다(1), 괄목하다(1), 고색창연하다(1), 소맥빛(1), 별빛(1), 원광(1), 넘보다(1), 우중충하다(1), 색채(1), 미명(1), 윤택(1), 짙다(1), 혈색(1), 붉히다(1), 황모(1), 흑판(1), 흑색(1), 생광(1), 진홍색(1), 검정(1), 저묾(1), 기웃거리다(1), 암암리(1), 적수공권(1), 색소(1), 환각(1), 낙조(1), 불콰하다(1), 백동화(1), 관망하다(1), 어둠(1), 견본(1), 일견(1), 흘겨보다(1), 구경꾼(1), 활연하다(1), 원망하다(1), 황포차(1), 은광(1), 광선(1), 흰빛(1), 새까맣다(1), 회색(1), 번쩍번쩍하다(1), 불변색(1), 벌게지다(1), 지키다(1), 색연필(1), 어슴푸레하다(1), 본숭만숭하다(1), 두리번거리다(1), 곁눈질(1), 광채(1), 거무튀튀하다(1), 컴컴하다(1), 색종이(1), 황금(1), 어두컴컴하다(1)
		청각		소리(62), 듣다(41), 목소리(27), 들리다(12), 조용하다(9), 노랫소리(7), 알아듣다(5), 음향(4), 기적(4), 울음(4), 기척(3), 사이렌(3), 종소리(3), 말소리(2), 엿듣다(2), 우렁차다(2), 소란하다(2), 기성(2), 큰소리(2), 묵묵하다(2), 쟁쟁하다(2), 발음(2), 소조하다(2), 잠잠하다(2), 야단(2), 대갈일성(2), 아늑하다(1), 청각(1), 발소리(1), 숨소리(1), 물소리(1), 야단스럽다(1), 고요하다(1), 곧이들리다(1), 시끄럽다(1), 울리다(1), 감각(1), 한적하다(1), 포성(1), 방청하다(1), 발성(1), 호곡(1), 적멸(1), 음성(1), 적막하다(1), 곧이듣다(1), 요란하다(1), 경청하다(1), 야단법석(1), 뒤법석(1), 수선(1), 고성(1)

		후각		냄새(14), 내(6), 향기(6), 체취(6), 맡다(4), 내음새(3), 고약하다(3), 새금하다(2), 향수(2), 꼬스르하다(1), 비린내(1), 찌르다(1), 쉬적지근하다(1), 방향(1), 냄새나다(1), 지각하다(1), 유취(1), 향기롭다(1), 악취(1)
		미각		맛(9), 싱겁다(5), 입맛(4), 맛보다(3), 맛있다(3), 쓰다(2), 단것(2), 맵다(2), 성미(2), 엉성하다(1), 고소(1), 씁쓸하다(1), 쌉싸름하다(1), 쓰디쓰다(1), 맹탕(1), 구수하다(1), 고언(1), 풍미(1), 무미건조하다(1), 달다(1)
		피부감각	촉각	굳다(9), 무겁다(7), 느끼다(7), 고집(5), 부드럽다(4), 축축하다(4), 든든하다(4), 살결(4), 무게(3), 촉각(3), 끈적끈적하다(3), 의젓하다(3), 스무드하다(3), 경편하다(2), 가볍다(2), 안타깝다(2), 튼튼하다(2), 묵직하다(2), 장중하다(2), 나스르르하다(2), 육중하다(2), 미끈하다(2), 거칠다(2), 중요하다(2), 황막하다(2), 배기다(1), 긴장(1), 가뿐하다(1), 거추장스럽다(1), 척척하다(1), 굳게(1), 껄껄하다(1), 두껍다랗다(1), 푼더분하다(1), 건조하다(1), 중병(1), 지중하다(1), 암팡지다(1), 거뜬하다(1), 번지레하다(1), 황량(1), 황량하다(1), 윤택(1), 체중(1), 경증(1), 위압(1), 윤(1), 경기구(1), 습기(1), 미끄럽다(1), 무미건조하다(1), 간질간질하다(1), 갑갑하다(1), 끈기(1), 경조부박하다(1), 중압(1), 보숭보숭하다(1)
			온도각	따뜻하다(17), 쾌감(8), 상쾌하다(7), 신선하다(6), 유쾌하다(5), 춥다(5), 서늘하다(4), 한심하다(4), 냉수(4), 선선하다(3), 훗훗하다(3), 차디차다(3), 냉회(2), 오한(2), 선뜩하다(2), 신열(2), 차다(2), 불유쾌하다(2), 온천(2), 덥다(2), 후덥지근하다(2), 시원하다(2), 훈풍(2), 온도(2), 뜨끔하다(2), 정열(2), 쾌활하다(2), 온돌방(2), 기온(1), 따끈따끈하다(1), 촉감(1), 써늘하다(1), 느끼다(1), 척척하다(1), 떨리다(1), 뜨겁다(1), 삭풍(1), 통쾌하다(1), 달다(1), 이글이글하다(1), 후끈후끈하다(1), 화끈화끈하다(1), 한등(1), 열성(1), 소슬하다(1), 빈한하다(1), 싸늘하다(1), 한산(1), 찬밥(1), 추상열일(1), 온아종용(1), 열렬하다(1), 따끈하다(1), 냉담하다(1), 선뜩선뜩하다(1), 발열(1), 열심(1), 쾌하다(1), 온순하다(1), 뜨뜻하다(1), 백열화하다(1), 활연하다(1), 산뜻하다(1), 미적지근하다(1), 체온(1), 열중(1), 선뜻하다(1), 냉장고(1), 온기(1)
			통각	아프다(12), 귀찮다(11), 앓다(10), 미안하다(6), 가난하다(5), 가렵다(4), 고통(4), 안일하다(3), 성가시다(3), 편하다(3), 고민(3), 근질근질하다(3), 경편하다(2), 거북살스럽다(2), 편리하다(2), 안타깝다(2), 불안(2), 거추장스럽다(2), 편안하다(2), 안녕하다(2), 불안하다(2), 저리다(2), 편의(2), 안전(2), 침통하다(2), 안심하다(2), 번민하다(2), 괴로움(2), 괴로워하다(2), 아늑하다(1), 쓰라리다(1), 안심(1), 느끼다(1), 재난(1), 불편하다(1), 안면(1), 통절하다(1), 통생(1), 거뜬하다(1), 간편하다(1), 고행(1), 심통(1), 심통하다(1), 산비하다(1), 고민하다(1), 간지럽다(1), 간질간질하다(1), 편두통(1), 인간고(1), 통봉(1), 터지다(1), 간지르다(1), 거북하다(1)
	내부	평형		취하다(12), 어지럽다(7), 혼도(4), 아뜩하다(3), 음란하다(3), 혼도하다(3), 혼란하다(2), 현기증(2), 어찔어찔하다(2), 수선(2), 소란하다(2), 문란하다(2), 돌다(2), 현란하다(2), 현란(1), 진동(1), 난마(1), 아찔하다(1)
		유기		피곤하다(18), 고프다(7), 피로(7), 배고프다(6), 고생(6), 졸리다(5), 공복(5), 끼치다(4), 혼곤하다(3), 졸음(3), 답답하다(3), 거북살스럽다(2), 느끼다(2), 동기(2), 노곤하다(2), 고단하다(2), 부르다(2), 가쁘다(2), 목마르다(2), 취중(2), 피로하다(1), 싫증(1), 울렁거리다(1), 뚝딱

				거리다(1), 찌뿌두둑하다(1), 메스껍다(1), 산란하다(1), 불편하다(1), 마렵다(1), 고달프다(1), 취기(1), 반사운동(1), 오예감(1), 두근두근하다(1), 곤하다(1), 맥박(1), 허기(1), 거북하다(1)

데이터 논문의 특성상, 본 데이터셋의 구체적인 활용 방안은 '이상 단편소설 기초·감각 데이터셋'을 통해 문학 텍스트 분석을 수행한 지해인(2024)[17]의 사례를 드는 것으로 갈음하고자 한다. 지해인(2024)은 이상 단편소설 기초·감각 데이터셋을 바탕으로 이상 단편소설 텍스트에 나타난 감각 양상을 살피고, 딥러닝 기반의 문장 단위 감정 분석을 병행하여 이상 텍스트에 나타난 감각 양상과 감정 양상 간 상관성과 그 함의를 살펴본 바 있다.

3. 결론

본 논문에서 소개한 '이상 단편소설 기초·감각 데이터셋'은 그 분량상 스몰 데이터(Small data)로, "양적으로 많지 않더라도 대상 분야 전문가에 의해서 적합한 의미가 부여돼 통계적 연구방법의 적용이 유의미한, 공개된 기계가독형 데이터(machine-readable data)"[18]라고 할 수 있다.

앞서 언급한 지해인(2024)[19]에서, 본 데이터셋을 바탕으로 감각 표현의 빈도와 그 양상에 대한 정량적 탐구를 감정 분석과 병행해 수행하였듯이, OSF 프로젝트로 공개한 본 데이터셋을 전초 기지 삼아 누구든 이상 단편소설에 대한 추가적인 기계적 분석을 수행할 수 있다. 예시로, 본 데이터셋은 현대어 역에 가까운 이상 단편소설 민음사 판본 텍스트를 모두 대조하여 기입하였으므로, 이에 형태소 분석을 활용한 기초적인 어휘 통계 분석을 적용해 볼 수 있다. 뿐만 아니라 원문에 근사하게 교열된 소명출판 판본의 이상 단편소설 텍스트를 실물 서적과 대조하여 기입하였으므로, Kiwi[20]의 서브워드 토크나이저(Subword Tokenizer)[21]를 활용해 해당 텍스트를 통계 기반의 서브워드(Subword) 단위로 분절하고 분석하는 것 역시 가능하다. 근대 문학 텍스트, 그중에서도 이상 단편소설 텍스트를 현대적 표기에 맞게 수정하지 않고 원문 그대로 분절하여 살필 수 있다는 점에서, 그간 한국 근대 문학 텍스트에 곧이곧대로 적용하기 힘들었던 형태소 분석의 한계를 일정 부분 극복한 새로운 접근이 점차로 가능할 것으로 사료된다. 또한 본 연구에서 문학 데이터셋의 구조와 설계에 대해 상술한바, 본 데이터셋이 향후 문학 데이터를 설계하고 구축하는 데 참고할 수 있는 선례로 남을 것으로 기대한다.

'이상 단편소설 기초·감각 데이터셋'은 대상 텍스트로 이상의 단편소설만을 선정하여 구축되었다. 구축 대상을 특정 작가의 작품에만 한정할 필요는 없으므로, 동시기 여러 작가와 작품을 아우르는 확장된 문학 데이터셋 구축함으로써 한국 근대문학의 비교 연구 수행을 위한 토대를 마련하는 것이 추후 과제라고 할 수 있다. 이러한 성취는 연구자 개인의 힘만으로 달성하기 어렵기 때문에, "연구자들에게 다양한 편찬을 직접 수행할 수 있는 플랫폼을 제공하고, 그 편찬에 대한 공로를 인정할 수 있는 편찬자 명기 등의 방안을 모색"할 필요가 있다는 김바로(2024)[22]의 제언을 한층 무겁게 숙고할 필요가 있다. 지속적인 데이터 작업을 토대로, 연구자는 한국 근대문학에 나타난 감각과 감정 양상에 대한 보다 총체적인, '멀리서 읽기'(Distance reading)를 접목한 연구를 수행하고자 하며, 이를 통해 개인의 감각과 감정을 넘어선, 시대적 맥락까지도 새롭게 읽어 낼 수 있으리라 기대한다.

[1] 지해인(2024). “디지털 감각·감정 분석을 통한 이상 문학의 에피파니 연구”. 한국학중앙연구원 한국학대학원 석사학위논문. https://www.riss.kr/link?id=T17070127

[2] 홍수현(2004). “李箱의 유일한 동화 ‘황소와 도깨비’ 창작 아닌 번안작품일 수도”. <이상리뷰>. 205-206. https://www.riss.kr/link?id=A103304974

[3] 다음의 문장은 민음사 판본 기준 549 자에 달하는 「休業과事情」의 한 문장이다. 문장의 길이로 인해 두 개 행(ID:813-814)에 걸쳐 입력하였다.

“잉크와 펜 원고지에 적히는 첫 자가 오자로 생겨 먹고 마는 것을 화를 내는 것 잡히지 않는 보산의 마음에 매어달려 데룽데룽하는 보산의 손이 종이를 꼬깃꼬깃 구겨서는 마당 한가운데에 홱 내어던진다는 것이 공교스러이도 SS 가 오늘 아침에 배앝아 놓은 침에서 대단히 가까운 범위 안에 떨어지고 만 것이 보산을 불유쾌하게 하여서 보산은 얼른 일어나 마당으로 내려가서는 그 구긴 종이를 다시 집어서는 보산이 인제 이만하면 적당하겠지 생각하는 자리에 갖다 떡 놓고 나서 생각하여 보니 그것은 버린 것이 아니라 갖다가 놓은 것이라 보산의 이 종이에 대한 본의를 투철치 못한 위반된 것이 분명하므로 그러면 이것을 방 안으로 가지고 돌아가서 다시 한 번 버려 보는 수밖에 없다 하여 그렇게 이번에야 하고 하여 보니 너무나 공교스러운 일에 공교스러운 일이 계속되는 것은 이것도 공교스러운 일인지 아닌지 자세히 모르는 것 같은 것쯤은 그대로 내어버려 두어도 관계치 않고 우선 이것을 내가 적당하다고 인정할 때까지 고쳐 하는 것이 없는 시간에 급선무라 하여 자꾸 해도 마찬가지고 고쳐 해도 마찬가지였다.”

다음 문장은 민음사 판본 기준 608 자에 달하는 「집팽이轢死」의 한 문장이다. 마찬가지로 두 행(ID:831-832)에 걸쳐 입력하였다.

“나는상에 노힌송아지고기를다먹은뒤에 냉수를청하얏드니 아주머니가 손소가저오는지라 죄송스럽다고그리닛가 이냉수한지게에 오전하는줄은 김상이 서울살아도—서울사닛가 몰으리라고그리길내 그것은쏘엇쌔서 그럿케냉수가갑이빗싸냐고그랫드니 이온천일대가 어듸를파든지 펄々슬는물밧게는 안솟는하느님헌테 죄바든쌍이되여서 냉수가먹고십흐면 보통갓흐면거저주는온천물을 듬뿍길어다가 잘식혀서 냉수를만들어서먹을것이로되 硫黃내음새가몹씨나는고로 서울서수도물만홀짝〳 마시고살아오든손님들이 짝질색들을하는고로 부득이지게를지고한마장이나 넘는정거장까지 냉수를한지게에오전식을주고사서 길어다먹는데 넘우거리가멀어서물통이 좀새든지하면 오전어치를사도이전어치밧게못어더먹으니 세음을싸지고보면 이냉수는한대접에 일전식은바다야경우가 올흔것이아니냐고 아주머니는그리는지라 그것참수고가만흐시다고 그럼이냉수는특별이 조심〳하야서 마시겟다고그랫드니 그럿치만냉수는 얼마든지거저들일것이니 념녀말고쑬썩〳 먹으라고그리는말을 듯고서야 S 와둘이비로소마음놋코 벌덕〳 먹엇슴니다”

다음 문장 역시 민음사 판본 기준 1,198 자에 달하는 「집팽이轢死」의 한 문장이다. 특히 긴 문장이므로 원활한 처리를 위해 세 행(ID:861-863)에 걸쳐 입력하였다.

“내가너무 『모나리사』 만을 바라다보닛가 마즌편에안젓는 항나적삼을 넙은비둘기가 참못난사람도다만타는듯이 내얼골을보고 나는그까짓일에 붓그러워할일은 아니닛가 막 『모나리사』를 보고십흔대로보고 『모나리사』는 내얼골을 보는 비둘기부인을 쏘좀조소하는듯이 바라보고 들어누어잇는 밧갓비둘기가 가만히보닛가 건너편에 안저잇는 『모나리사』가 자기안해를 그럿케 업슨역여보는것이 마음에 좀흡족하지못하야서 화를내이는기미로 벌덕닐어나안는바람에 들어눕느라고버서노흔구두에발이잘들어맛지안아서그만양말로담배꽁다리를밟을것을 S 가보고 싱그레웃으닛가 나도그눈치를채이고 S 를향하야마조싱그레우섯드니 그것이대단이실례행동갓고 쏘한편으로무슨음모나아닌가 퍽수상스러워서 저편에안저잇는금시게줄과 진흙무든흰구두가 눈을똥그럿케쓰고 이쑥을노려보닛가 당것장수할머니는쏘이쑥에무슨괴변이나 나지안앗나해서 역시눈을두리번〳하다가아모일도업스닛가 싱거워서 눈을도로그마즌편의 금시게줄로 옴겨 노흘적에 S 는보든신문을쳑々접어서 인생관가방속에다가집어늣트니 정식으로 『모나리사』와 비둘기는 어는편이더어엿쏀가를판단할작정인모양으로 안경을바로잡드니 참세게에이런기차는 다시업스리라고한마데하닛가 비둘기와 『모나리사』 가 S 쑥을일시에보는지라 나는쏘창밧갓 논속에허수아비갓흔황새가 한마리나려안젓스니 저것좀보라고 소리를질넛드니 두미인은쏘일시에시선을나잇는창밧갓흐로 옴겨보앗는데 결국아모것도 보히지안으닛가 싱그레우스면서 내얼골을한 번식보드니 『모나리사』 는생각난듯이 겻헤 『비프스테이크』 갓흔밧갓어룬의기름씨흘으는코잔등이 근처를 한번 들여다보는것을본나는 속마음으로 참앗갑도다

그럿케생각하고잇는데 S 는 무슨생각으로앗랫는지 개발에편자라는말이잇지안으냐고그리면서 나에게해태한개를주는지라 성냥을그어서불을부치려닛가 내겻헤안젓는갓쓴해태가성냥을좀달나고 그리길내주엇드니 서울서주머니에너허가지고간『캬페』석냥이되여서이상스럽다는듯이 두어번두집어보드니 집고들어온 길고도굵은 얼는보면 몽둥이가튼 집행이를 방헤안되도록한쪽으로치워노려고노차마자 꾀크게와직근하는소리가나면서 그길다란집행이가 간데온데가업습니다"

[4] 딥러닝 분석을 위해 채택한 KcELECTRA 모델이 처리할 수 있는 문장의 길이, max_len 이 512 여서 한 번에 512 자까지 분석할 수 있기 때문에 이러한 방식을 택한 것이다.

[5] 전두영, "KOTE (Korean Online That-gul Emotions) Dataset", https://github.com/searle-j/KOTE

[6] Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

[7] 이준범, "KcELECTRA: Korean comments ELECTRA", https://github.com/Beomi/KcELECTRA

[8] 지해인(2024). "디지털 감각·감정 분석을 통한 이상 문학의 에피파니 연구". 한국학중앙연구원 한국학대학원 석사학위논문. https://www.riss.kr/link?id=T17070127

[9] 예를 들어, 구축한 '이상 단편소설 기초 데이터셋'을 바탕으로 지해인(2024)에서 문장(열) 단위 딥러닝 감정 분석을 수행한 바 있다. 해당 데이터셋을 코랩(Colab) 환경에 불러와 조작하기 위해서는 다음의 파일을 참고할 수 있다. 지해인. "Analysis_이상단편소설_감정분석.ipynb". https://osf.io/ca64z. 다음은 해당 주피터 노트북 파일의 코드 셀 일부분을 가져온 것이다.

```
FILEID = "1N-Qcqe8q5GUuKwaL1Qd6A52GxOexBOCK"
FILENAME = "Rawdata_이상단편소설_기초데이터.tsv"
!wget --load-cookies ~/cookies.txt "https://docs.google.com/uc?export=download&confirm=$(wget --quiet --save-cookies ~/cookies.txt --keep-session-cookies --no-check-certificate 'https://docs.google.com/uc?export=download&id={FILEID}' -O- | sed -rn 's/.*confirm=([0-9A-Za-z_]+ ).*/₩1₩n/p')&id={FILEID}" -O {FILENAME} && rm -rf ~/cookies.txt
```

구글 드라이브에 업로드한 데이터셋을 활용하기 위해, 상기 코드가 담긴 코드 셀을 실행시킬 필요가 있다. 상기 코드는 해당 데이터셋을 Wget 패키지를 활용하여 'Rawdata_이상단편소설_기초데이터.tsv'라는 이름으로 불러오기 위한 코드이다.

```
import pandas as pd

# "rawdata.txt" 파일의 내용을 "원본데이터" 변수로 불러오기
원본데이터 = pd.read_csv(FILENAME, sep="₩t")
원본데이터
```

이후 상기 코드 셀을 실행시키면, 판다스(Pandas) 라이브러리로 불러온 'Rawdata_이상단편소설_기초데이터.tsv' 파일을 표처럼 행과 열로 구성된 데이터프레임(Dataframe) 형식으로 변환할 수 있다. 이후 판다스(Pandas)의 여러 메서드를 활용해 데이터셋을 원하는 형태로 가공하거나, 기타 라이브러리를 임포트하여 감정 분석 등의 분석 방법론을 적용할 수 있다.

[10] 최현배(1937). <우리말본>. 延禧專門學校出版部. 647-649. https://lod.nl.go.kr/page/KMO000013361

[11] 임윤정(2019). "한국어 촉각 형용사의 확장 의미 연구". 경희대학교 석사학위논문. 14. https://www.riss.kr/link?id=T15347853

[12] 임윤정(2019). "한국어 촉각 형용사의 확장 의미 연구". 경희대학교 석사학위논문. 15. https://www.riss.kr/link?id=T15347853

[13] 김창섭(1985). "시각형용사의 어휘론". <관악어문연구>. 149-176. https://www.riss.kr/link?id=A3284058

[14] 전은진(2011). "후각 형용사의 의미론적 연구". <국제어문>. 9-47. https://www.riss.kr/link?id=A103032634

[15] 국립국어연구원(2000). "표준국어대사전 편찬 지침 2". 76. https://www.korean.go.kr/front/etcData/etcDataView.do?mn_id=46&etc_seq=31

[16] 해당 데이터셋을 코랩(Colab) 환경에 불러와 조작하기 위해 다음의 파일을 참고할 수 있다. 지해인. "Analysis_이상단편소설_감각분석.ipynb". https://osf.io/p6y7c. 다음은 해당 주피터 노트북 파일의 코드 일부분을 가져온 것이다.

```
FILEID = "11K99r1yyda8kAf4WoV0e1A9WYvzPJq2M"
FILENAME = "Rawdata_이상단편소설_감각데이터.tsv"
!wget --load-cookies ~/cookies.txt "https://docs.google.com/uc?export=download&confirm=$(wget --quiet --save-cookies ~/cookies.txt --keep-session-cookies --no-check-certificate 'https://docs.google.com/uc?export=download&id={FILEID}' -O- | sed -rn 's/.*confirm=([0-9A-Za-z_]+ ).*/₩1₩n/p')&id={FILEID}" -O {FILENAME} && rm -rf ~/cookies.txt
```

구글 드라이브에 업로드한 데이터셋을 활용하기 위해, 상기 코드가 담긴 코드 셀을 실행시킬 필요가 있다. 상기 코드는 해당 데이터셋을 Wget 패키지를 활용하여 'Rawdata_이상단편소설_감각데이터.tsv'라는 이름으로 불러오기 위한 코드이다.

```
# "rawdata.txt" 파일의 내용을 "감각_원본데이터" 변수로 불러오기
감각_원본데이터 = pd.read_csv(FILENAME, sep="₩t")

# 필요 없는 열 제거
감각_원본데이터 = 감각_원본데이터.loc[:, ~감각_원본데이터.columns.str.contains('^Unnamed')]
감각_원본데이터
```

이후 상기 코드 셀을 통해, 판다스(Pandas) 라이브러리로 불러온 'Rawdata_이상단편소설_감각데이터.tsv' 파일을 행과 열로 구성된 데이터프레임(Dataframe) 형식으로 변환하고, Unnamed 로 임시 명명된, 불필요한 열을 제거할 수 있다. 이후의 감각 분석 프로세스는 본 논문에서는 상술하지는 않으나, 해당 주피터 노트북 파일을 통해 구체적인 소스 코드를 살펴볼 수 있다.

[17] 지해인(2024). "디지털 감각·감정 분석을 통한 이상 문학의 에피파니 연구". 한국학중앙연구원 한국학대학원 석사학위논문. https://www.riss.kr/link?id=T17070127

[18] 윤미선(2023). "인문 "스몰 데이터" 연구 방법론과 사례 연구: 19 세기말 영국 정기간행물 비평 담론 - 주간지 『런던』을 중심으로". <영미문학연구>. 83-135. https://doi.org/10.46562/jesk.44.4

19 지해인(2024). "디지털 감각·감정 분석을 통한 이상 문학의 에피파니 연구". 한국학중앙연구원 한국학대학원 석사학위논문. https://www.riss.kr/link?id=T17070127

20 이민철(2024). "Kiwi: 통계적 언어 모델과 Skip-Bigram 을 이용한 한국어 형태소 분석기 구현". <디지털인문학>. 109-136. https://doi.org/10.23287/KJDH.2024.1.1.6; 이민철. "Kiwi : 지능형 한국어 형태소 분석기(Korean Intelligent Word Identifier)". https://github.com/bab2min/Kiwi

21 이민철. "Module kiwipiepy.sw_tokenizer". https://bab2min.github.io/kiwipiepy/v0.17.1/kr/sw_tokenizer.html

22 김바로(2024). "국사편찬위원회 한국근현대잡지자료 데이터(2024.03.27.)". <디지털인문학>. 151. https://doi.org/10.23287/KJDH.2024.1.1.8

참고문헌

- 연구 데이터

지해인. "디지털 감각·감정 분석을 통한 이상 문학의 에피파니 연구". https://osf.io/964nc

지해인. "Analysis_이상단편소설_감각분석.ipynb". https://osf.io/p6y7c

지해인. "Analysis_이상단편소설_감정분석.ipynb". https://osf.io/ca64z

지해인. "Rawdata_이상단편소설_기초데이터.tsv". https://osf.io/a69ec

지해인. “Rawdata_이상단편소설_감각데이터.tsv”. https://osf.io/5st4r

• 저서

이상(2012). 권영민 편. <이상 소설 전집>. 민음사. https://lod.nl.go.kr/resource/KMO201251206

이상(2009). 김주현 편. <정본 이상문학전집>. 소명출판. https://lod.nl.go.kr/page/KMO201002593

최현배(1937). <우리말본>. 延禧專門學校出版部. https://lod.nl.go.kr/page/KMO000013361

• 논문

김바로(2024). “국사편찬위원회 한국근현대잡지자료 데이터(2024.03.27.)”. <디지털인문학>. 143-156. https://doi.org/10.23287/KJDH.2024.1.1.8

김창섭(1985). “시각형용사의 어휘론”. <관악어문연구>. 149-176. https://www.riss.kr/link?id=A3284058

윤미선(2023). “인문 “스몰 데이터” 연구 방법론과 사례 연구: 19 세기말 영국 정기간행물 비평 담론 - 주간지 『런던』을 중심으로”. <영미문학연구>. 83-135. https://doi.org/10.46562/jesk.44.4

이민철(2024). “Kiwi: 통계적 언어 모델과 Skip-Bigram 을 이용한 한국어 형태소 분석기 구현”. <디지털인문학>. 109-136. https://doi.org/10.23287/KJDH.2024.1.1.6

임윤정(2019). “한국어 촉각 형용사의 확장 의미 연구”. 경희대학교 석사학위논문. https://www.riss.kr/link?id=T15347853

전은진(2011). “후각 형용사의 의미론적 연구”. <국제어문>. 9-47. https://www.riss.kr/link?id=A103032634

지해인(2024). “디지털 감각·감정 분석을 통한 이상 문학의 에피파니 연구”. <상허학보>. 753-827. https://doi.org/10.22936/sh.72..202410.021

지해인(2024). “디지털 감각·감정 분석을 통한 이상 문학의 에피파니 연구”. 한국학중앙연구원 한국학대학원 석사학위논문. https://www.riss.kr/link?id=T17070127

홍수현(2004). “李箱의 유일한 동화 ‘황소와 도깨비’ 창작 아닌 번안작품일 수도”. <이상리뷰>. 205-206. https://www.riss.kr/link?id=A103304974

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

• 웹 자원

국립국어연구원(2000). “표준국어대사전 편찬 지침 2”. https://www.korean.go.kr/front/etcData/etcDataView.do?mn_id=46&etc_seq=31

이민철. “Kiwi : 지능형 한국어 형태소 분석기(Korean Intelligent Word Identifier)”. https://github.com/bab2min/Kiwi

이민철. “Module kiwipiepy.sw_tokenizer”. https://bab2min.github.io/kiwipiepy/v0.17.1/kr/sw_tokenizer.html

이준범, “KcELECTRA: Korean comments ELECTRA”, https://github.com/Beomi/KcELECTRA

전두영, “KOTE (Korean Online That-gul Emotions) Dataset”, https://github.com/searle-j/KOTE