서술 시간(narrated time) 추정의 언어 간 검증: 한국어 문학 텍스트와 LLM의 비교

이시헌[1], 황태검[1], 박혜승[1*]
[1] *협성대학교 소프트웨어공학과, 화성, 대한민국*
[*] *Corresponding Author* : hs2000park@uhs.ac.kr

1. 배경 및 목적

서술 시간은 작품 속에서 흐른 시간을 말한다. 발화의 경우는 실제로 말하는데 소요되는 시간이 되고, 회상이 전개된다면 회상 속의 시간 흐름인 것이다.

이 시간 흐름은 인간이 갖고 있는 일반 배경 지식에 기반하여 추정될 수 있다. 훈련된 독자가 아니더라도, 본능적으로 그 시간 흐름을 느낀다. 소설 속 주인공이 급하게 면도를 하고 집 밖으로 뛰어나가는 장면이 있다고 하자. 묘사 속 행위들의 소요시간을 대략적으로 추정하는 일은 어렵지 않다. 그러나 이 '서술 시간'에 기반한 문학 작품의 분석은 굉장히 까다롭다. 매 순간의 시간 흐름을 모두 기록하는 것은 지난한 작업이고, 그것을 여러 작품에 대하여 수행하기는 불가능에 가깝다.

한편, 문학 텍스트를 자동적으로 분석(automate content analysis)하는 기존의 방법론들이 없는 것은 아니다. 대표적인 것은 'Named entity extraction'과 'Topic modeling'이 있다. 그러나 이들 방법론은 기본적으로 '단어를 세어' 통계량을 얻는 것이고, 명백한 한계를 갖는다. 예를 들면 "소설 한 페이지에 평균적으로 얼마나 긴 시간이 흐르는가?"와 같은 질문이 있다. 다음 문장을 보자.

"wow, it's been thirty years but feels like yesterday"

Ted Underwood(2023)는 위와 같은 문장에서 30년의 시간이 흘렀다고 계산해서는 안된다는 점을 지적하며, GPT-4를 사용한 시간 흐름 추정이 유의미하다는 견해를 제시한다. 그러나 Underwood의 연구는 그 결과물이 디지털 문학 연구에 유효함을 입증했음에도, 여전히 개념 증명 단계에 머무르고 있으며, 언어의 차이 또한 존재한다. **이에 본 연구는 LLM의 서술 시간 측정이 한국어 텍스트에 대하여 유효한지를 검증하고, 연구 결과를 간단히 사용할 수 있는 파이썬 라이브러리[1]로 만들어 배포하는 것을 목적으로 한다.**

2. 작품 선정

Underwood는 그의 2018년도 연구에서 90편의 작품을 선정하였다. 그는 1719년부터 1996년까지의 범위에서 정전 작품(canonical works), 대중 소설(popular fiction), 무작위 선정 작품의 세 분류에 맞추어 90편을 선정한 것이다. 이후 각 작품을 250단어 길이로 잘라서, 앞의 두 조각, 뒤의 두 조각과 중간의 무작위 12조각을 샘플링 하여, 총 16조각의 데이터로 가공하였다. 그리고 인간 연구자 세 명이 독립적으로 각 조각의 서술 시간을 기록하였다.

본 연구 또한 동일한 분류 기준과 샘플링 방법론을 적용하여, 총 80개의 한국어 작품을 선정하고, 세 명의 인간 주석자를 확보하여 서술 시간을 측정하고 있다.

소설 선정은 1906년 발표된 이인직의 '혈의 누' 부터, 2019년 발표된 김초엽의 '우리가 빛의 속도로 갈 수 없다면' 까지의 113년 범위로 정했다. 1988년 발표된 신영복의 '감옥으로부터의 사색'과 같은 비소설 작품이 12편 선정되어 있으며, 10년 단위로 작품의 개수를 최대한 균일하게 맞추었다.

3. 현황 및 방법론

2026년 2월 1일 기준으로 41작품(656조각)에 대한 샘플링이 완료되었고, 총 1280조각을 기준으로 평균 20.8%만큼 주석 작업이 완료되었다. 주석 작업은 전용 웹페이지를 제작하여 주석자들이 원격을 접속하여 수행한다. 미리 가공하여 서버에 저장된 텍스트 조각 하나씩을 웹페이지에서 제공하고, 주석자는 해당 문장을 읽고 서술 시간을 분 단위[2]로 평가한다. (Figure 1)

이때 한국어 문장의 특성을 고려하여, 영문 250단어 대신 한글 500자를 기준으로 샘플링을 수행했다. 다소 길이가 짧아 보일 수 있으나, 교착어의 특성상 문장의 길이에 비해 더 정보량이 많다는 것, 약 300자 정도의 길이만으로도 충분한 시간 흐름이 관찰된다는 점 등을 고려하여 350~650자 범위에서 원문 텍스트를 쪼개도록 코드를 구현하였다.

주석자가 입력한 세 가지 시간은, 주석자가 읽은 텍스트와 함께, Underwood(2018)의 방법론대로 정리되어, 서버에 JSON 형태로 기록된다.[3]

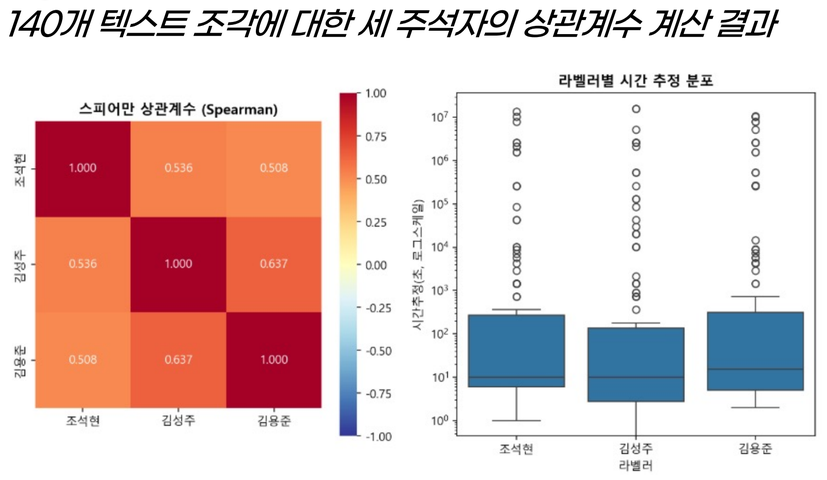
현재 세 주석자가 공통으로 작업한 조각은 140개로, 해당 데이터에 대하여 스피어만[4](Spearman) 상관계수를 계산하면 'Figure 2'와 같다. Underwood(2023)의 127개 청크에 대한 두 작업자의 상관계수(r=.74)보다는 낮은 r=.5~.64 ($p < .001$)의 값이 계산되었으나, 서술 시간을 관찰하기에는 충분히 유효하다.

한편, 디지털 텍스트가 없는 경우는 실물 자료에서 무작위 샘플링을 수행해야 했다. 이를 위해 무작위 추출한 페이지에 대하여 OCR 기술로 실물 책의 '페이지 당 글자 수'를 산출하고, 자료가 갖는 행 개수를 기반으로 무작위 추출을 시작할 12개의 '페이지 번호와 행 번호' 쌍을 계산하였다.

Figure 1
원본 텍스트와 서술 시간의 JSON 형식 기록 예시


```
"segments": [
  {
    "idx": 0,
    "text": "혈(血)의 누(淚)\n\n일청전쟁(日淸戰爭)의 총소리는 평양 일경이 떠나가는 듯하더니, 그 총소리가 그치매 사람의 자취는 끊어지고 산과 들에 비린 티끌뿐이라.\n평양성의 모란봉에 떨어지는 저녁 볕은 뉘엿뉘엿 넘어가는데, 저 햇빛을 붙들어매고 싶은 마음에 붙들어매지는 못하고 숨이 턱에 닿은 듯이 갈팡질팡하는 한 부인이 나이 삼십이 될락말락하고, 얼굴은 분을 따고 넣은 듯이 흰 얼굴이나 인정 없이 뜨겁게 내리쪼이는 가을 볕에 얼굴이 익어서 선앵둣빛이 되고, 걸음걸이는 허둥지둥하는데 옷은 흘러내려서 젖가슴이 다 드러나고 치맛자락은 땅에 질질 끌려서 걸음을 걷는 대로 치마가 밟히니, 그 부인은 아무리 급한 걸음걸이를하더라도 멀리 가지도 못하고 허둥거리기만 한다.\n남이 그 모양을 볼 지경이면 저렇게 어여쁜 젊은 여편네가 술 먹고 한길에 나와서 주정한다 할터이나, 그 부인은 술 먹었다 하는 말은 고사하고 미쳤다, 지랄한다 하더라도 그따위 소리는 귀에 들리지 아니할 만하더라.",
    "word_count": 116,
    "char_count": 370,
    "complete": true,
    "narratedtime": 5,
    "ellipsistime": 0,
    "subjectivetime": 0,
    "ellipsisphrase": "",
    "subjectivephrase": "",
    "labeled_at": "2025-12-23 13:45:04"
  },
```


Figure 2
140개 텍스트 조각에 대한 세 주석자의 상관계수 계산 결과



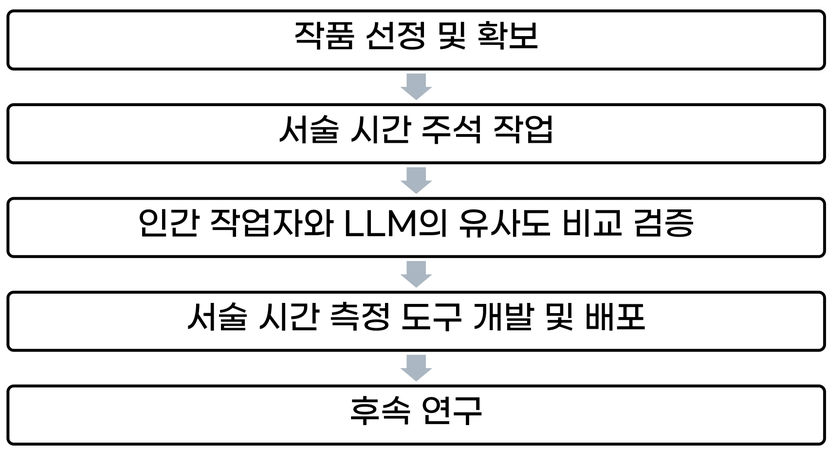
4. 의의 및 기대효과

본 연구는 'Figure 3'과 같은 절차로 진행되며, **최종적으로 '서술 시간이 기록된 한국어 텍스트 데이터셋'과, '서술 시간을 측정할 수 있는 연구 도구'를 산출한다.**

본 연구가 산출하는 1280개 조각의 텍스트는 총 640,000자 분량으로, 한국 문학의 역사적 흐름을 원거리 읽기에 기반하여 조망하기에 충분한 크기를 갖추었다. 이를 통해 **다양한 문학사적 연구가 촉발되리라 기대한다.**

본 연구는 산출한 데이터셋을 통해, 한국어 소설의 서술 시간을 LLM이 추정할 수 있음을 증명한다. 막연한 직관을 넘어, 실제로 LLM을 통해 서술 시간을 추정할 수 있음이 증명된다면, 문학 작품의 특징과 구조를 분석할 새로운 도구로서 LLM을 사용할 수 있게 된다. **이를 통해 대량의 문학 작품을 서술 시간의 관점에서 연구할 수 있게 될 것이다.**

Figure 3
연구 절차 도식



1) fic-time 이라는 이름으로 TestPyPy에 게시되어 있다 (이시헌, 2025). Underwood(2023)의 작업물을 파이썬 라이브러리로 옮겨 놓은 것으로, 영문에 대해서는 검증되었다.
2) Underwood(2018) 연구는 서술 시간을 분(Minutes) 단위로 측정해야 한다고 밝힌다.
3) 각 주석자별로 별도의 파일이 생성되며, Underwood(2018)의 작업과 기본 형태가 같다.
4) 아주 긴 시간 평가값이 이상치로 동작한다고 판단하여, 이상치에 대하여 조금 더 강건한 Spearman Correlation을 사용하였다. 데이터에 대한 정제 작업 없이, 현재까지 완성된 평가 값을 일괄적으로 사용한 결과이며, 추후 데이터셋이 완성되면 더 높은 상관계수가 나올 것으로 예상한다.

참고문헌

- Underwood, T. (2018). Why literary time is measured in minutes. New Literary History, 85(2), 351-365.
- Underwood, T. (2023, March 19). Using GPT-4 to measure the passage of time in fiction. https://tedunderwood.com/2023/03/19/using-gpt-4-to-measure-the-passage-of-time-in-fiction/
- 이시헌 (2025). *fic_time* [GitHub repository]. GitHub. https://github.com/lamaBread/fic_time