번역학연구, 27(1), 163-191 (2026년 3월 31일)
https://doi.org/10.15749/jts.2026.27.1.005

LLM 문학 번역의 창의적 변이에 대한 계량적 분석: 온도와 프롬프트 전략의 상호작용을 중심으로*

임진(이화여자대학교)

1. 서론

창의적 특성으로 인해 “인간 번역의 마지막 보루(the last bastion of human translation)”(Toral & Way, 2014, p. 714)라 일컬어지는 문학 번역에 기계번역(MT)을 적용하는 연구가 증가하는 추세이다. 이는 번역학뿐만 아니라 전산언어학을 포함한 대형언어모델(LLM) 연구 전반에서 나타나는 창의적 글쓰기 및 예술적 작업에 대한 높은 관심에서 비롯된 결과라고 볼 수 있다. 번역은 LLM의 일차적인 핵심 기능은 아님에도 불구하고, LLM 번역이 인간의 번역 실천 방식에 지대한 영향을 미치고 있다는 점에서 번역학 내에서도 이에 대한 학술적 관심과 체계적인 탐색이 지속되고 있다.

기존 번역 전용 모델(NMT)의 경우 원문을 제공하는 것 이외에 번역 생산에 있어 인간의 개입이 제한적이기에 문학 번역의 문체적 정교함과 창의성 구현에 있어 여러 한계를 노정해 왔다(Abdelhalim et al., 2025; Guerberof-Arenas & Toral, 2020, 2022). 그러나 대규모 언어모델(LLM) 기반의 번역은 프롬프트(Prompt)를 통해 사용자가 출력의 성격을 직접 조정할 수 있다는 점에서 창의적 텍스트 번역에 대한 새로운 가능성을 열어주었으

* 본 논문에 소중한 통찰을 나누어 주신 익명의 심사자 세 분께 깊이 감사드립니다.

며, 국내외 학계에서도 이에 대한 논의가 본격적으로 확대되는 추세이다(김현웅과 이상빈, 2025; 마승혜, 2025; Manapbayeva et al., 2024).

번역에서의 창의성은 원문을 단순히 변형하는 것이 아니라, 의미를 수용 가능한 수준에서 보존하면서도 번역문의 예술적 참신성을 확보하는 과정으로 이해된다(Guerberof-Arenas & Toral, 2020, 2022). LLM 번역에서 이러한 창의성은 단어 생성 시 확률 분포의 무작위성을 조정하는 내부 매개변수인 온도(Temperature)와 외부적 지시 체계인 프롬프트의 상호작용을 통해 발현된다. 높은 온도 설정이 창의적 번역에 도움이 된다는 소수 연구가 존재하지만(마승혜, 2025; Du et al., 2025) 온도가 번역에 미치는 구체적인 영향이나 프롬프트 효과와의 상호작용 등에 대한 연구는 아직 미진하기에 그 필요성이 제기되고 있다(Li et al., 2025, p. 2). 이러한 관점에서 LLM 문학 번역을 논의하기 위해서는 인간 번역과의 단순한 성능 비교를 넘어, 창의성 유도에 대한 모델의 반응 양상을 구조적으로 이해할 필요가 있다.

본고는 프롬프트 전략과 온도 변수의 상호작용이 LLM 번역 출력에 어떠한 언어적 변이를 산출하는지를 계량적으로 분석함으로써, LLM 문학 번역의 작동 원리를 구체화하고자 한다. 아울러 통제된 실험 설계를 통해 분석의 신뢰성을 확보함으로써, 기술 중심의 AI 번역 담론 속에서 번역학적 관점의 분석 지평을 확장하는 데 기여하고자 한다.

2. LLM 번역의 창의성

본고의 목적은 프롬프트 및 온도 변수를 통해 창의적 번역을 유도하는 경우 문학 텍스트의 LLM 번역 출력에 어떠한 변화가 일어나는지를 탐색하는 것이다. 이를 위해 우선 번역에서 창의성에 대한 개념을 정의하고, LLM 출력의 창의성에 대한 기존 논의를 개괄함으로써 본고의 분석 기준을 정립하고자 한다.

2.1 창의성과 번역

보덴(Boden, 2004, pp. 1-10)에 의하면 창의성은 새롭고(new) 놀라우면서

도(surprising) 가치있는(valuable) 인공물(artifacts)을 만들어내는 능력이다. 일반적인 맥락에서의 창의성과 달리 번역에서의 창의성에는 원문과의 관계가 개입된다. 번역에서 창의성은 기존의 것과 다른 독창성(originality), 새로움(novelty)과 함께 해당 맥락에서 가치가 있거나(valuable) 유용하고(useful) 적절한(appropriate)한 것을 뜻한다(Aranda, 2009). 이는 참신성(novelty)과 수용성(acceptability) 간의 균형으로 정의된다(Bayer-Hohenwarter, 2010; Guerberof-Arenas & Toral, 2020, 2022).

다시 말하면, 독창적인 대안을 사용한다고 해서 무조건 창조적 번역이 되는 것은 아니다. 창조적 번역은 참신하면서도 번역의 목적에 부합하고 목표 문화의 맥락에서 유용하고 적절하게 수용될 수 있어야 한다(Guerberof-Arenas & Toral, 2020, 2022).

번역학에서 창의적 번역이란 주로 번역을 원문과 비교하여 어느 정도의 변이(shift)가 발생했는지를 평가하는 방식으로 규정되어 왔다(O'Sullivan, 2013). 이를 창조적 변이(creative shifts)(Bayer-Hohenwarter, 2011)라 하는데, 여기에는 원문을 보다 일반적이고 추상적으로 바꾸는 추상화(abstraction), 원문을 다른 대안으로 변경하는 변형(modification), 원문 내용을 더 구체적으로 명시화하는 구체화(concretization)가 포함된다. 창조적 변이에 번역 오류가 없고, 이러한 번역이 목표 언어에서 수용 가능하다고 판단될 때 창의적 번역으로 평가된다.

보다 구체적으로는 창조적 변이에 해당되는 번역을 질적 분석을 통해 식별하여 가산점을 부여하고 오류에 대해서는 감점을 통해 창의성 점수를 산출하는 방식으로 번역의 창의성을 평가해 왔다(Bayer-Hohenwarter, 2010; Guerberof-Arenas & Toral, 2020, 2022). 그런데 번역의 창의성에 대한 판단은 어디까지나 주관적일 수밖에 없다. 복수의 인간 평가자가 참여하고 평가자 간 신뢰도를 검증하는 방식이 활용되어 왔지만, 세부 범주에 대해서는 상당한 판단의 차이가 존재하는 것으로 알려져 있다(Guerberof-Arenas & Toral, 2022). 문학 번역과 같은 창의적 번역 평가에 존재하는 본질적인 어려움은 MT 문학 번역 평가에서도 이어지고 있으며, 기존 평가 지표와 차별화된 지표의 필요성이 꾸준히 제기되고 있다(예: Zhang et al., 2025).

2.2 MT 문학번역의 창의성

번역의 창의성 논의는 NMT나 LLM 번역 품질 향상과 함께 본격화되었다. NMT는 문학 텍스트와 같이 창의적인 접근을 요하는 텍스트 번역에서 인간 번역가에 버금가는 품질의 번역을 구현하는 데 한계가 있는 것으로 지적되어 왔다(Abdelhalim et al., 2025; Guerberof-Arenas & Toral, 2022; Hu & Li, 2023). 지나친 직역과 획일화된 문체, 문화적 뉘앙스 처리의 한계(Abdelhalim et al., 2025)와 함께 문맥 파악의 어려움(Guerberof-Arenas & Toral, 2022)으로 번역 문제가 발생할 가능성이 높다는 것이다. 또한 NMT 결과물을 인간 번역가가 편집(post-editing)하는 경우, 인간 번역가는 NMT가 선택한 어휘나 문장 구조를 벗어나 창의적인 대안을 제시하기 어려우며(Guerberof-Arenas & Toral, 2022) 인간 번역가가 처음부터 번역하는 것보다 창의성이 떨어지는 번역에 안주하게 되는 것으로 알려져 있다(Guerberof-Arenas & Toral, 2020).

그러나 사용자가 프롬프트를 통해 출력을 세밀하게 조정할 수 있는 LLM의 등장은 이러한 한계를 극복하기 위한 새로운 가능성을 제시하고 있다. LLM 번역 출력의 창의성을 조절하는 핵심 기제로는 프롬프트 전략과 온도(Temperature) 설정을 꼽을 수 있다. 온도는 LLM이 단어를 생성할 때 확률 분포의 무작위성(randomness)을 제어하는 매개변수이다(Peeperkorn et al., 2024). 원래 통계 역학에서 쓰이던 개념을 애클리 외(Ackley et al., 1985)에서 최초로 확률적 신경망 모델에 도입하였다. LLM에서 낮은 온도(<1)는 결정론적인 출력을, 높은 온도(>1)는 보다 창의적이고 다양한 출력을 유도하는 것으로 알려져 있다(Peeperkorn et al., 2024). 일반적인 웹 인터페이스 환경에서는 온도를 조정할 수 없지만, API 호출을 통해 LLM에 접근하면 온도 설정이 가능하다. 온도는 LLM의 성능에 직결되는 핵심 매개변수임에도 불구하고, 실제 과업 수행에 미치는 영향에 대해서는 체계적인 탐색이 부족한 실정이다. 특히 특정 과업에 최적화된 온도 가이드라인이 부재하여 이에 대한 연구의 필요성이 지속적으로 제기되고 있다(Li et al., 2025, p. 2).

온도의 특성은 LLM 번역 출력에서도 유사하게 나타나는 것으로 보고된다. 높은 온도 설정은 출력의 다양성을 풍부하게 하지만, 보편적인 번역 과업에서는 오히려 번역 품질 점수의 저하를 초래하기도 한다(Peng et al.,

2023; Li et al., 2025). 반면, 창의적 성격이 강한 문학 번역에서는 높은 온도가 긍정적인 영향을 미치기도 한다(마승혜, 2025; Du et al., 2025). GPT를 활용한 서구권 및 아시아권 언어 번역 실험에서는 온도가 1.0일 때 창의적 변이가 극대화되지만, 낮은 온도에서 생성된 번역의 경우 문법적 정확성은 높지만 창의성이 떨어지는 것으로 나타났다(Du et al., 2025). 마승혜(2025) 역시 한국어 문학 텍스트 은유와 직유 표현을 적절한 프롬프트와 높은 온도 설정을 결합하여 효과적으로 번역할 수 있는 것으로 보고한 바 있다. 그러나 높은 온도는 환각으로 인한 오류 생성 가능성을 높인다는 점(Li et al., 2025)에서 창의성과 정확성 사이의 균형점을 찾는 것이 중요하다.

또한 LLM 번역 출력을 조정하는 대표적인 방법으로 활발한 연구가 이루어지고 있는 것이 프롬프트이다. 제로샷(Wei et al., 2022), 퓨샷(few-shot) (Gao et al., 2023), 페르소나 등 역할 부여(He, 2024), 맥락 정보(Peng et al., 2023), 모델의 논리적 추론 과정을 유도하는 CoT(Chain-of-Thought) 프롬프트 (Zanina-Seck & Groener, 2025), 역할, 목표, 제약을 지시하는 RGB(Akinwale, 2024) 프롬프트, 맥락, 동작, 역할, 예시를 제공하는 CARE 프롬프트 (Akinwale, 2023) 등 다양한 프롬프트 구조가 LLM의 과업 수행에 도움이 되는 것으로 알려져 있다. 그러나 프롬프트에 너무 복잡한 번역 지침이 포함되는 경우 오히려 번역 품질이 저해된다는 연구 결과도 존재한다(He, 2024). 또한 다양한 프롬프트에 대한 실증적 분석 결과는 언어 쌍, 텍스트의 도메인, LLM 모델 및 버전에 따라 상이하며 특히 한국어와 같은 저자원 언어의 경우 고자원 언어와 양상이 다를 수 있으므로, 다양한 변수에서 프롬프트 효과를 실험하는 연구가 활발히 이루어지고 있다.

상기 논의를 종합할 때, 최근 비약적인 발전을 거듭하고 있는 LLM 기반 번역은 기존 NMT가 보여준 문학 번역의 한계를 보완하고 창의적 표현 가능성을 확장할 잠재력을 제시하고 있다. 그러나 기존 연구들은 대체로 인간 번역을 기준점(baseline)으로 설정하고, 기계번역의 품질이 이에 얼마나 근접하는지 평가하는 데 초점을 맞추어 왔다. 이러한 접근은 번역 결과의 우열을 비교하는 데에는 유의미하나, LLM이 창의성 지시를 받을 때 내부적으로 어떠한 언어적 선택과 변이를 산출하는지에 대한 이해를 충분히 제공하지 못한다. 특히 프롬프트 전략과 온도와 같은 생성 매개변수가 결합될 때,

LLM이 어떠한 방식으로 어휘적 구문적 변이를 확대하는지는 아직 체계적으로 분석되지 않았다.

따라서 본고는 인간 번역과의 단순 비교를 넘어, 프롬프트와 온도 변수를 통해 창의적 번역을 유도할 때 LLM이 어떠한 언어적 반응을 보이는지를 계량적으로 규명하는 데 목적을 둔다. 이를 통해 창의성 유도에 대한 LLM의 반응 양상에 대한 이해를 깊이 하고, 문학 번역에서 매개변수 설계가 갖는 의미를 실증적으로 고찰하고자 한다.

2.3 번역의 창의성 분석 범주

본고에서는 기존 연구(Boden, 2004; Guerberof-Arenas & Toral, 2022)의 창의성 개념을 차용하여 LLM 번역에서의 창의성을 “원문의 의미를 충실하게 보존하면서도(Acceptability), 관습적인 번역 선택에서 벗어나 새로운 어휘적 구문적 변이를 구현하는(Novelty) 능력”으로 조작적 정의를 내린다. 본 연구는 인간 번역을 기준점으로 삼는 기존 연구와 달리 창의성 지시가 배제된 LLM 번역을 기준점으로 설정하여 창의성 프롬프트와 온도 설정 변경에 따른 모델의 반응 양상을 다음과 같이 고찰한다.

첫째, 번역의 참신성은 창의성 지시에 따른 LLM 번역의 어휘적 구문적 변이 정도를 통해 측정한다. 구체적인 지표로는 우선 어휘적 변이를 탐지하기 위하여 문장 간 유사도 측정에 최적화된 BERTScore(Zhang et al., 2020)와 Bleu(Papineni et al., 2002)를 역으로 활용하여 LLM 번역이 기준 번역에서 얼마나 이탈하는지를 측정한다. 어휘 다양성의 변화는 타입토큰비율(TTR), 원문의 길이에 영향을 받지 않는 MATTR(moving average type-token ratio)을 통해 고찰한다. 또한 코퍼스 기반 문체 변이를 탐지하는 지표로 널리 활용되어 온 명사화 비율, 수동태 비율과 기능어와 내용어 비중을 측정한다(Biber et al., 1998). 구문적 측면에서는 문장 구조의 복잡도를 측정한다. 이를 위해 부사절 빈도(advcl)와 종속절 활용 양상, 평균 문장 길이, 그리고 단어 간 문법적 연결 거리를 나타내는 평균의존거리(MDD; Liu, 2008)를 분석하여 원문 대비 구문적 재구성 양상을 파악한다.

둘째, 번역의 수용성은 창의적 시도가 원문의 의미를 훼손하지 않고 번역으로서의 가치를 유지하고 있는지를 의미한다. 본고에서는 이를 원문-번

역문 간의 의미적 유사성으로 간주한다. 이를 검증하기 위해 기준 번역 없이(reference-free) 한국어 원문과 영어 번역문을 직접 대조하는 자동화 평가를 수행한다. 구체적으로 다국어 처리에 특화된 mBERT 기반의 BERTScore를 활용하는데, 이러한 방식은 특히 문체적 변이가 중요한 문학 번역 평가에서 특정 참조 번역에 의존하는 방식보다 더 안정적이고 우수한 성능을 보이는 것으로 입증된 바 있다(Park & Padó, 2024). 물론 인간 평가를 통한 종합적 품질 판단은 문학 번역의 미학적 문화적 적합성을 평가하는데 중요하겠지만 본고의 목적은 번역의 절대적 우열을 가리는 데 있지 않고, 창의성 유도 조건에서 LLM의 의미 보전 범위가 어떻게 변동하는지를 계량적으로 추적하는 데 있다. 인간 평가는 해석적 판단과 평가자 간 변이를 수반할 수 있는 반면, 본 연구는 조건 간 의미적 이동의 상대적 변화를 일관된 기준으로 측정하는 데 초점을 둔다. 따라서 mBERT 기반 BERTScore는 절대적 품질 지표가 아니라, 창의적 발산 상황에서 의미 변화의 움직임을 추적하기 위한 보조적 지표로 활용된다.

이와 같은 분석 틀을 바탕으로, 본 연구는 한국 문학 단편 소설 번역 사례를 통해 창의성 관련 설정의 변화가 LLM의 출력물에서 야기하는 어휘적 구문적 변이를 정량적으로 탐색한다. 이는 단순히 번역의 우열을 가리는 것을 넘어 LLM을 외부 명령에 따라 언어적 층위를 스스로 재구성하는 동적 시스템으로 간주하고 그 메커니즘을 탐색한다는 점에서 기존 연구와 차별화된다. 궁극적으로 본 연구는 LLM 번역의 창의성 발현 특성에 대한 깊이 있는 이해를 제고하고, 향후 AI 번역의 창의성 평가 및 제어 가능성에 대한 분석적 토대를 마련하고자 한다.

3. 연구 방법

3.1 분석 텍스트

본고의 분석 대상 텍스트는 이상의 단편 소설『날개』(1936)이다. 한국 근대 모더니즘의 대표작이라 할 수 있는 이 작품은 식민지 시대 한 지식인

의 분열된 자아를 의식의 흐름과 고도의 상징을 담은 독특한 문체로 탐구하고 있어, LLM 모델 출력의 창의성을 배가하는 온도 및 프롬프트에 따라 문학적 비유와 뉘앙스가 어떻게 재구성되는지 관찰하기에 적합한 텍스트라고 판단하였다.

원작은 총 5,874 토큰(형태소 분리되지 않은 띄어쓰기 기준)으로, 이를 총 55개의 텍스트 단위로 나눈 후 LLM 번역을 수행하였다. 다양한 번역 단위에 대한 번역 생성 결과, 창의적인 번역을 요구한 프롬프트 및 온도에서 번역 이외의 텍스트가 출력되는 현상이 일부 발생하여, 문맥 단절을 최소화하면서도 번역 출력을 얻을 수 있는 단위로 조정한 결과임을 밝힌다.

3.2 번역 생성 절차

표 1

실험 조건 및 사용 패키지(2026.1.15~2.14)

종류	내용
번역 생성 조건	OpenAI GPT-4o, temperature=0.2, 0.7, 1.2, seed=42, top-p=1
데이터 분석	Python (3.12), Stanza (1.11.0), SacreBLEU (2.6.0), BERTScore (0.3.13, bert-base-multilingual-cased, roberta-large), Pandas (2.2.2), NumPy (2.0.2)
통계 분석	Statsmodels (0.14.6), SciPy (1.16.3), Pingouin (0.5.5), Sklearn (1.4.0)
시각화	Matplotlib (3.10.0), Seaborn (0.13.2)

본 연구에서는 파이썬(Python)을 통한 API(Application Programming Interface) 호출 방식으로 LLM 번역을 생성하였다. 일반 사용자 환경인 웹 인터페이스는 사용자의 이전 대화 이력, 계정별 개인 설정, 웹 세션에 축적된 상태 정보를 지속적으로 반영하기 때문에 연구자가 의도하지 않은 데이터의 오염이나 비결정적 편향이 개입될 여지가 있어 연구 데이터로 활용하기에는 재현성과 신뢰성 측면에서 한계를 지닌다. 따라서 본 연구는 이러한 외부 요인을 효과적으로 차단할 수 있도록 모델, 입력 텍스트, 프롬프트, 온도, top-p(=1), 난수 생성 시드(=42)를 명시적으로 고정하는 API 호출을 통해 모델의 순수한 언어적 반응만을 정밀하게 포착하고 재현성을 확보하고자

하였다. 번역 생성에 관련된 변수와 API 호출 및 결과 분석에 사용된 Python 패키지 및 버전 정보는 <표 1>과 같다. 온도와 유사하게 출력 단어 범위 선택과 관련된 또 다른 매개변수인 top-p는 기본값으로 설정했는데, 이는 온도 설정을 조정하는 경우 top-p를 함께 조정하는 것은 권장하지 않는 OpenAI의 가이드라인(OpenAI, n.d.)에 따른 것이다.

본고에서는 창의적 번역을 지시하기 위해 세 가지 온도(0.2, 0.7. 1.2), 네 종류의 프롬프트(P1, P2, P3, P4)를 조합한 총 12가지 조건에서 번역을 생성하였다. P1은 별도의 지시가 없는 제로샷, P2는 창의적 번역 지침을 제시하였으며, P3는 역할, 목표, 제약(RGB), P4는 맥락, 행동, 역할, 예시(CARE) 프롬프트로 구성하였다. 공통 제약 조건(CONSTRAINT)으로는 원문 텍스트의 다음 줄거리를 예측하여 확장하거나 번역 선택에 대한 설명 등을 덧붙이지 않아야 한다는 내용을 추가하였다.

또한 P2, P3, P4에서는 창의적인 번역에 대해 연구자가 작성한 학술적 정의(ACADEMIC_DEF)를 추가하였으며 P4의 경우 창의적인 번역의 예로 이효석(1936)의 『메밀꽃 필 무렵』의 번역(Yi, 2023) 중 도착어의 관용적 표현을 활용한 번역, 번역가의 맥락적 해석이 반영된 번역, 리듬을 살리기 위해 구문 상의 변화를 도모한 번역 세 문장을 제시하였다(EXAMPLES)(<부록 1> 참조).

3.3 분석 및 통계 검정

본 연구 결과의 언어적 특성을 다각도로 분석하고 통계적 유의성을 검정하기 위해 다음과 같은 절차를 수행하였다.

첫째, 산출된 개별 어휘 및 구문 지표들을 대상으로 기술통계 분석 및 분산분석(ANOVA)을 실시하여 프롬프트와 온도 변화에 따라 유의미한 변동을 보이는 지표를 1차적으로 선별하였다. 분석 대상 지표 간의 다중공선성(Multicollinearity) 문제를 방지하기 위해 VIF(Variance Inflation Factor) 지수를 산출하여 변수의 적합성을 검증하고 이를 통해 각 변수가 독립적이면서도 고유한 언어적 정보를 유지하도록 설계하였다. 선별된 지표들을 대상으로 주성분분석(PCA)을 실시하여 다차원적인 언어적 변이를 소수의 주성분으로 축소하였다(Biber, 1988).

둘째, 해당 주성분(PC1, PC2)에 대해서는 변동성과 프롬프트 및 온도의 효과를 검정하기 위해 선형혼합효과모형(linear mixed-effects model, LMM)(Baayen et al., 2008)을 사용한다. 이는 동일한 원문 문장에 대해 프롬프트 전략과 온도 설정의 조합(P1~P4 x T0.2~T1.2)에 따른 번역이 반복 측정되는 데이터를 고려하여 문장별 고유 특성을 통제하면서 매개변수 변화에 따른 변이의 효과를 추정하기 위한 것이다. 모형에서는 원문 문장 ID를 랜덤효과(random intercept)로 설정하였으며, 프롬프트 종류와 온도 설정, 그리고 두 변수 간의 상호작용(interaction)을 고정효과(fixed effect)로 설정하였다. 이를 통해 프롬프트의 지시 효과가 온도라는 확률적 변수와 결합했을 때 나타나는 비선형적 변화를 포착하고자 하였다.

고정효과의 기준(reference)은 번역의 기준점인 제로샷(zero-shot)(P1) 및 온도 0.2(T0.2)의 조합(P1_T0.2)으로 설정하여, 창의적 프롬프트와 온도 상승이 기준점 대비 어떠한 구문적, 어휘적 변이를 유도하는지 추정하였다. 모형의 추정은 제한최대우도법(REML)으로 수행하였으며, 고정효과의 유의성은 Wald F-검정(또는 Likelihood Ratio Test)으로 평가하였다. 전역효과가 유의할 경우, Bonferroni 보정을 적용한 사후비교(pairwise contrast)를 통해 기준점과 각 번역 간의 차이를 확인하였다.

4. 분석 결과

본 연구에서 설정한 세 가지 온도(T02, T07, T12) 및 네 가지 프롬프트(P1, P2, P3, P4)의 조합으로 총 12종의 번역을 생성한 다음 총 20개 분석 범주로 언어적 특성을 분석한 결과, 온도가 상승함에 따라 기준 번역(P1_T0.2)으로부터 이탈하여 언어적 복잡성이 증가하는 경향이 관찰되었다. 상세한 기술통계 수치는 <부록 2>에 제시하였다.

주성분 분석(PCA) 및 혼합효과모형(LMM) 검정에 앞서, 각 지표의 타당성을 검토하기 위해 다중공선성 및 프롬프트와 온도별로 각 변수의 유의성을 분석하였다. 유의성 분석 결과 총 12개의 변수(BLEU, RIBES_Lemma, BERTScore_vs_Baseline, avg_sent_len, token_count, TTR, MATTR,

Content_Word_Count, Function_Word_Count, Passive_Ratio, Nominalization_Ratio, Nominalization_Count)가 후보로 도출되었다. 이 가운데 BLEU와 RIBES_Lemma는 모두 기준 번역과의 표층적 유사도를 반영하는 지표로서 BERTScore_vs_Baseline만을 채택하였다. 또한 어휘 다양성 지표 token_count, TTR, MATTR 중 텍스트 길이의 영향을 상대적으로 덜 받는 MATTR만을 유지하였다. 또한 텍스트 길이의 영향을 받는 절대지표(Nominalization_Count, Content_Word_Count, Function_Word_Count)를 제외하였다. 반면 BERTScore_F1은 통계적으로 유의한 차이가 나타나지 않았으나 원문과 번역문 간 의미 보존 정도를 직접적으로 반영하는 핵심 지표이므로 분석 변수에 포함하였다. 평균 의존거리(mdd) 역시 유의 후보 변수 수준이었으나 VIF 지수가 낮아 독립적인 구문 복잡성 정보를 제공하는 변수로 판단되어 함께 포함하였다.

이러한 과정을 거쳐 최종적으로 원문 대비 의미 변화(BERTScore_F1), 기준 번역 대비 의미 변화(BERTScore_vs_Baseline), 어휘 다양성(MATTR), 평균 문장 길이(avg_sent_len), 구문 복잡도를 의미하는 평균 의존거리(mdd), 명사화 비율(Nominalization_Ratio), 수동태 비율(Passive_Ratio)의 총 7개 변수가 PCA 분석에 투입되었다. 이들 지표의 VIF 지수는 1.09~2.34 범위로, 통계적 수용 한계치인 5 미만(Hair et al., 2019)을 기록하였다(<표 2> 참조).

표 2

PCA 분석 변수

변수명	통계적 유의성 (p-value)	VIF 지수
BERTScore_F1	0.304	1.14
BERTScore_vs_Baseline	< 0.001***	1.44
MATTR	< 0.001***	1.29
avg_sent_len	< 0.001***	2.34
mdd	유의미 변수 후보	2.23
Nominalization_Ratio	0.027*	1.21
Passive_Ratio	0.035*	1.09

주목할 점은 수용성 지표인 원문과 번역문 사이의 의미적 유사성을 의미하는 BERTScore_F1의 안정성으로, 모든 실험 조건에서 0.70 내외의 수치를 유지하였으며 통계적으로도 유의미한 차이가 나타나지 않았다($p = .304$). 이는 창의성 지시 하에서도 최소한의 품질을 유지했다는 것을 시사한다.

반면, 참신성 지표에서는 실험 조건에 따라 뚜렷한 변화가 감지되었다. 어휘 측면에서는 온도가 상승함에 따라 기준점 대비 유사도(BERTScore_vs_Baseline)가 1.00에서 0.92로 하락하고($p < .001$), 어휘 다양성(MATTR)은 0.88에서 0.91로 상승하여($p < .001$) 관습적 선택을 탈피한 풍부한 표현이 생성되었음을 확인할 수 있었다. 또한 명사화 및 수동태 비율의 유의미한 증가는 높은 온도에서 문체 재구성이 이루어졌음을 짐작케 했다. 구문적 측면에서도 역시 평균 문장 길이(avg_sent_len)가 14.91에서 16.85까지 길어지고($p < .001$), 평균의존거리(mdd)가 변동하는 등 구조적 복잡성이 심화된 것으로 나타났다. 이러한 결과는 번역의 수용성이 크게 저해되지 않는 범위 내에서 온도와 프롬프트의 조절을 통해 어휘와 구문 양측의 창의적 변이가 나타났음을 암시하는 것이라 볼 수 있다.

프롬프트 전략에 따른 차이를 분석한 결과, P2(지시형)가 높은 온도 설정과 결합할 때 문장 길이와 어휘 다양성이 극대화되며 가장 공격적인 변이가 나타난 반면, 구조화된 프롬프트인 P3(RGB)와 P4(CARE)에서는 온도 상승에 따른 이탈 폭이 P2에 비해 상대적으로 완만한 것으로 나타났다. 이는 퓨샷 등 상세한 프롬프트 설정이 모델의 무작위적 확산을 억제하는 기제로 작용하여 출력 일관성 유지에 기여했다는 해석이 가능하다.

4.1 PCA 분석

개별 지표들 간의 상관성을 바탕으로 LLM 번역의 변이 양상을 결정짓는 핵심 요인을 식별하기 위해 주성분 분석(PCA)을 실시한 결과 상위 2개의 주성분(PC1, PC2)이 전체 변량의 50.98%를 설명하는 것으로 나타나 데이터의 차원 축소가 효과적으로 이루어졌음을 확인하였다(<표 3> 참조).

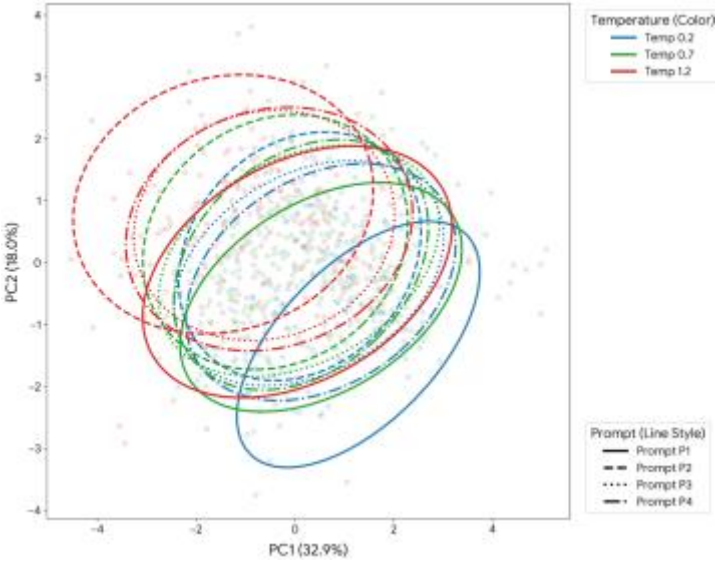
표 3

PCA 결과

	PC1	PC2
BERTScore_vs_Baseline	0.411537	-0.50346
BERTScore_F1	0.331032	0.135152
MATTR	-0.32284	0.506697
avg_sent_len	-0.5204	-0.3786
mdd	-0.4748	-0.48854
Nominalization_Ratio	-0.3363	0.277129
Passive_Ratio	-0.08644	0.112764
Explained Variance (%)	32.95%	18.04%
Cumulative Variance (%)	32.95%	50.98%

그림 1

온도/프롬프트별 번역에 대한 PCA 분석 결과



우선 전체 변동의 32.95%를 설명하는 PC1은 기준점 수렴 및 구문적 단순화 축으로 정의할 수 있다. PC1의 평균 문장 길이(avg_sent_len, -0.520), 평

균의존거리(mdd, -0.475)에서 강한 음의(negative) 부하량이 나타난 반면 기준 번역과의 유사도(BERTScore_vs_Baseline, 0.412)와는 양의(positive) 상관관계가 관찰되었다. 이는 PC1 값이 작아질수록 구문 복잡성이 증가하며 기준 번역에서 이탈한다는 것을 의미한다. 즉, PC1에서 음의 방향은 번역의 구문적 참신성이 증대되는 지표로 볼 수 있다.

PC2는 어휘적 다양성 및 참신성으로 정의할 수 있으며, 18.04%의 설명력을 가진다. PC2는 어휘 다양성 지표인 MATTR(0.507)에서 가장 높은 양의 부하량을, 기준 번역과의 유사도(BERTScore_vs_Baseline, -0.503)에서 강한 음의 상관관계가 나타났다. 이는 PC2의 값이 커질수록 기준 번역의 어휘 선택에서 탈피하여 더욱 풍부하고 다양한 어휘가 사용되었음을 시사한다. 따라서 PC2는 모델이 창의적 지시에 반응하여 구현하는 어휘적 참신성의 정도를 측정하는 축으로 활용될 수 있다.

<그림 1>은 이러한 두 주성분을 축으로 실험 조건별 번역문의 분포를 시각화한 결과로, 프롬프트와 온도 설정에 따라 번역 양상이 뚜렷하게 군집화되는 경향이 관찰되었다. 특히 온도 상승에 따라 데이터 포인트가 PC1 축의 음의 방향과 PC2 축의 양의 방향으로 이동하는 양상이 두드러졌는데, 이는 확률적 변수의 증가가 구문과 어휘 양측에서 참신성 생성을 가속화하는 핵심 동인임을 뒷받침한다.

한 가지 주목할만한 것은 수용성 지표인 BERTScore_F1(0.331)과 PC1의 양의 상관관계이다. 이는 번역 품질이 기준 번역과 어느 정도 관계가 있다는 점을 시사한다.

4.2 주성분에 대한 LMM 분석

추출된 두 주성분을 종속변수로 하여 프롬프트 전략과 온도 설정의 효과를 검정한 LMM 분석 결과는 <표 4>와 같다.

우선 PC1에 대한 분석 결과, 모든 프롬프트(P2~P4)와 온도 상승(T0.7, T1.2)은 기준점(P1_T0.2) 대비 PC1 수치의 유의미한 감소로 이어졌다($p < .001$). 이는 창의성 지시 프롬프트와 온도 상승이 공통적으로 구문적 단순화에서 벗어나 복잡성을 증대시키는 방향으로 작용했음을 의미한다. 특히 프롬프트 중에서는 P2($\beta = -1.184$)가, 온도 중에서는 T1.2($\beta = -1.233$)가 가장

강력한 음(-)의 효과를 보이며 구문적 이탈을 주도하였다. 상호작용 효과에서는 P3와 온도의 결합이 PC1 수치를 다시 상승시키는 양(+)의 유의성을 보였는데, 이는 특정 프롬프트 전략이 고온에서의 과도한 구문적 발산을 제어하는 기제로 작용할 수 있음을 시사한다.

PC2의 경우 모든 조건에서 기준점 대비 PC2 수치가 유의미하게 증가하였다($p < .001$). 프롬프트 전략 중 P2($\beta = 1.423$)가 어휘 다양성 증가에 대한 기여도가 높았으며, 온도 역시 증가할수록 어휘적 참신성이 선형적으로 증가하였다. 주목할 점은 상호작용 효과에서 나타나는 모든 계수가 음(-)의 유의성을 보인다는 것이다. 이는 온도 상승에 따른 어휘 다양성의 증가 폭이 기준점인 P1(기준점)에서 가장 가파르며, 이미 창의적 지시가 포함된 P2~P4 조건에서는 온도 상승에 따른 추가적인 어휘 변이의 폭이 상대적으로 완만해짐을 의미한다.

표 4

LMM 분석 결과

구분	변수	PC1	PC2
-	-	Estimate (SE)	Estimate (SE)
고정 효과	(Intercept)	1.285 (0.186)***	-1.320 (0.132)***
[Prompt] (Ref: P1)	Prompt P2	-1.184 (0.113)***	1.423 (0.108)***
	Prompt P3	-1.001 (0.113)***	1.151 (0.108)***
	Prompt P4	-0.760 (0.113)***	1.006 (0.108)***
[Temp] (Ref: 0.2)	Temp 0.7	-0.755 (0.113)***	0.762 (0.108)***
	Temp 1.2	-1.233 (0.113)***	1.171 (0.108)***
[Interaction]	P2 × Temp 0.7	0.261 (0.160)	-0.535 (0.152)***
	P3 × Temp 0.7	0.554 (0.160)***	-0.564 (0.152)***
	P4 × Temp 0.7	0.355 (0.160)*	-0.483 (0.152)***
	P2 × Temp 1.2	-0.317 (0.160)*	-0.331 (0.152)*
	P3 × Temp 1.2	0.328 (0.160)*	-0.402 (0.152)***
	P4 × Temp 1.2	0.190 (0.160)	-0.314 (0.152)*
랜덤 효과	Var: SubUnit (Intercept)	1.551	0.639
	Residual Var.	0.354	0.319

모델 적합도	Log-Likelihood	-714.42	-659.99
	AIC	1456.84	1347.98

주: SE=Standard Error. 상호작용의 기준점은 P1_T0.2. 유의성: *p < .05, **p < .01, ***p < .001.

이와 같은 분석 결과는 LLM의 창의적 번역을 유도하려면 온도 조절만으로는 한계가 있으며, 특정 프롬프트 전략이 특정 온도 설정과 결합할 때 비로소 그 양상이 극대화되거나 통제될 수 있다는 것을 의미한다. 특히 P2와 높은 온도의 결합은 구문과 어휘 양측에서 가장 파괴적인 참신성을 보인 반면 P3와 P4에서는 높은 온도에서도 번역의 구조적 틀이 일정 부분 유지되는 등, 정제된 창의성을 발현하는 동적인 반응 체계가 관찰되었다.

4.3 질적 분석

본 절에서는 양적 분석 결과가 실제 번역문의 언어적 층위에서 어떻게 실증되는지 고찰한다. 분석 결과, 높은 온도와 창의성 프롬프트의 결합은 단순한 어휘 교체를 넘어 문체의 정교화와 구문 구조의 재구성을 유도하는 것으로 나타났다.

<예시 1> D-ID06

원문: 여러 번 자동차에 치일 뻔하면서 나는 그래도 경성역으로 찾아갔다. 빈자리와 마주 앉아서 이 쓰디쓴 입맛을 거두기 위하여 무엇으로나 입가심을 하고 싶었다. 커피! 좋다. 그러나 경성역 홀에 한 걸음 들여놓았을 때 나는 내 주머니에는 돈이 한 푼도 없는 것을 그것을 깜박 잊었던 것을 깨달았다.

Baseline (P1-T0.2): Several times I almost got hit by a car, but I still made my way to Gyeongseong Station. Sitting across from an empty seat, I wanted to cleanse my palate with something to rid myself of this bitter taste. Coffee! That would be good. However, when I stepped into the hall of Gyeongseong Station, I realized that I had completely forgotten that I didn't have a single penny in my pocket.

P2-T1.2: After narrowly avoiding several collisions with cars, I still made my way to Gyeongseong Station. Faced with an empty seat, I yearned to rinse away this bitter taste lingering in my mouth with something—

anything. Coffee! That would be perfect. But as I set foot inside the station's grand hall, I was jolted by the realization that my pockets were completely void of money, a fact I had momentarily forgotten.

첫째, 어휘적 측면에서의 참신성(PC2)은 저빈도 어휘의 선택과 문학적 비유의 강화로 구체화되었다. <예시 1>에서는 기준 번역에서 “completely forgotten”이나 “I didn’t have a single penny in my pocket” 등 일상적 표현을 선택한 것과 대조적으로, 높은 온도(T1.2)가 창의성 프롬프트(P2)와 결합되자 “jolted by the realization,” “completely void of money”와 같은 출현 빈도가 상대적으로 낮은 어휘가 사용되었다. 이러한 변화는 어휘 다양성 지표인 MATTR의 상승과 기준점 이탈도(BERTScore_vs_Baseline)의 증가가 실제 번역 산출에서 어휘 선택의 다양화와 문체적 변이로 이어졌음을 보여준다.

특히 저빈도 어휘의 사용은 창의성 프롬프트와 높은 온도 설정 하에서 전반적으로 나타났다. “해가 들지 않는다”는 “Sunlight does not enter”(P1_T0.2)에서 “Sunlight shies away”(P2_T1.2)로, “낮잠”은 “take naps”(P1_T0.2)에서 “take refuge in daytime slumbers”(P2_T1.2)로, “꾸지람”은 “scolding”(P1_T0.2)에서 “rebukes”(P2_T1.2)와 같이 의인화, 명사화, 격상된 어휘로 바뀌는 것을 찾아볼 수 있었다. 또 다른 예에서는 밤이 찾아오는 장면을 “nocturne descends”(B-ID01, T1.2-P2)로 묘사하여, 원문의 핵심 의미는 유지한 가운데 창의적 변이가 시도된 것을 발견할 수 있었다.

<예시 2> C-ID03

원문: 나는 벽에 기대어 앉아서 담배를 한 대 피워 물고 그리고 벼락이 떨어질 테거든 어서 떨어져라 하고 기다렸다. 오 분! 십 분! 그러나 벼락은 내리지 않았다. 긴장이 차츰 풀어지기 시작한다. [...] 걸을 때는 몰랐더니 숨이 차다. (1) 등에 식은땀이 쭉 내배인다. (2) [...] 우리 부부는 이야기하는 법이 없었다. (3) 밥을 먹은 뒤에도 나는 말이 없이 부스스 일어나서 내 방으로 건너가 버렸다. (4) 아내는 나를 붙잡지 않았다. (5)

Baseline(P1-T0.2): I sat leaning against the wall, lit a cigarette, and waited, thinking if lightning was going to strike, let it strike soon. Five minutes! Ten minutes! But the lightning didn't strike. The tension gradually began to

ease. [...] I didn't notice it while walking, but I was out of breath. (1) Cold sweat trickled down my back. (2) [...] We, as a couple, never talked. (3) Even after eating, I got up silently and went over to my room. (4) My wife didn't stop me (5).

P2-T1.2: There, leaning against the wall, I lit a cigarette and waited, daring the thunderbolt to strike if it ever would. Five minutes. Ten. Yet, lightning refused to fall. Tension slowly unraveled. [...] All seemed calm while walking, but now each breath was labored, cold sweat tracing down my back. (1) [...] Conversation wasn't our practice; we dined quietly as ever, after which I rose in silence and shuffled to my room without a word, she never stopping me. (2)

둘째, 구문적 측면에서의 복잡성(PC1)은 문장 통합 및 구조적 재배치를 통해 나타났다. <예시 2>는 양적 분석에서 나타난 평균 문장 길이(avg_sent_len)의 증가가 실제 어떠한 구문적 변화를 수반하는지 보여준다. <예시 2>에서는 원문에서 마침표로 구분되어 있던 두 문장이 낮은 온도(0.2)에서는 원문과 동일한 문장 구분으로 처리되었지만, 높은 온도와 창의성 프롬프트에서는 관계절이나 접속사를 통해 분절된 문장들이 하나의 긴 호흡으로 통합된 것을 관찰할 수 있었다. 이는 온도가 상승함에 따라 모델이 문장 간의 논리적 연결성을 능동적으로 재구성한다는 것을 시사한다.

셋째, 문화소 번역 전략의 변화는 수용성과 참신성 사이의 균형을 보여준다. <예시 3>과 같이 낮은 온도에서의 단순 음차 전략은 높은 온도와 정교한 프롬프트(P4) 하에서 원문의 정보를 명시화하는 전략으로 전환되었다.

<예시 3> D-ID06

원문: [...] 내가 미쓰꼬시 옥상에 있는 것을 깨달았을 때는 거의 대낮이었다.

Baseline (P1-T0.2): [...] when I realized I was on the rooftop of Mitsukoshi,

P2-T1.2: [...] did I come to my senses on the rooftop of Mitsukoshi, almost washed by the midday light.

P4-T1.2: [...] I found myself on the Mitsukoshi Department Store's roof garden

이는 <표 2>에서 수용성 지표인 BERTScore_F1이 유의미하게 하락하지 않았던 결과와 연결된다. 모델은 온도가 상승함에 따라 기준점으로부터 이탈하되, 독자의 이해도를 고려한 대안적 번역을 제시함으로써 참신성과 번역의 품질을 동시에 확보하는 양상을 보였다.

5. 결론

본 연구는 생성형 AI의 비결정론적 특성을 활용하여 문학 번역에서 창의적 변이를 유도할 수 있는 최적의 프롬프트 전략과 온도 설정의 조합을 탐색하였다. 연구 결과를 요약하면 다음과 같다.

첫째, 양적 분석 결과 LLM은 프롬프트의 지시 강도와 온도가 높아질수록 기준점에서 유의미하게 이탈하며 어휘적 다양성(PC2)과 구문적 복잡성(PC1)을 확보하는 것으로 나타났다. 특히 지시형 프롬프트(P2)와 높은 온도 설정(T1.2)의 결합은 가장 강력한 참신성을 유도하는 것으로 나타났다. 그러나 역할, 목표, 제약을 제시한 프롬프트(P3)와 CARE 프롬프트(P4)는 오히려 모델의 출력을 특정 범위 내로 수렴시키며 창의적 발산을 제어한 것으로 나타났다.

이러한 결과는 복잡한 프롬프트가 번역 성능을 저해한다는 선행연구(He, 2024)와 일치하는 결과로도 볼수 있지만, 높은 온도 설정에서 알려지는 것으로 나타나는 의미 이탈이나 과도한 서사적 확장이 P3, P4 프롬프트로 인해 일정 범위 내로 조정되었을 가능성을 시사한다. 실제로 높은 온도 조건에서도 BERTScore_F1이 비교적 안정적으로 유지된 점은, 제약적 프롬프트가 의미의 핵심을 보존하는 안전 장치로 기능했음을 시사한다. 즉, 어휘적 구문적 변이를 확대하면서도 의미 보전을 일정 수준 유지하는, 수용성과 참신성 사이의 균형 지점이 형성되었을 가능성을 보여준다.

그럼에도 불구하고, 이러한 수렴 효과가 모델의 본질적 특성이라기보다 프롬프트 설계 방식에 기인했을 가능성 역시 배제할 수 없다. 창의적 번역의 예시를 제시하고 이를 참조하도록 요구하는 방식은 모델의 탐색 공간을 구조적으로 제한하며, 결과적으로 특정 문체적 패턴을 모방하는 방향으로

출력을 유도했을 수 있다. 또한 고온 조건에서의 과도한 확장을 방지하기 위해 추가한 번역 행위에 대한 통제 지시 역시 발산적 창의성을 제약했을 가능성이 있다. 따라서 본 연구에서 관찰된 구조화 프롬프트에서의 수렴 현상은 창의성 억제와 의미 안정화라는 두 가능성을 동시에 내포하며, 프롬프트 설계 변인을 보다 정교하게 분리 통제한 후속 연구를 통해 추가적으로 검증될 필요가 있다.

둘째, 질적 분석을 통해 이러한 양적 변이가 실제 문체적 재구성의 가시화로 이어진다는 것을 확인하였다. 높은 온도 설정 하에서 모델은 저빈도 어휘를 선택하거나 분절된 문장을 통합하여 복잡한 구문 구조를 재구성하였으며, 문화소 번역에 있어서도 단순 음차를 넘어 정보 명시화 및 의역 전략을 구사하였다. 주목할 점은 이러한 참신성의 증대에도 불구하고 수용성 지표가 비교적 안정적으로 유지되었다는 사실이며, 이는 LLM이 번역 품질을 유지하는 임계점 내에서 문체를 재구성할 수 있다는 점을 시사한다.

이는 NMT가 번역가의 창의적 선택에 제약으로 작용할 수 있다는 선행 논의(예: Guerberof-Arenas & Toral, 2022)와 대비된다. 본 연구 결과는 매개변수 조절을 통해 창의적 변이를 확장한 LLM 번역이 단순한 자동화 도구를 넘어, 인간 번역가의 표현 탐색을 지원하는 발판으로 기능할 수 있음을 보여준다. 다시 말해, LLM 번역은 완결된 대체물이 아니라, 번역가가 어휘적 문체적 가능성을 실험하고 확장하는 과정에서 참조 가능한 발판이자 사고의 보조장치로 활용될 수 있음을 의미한다. 이러한 점에서 본 연구는 LLM 문학 번역이 인간과 대적하기보다 인간과의 협업적 생산 구조 속에 통합될 가능성을 암시한다.

셋째, LLM의 창의적 발산이 지닌 위험성 또한 관찰되었다. <예시 4>에서 확인되듯, 창의성 지시와 고온의 결합은 원문에 대한 초기 의미 해석에서 오독이 발생한 경우, 이를 보완하려는 과정에서 원문에 존재하지 않는 논리적 개연성을 모델이 스스로 구성하는 환각 현상을 야기할 수 있다. 이는 LLM과 인간 번역가와의 협업 가능성을 보여주는 동시에 과번역 및 오역에 대한 비판적 검토가 상시 동반되어야 함을 시사한다. 또한 정확히 어떠한 온도와 프롬프트에서 이러한 현상이 나타나는지에 대해서는 본고의 논의를 넘어 후속 연구가 필요하다고 사료된다.

<예시 4> ID499
원문: 이 18가구를 대표하는 대문이라는 것이 일각이 져서 외따로 떨어지기는 했으나, 있다.
Baseline (P1-T0.2): The gate representing these 18 households is slightly tilted and stands apart, but it exists.
P2-T1.2: The main gate... leans somewhat due to a broken hinge, yet it stands.

본 연구는 인간 평가를 통한 주관적 품질 검증을 수행하지 못했다는 한계를 지니나, 창의성 지시에 따른 LLM 출력의 변화를 객관적 지표를 통해 정량적으로 탐색했다는 점에서 의의를 찾고자 한다. 특히 창의적 번역에 대한 학술적 합의가 부재한 상황에서, 기존 번역 연구에서 효과적인 것으로 알려진 퓨샷 등 구조화된 프롬프트 전략이 창의적 변이 측면에서는 오히려 출력을 제약할 수 있다는 가능성을 제시하였다.

마지막으로 본 연구는 LLM 번역 연구의 방법론적 엄정성 확보를 강조한다. 본고는 세션 히스토리 누적 및 샘플링 변이를 통제하기 위한 절차를 마련함으로써 재현성을 확보하고자 하였다. 향후 번역학 분야에서 AI 번역을 본격적인 연구 대상으로 다루기 위해서는 번역 생성 환경에서의 초기화 여부, 시드 고정, 호출 환경의 명시화 등 실험 절차에 대한 엄격한 보고 기준이 마련되어야 한다. 이러한 절차적 정당성은 AI 번역을 둘러싼 학제적 논의에서 번역학의 학문적 정합성과 발언권을 확보하는 중요한 기반이 될 것이다.

참고문헌

<1차 자료>
이상. (1936/2012). 날개. 더플래닛.

<2차 자료>
김현웅, 이상빈. (2025). AI 문학번역에서 프롬프트 엔지니어링이 번역 오류

와 창의성에 미치는 영향. 번역학연구, 26(3), 147-171. https://doi.org/10.15749/JTS.2025.26.3.005

마승혜. (2025). AI 기반 한국문학 번역 전략 탐색 — Google AI Studio의 Temperature 조절과 프롬프트 설계를 중심으로. 번역학연구, 26(4), 229-258. https://doi.org/10.15749/JTS.2025.26.4.008

Abdelhalim, S. M., Alsahil, A. A., & Alsuhaibani, Z. A. (2025). Artificial intelligence tools and literary translation: A comparative investigation of ChatGPT and Google Translate from novice and advanced EFL student translators' perspectives. *Cogent Arts & Humanities, 12*(1), 1-20. https://doi.org/10.1080/23311983.2025.2508031

Ackley, D., Hinton, G., & Sejnowski, T. (1985). A learning algorithm for boltzmann machines. *Cognitive Science, 9*(1), 147-169. https://doi.org/10.1016/S0364-0213(85)80012-4

Akinwale, P. (2023). *AI basics and the RGB prompt engineering model: Empowering AI & ChatGPT through effective prompt engineering*. Praizion Media.

Aranda, L. V. (2009). Forms of creativity in translation. *Cadernos de Tradução, 1*(23), 23-37. https://doi.org/10.5007/2175-7968.2009v1n23p23

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412. https://doi.org/10.1016/j.jml.2007.12.005

Bayer-Hohenwarter, G. (2010). Comparing translational creativity scores of students and professionals: Flexible problem-solving and/or fluent routine behaviour? In S. Göpferich, F. Alves & I. M. Mees (Eds.), *New approaches in translation process research* (pp. 83-111). Samfundslitteratur.

Bayer-Hohenwarter, G. (2011). "Creative shifts" as a means of measuring and promoting translational creativity. *Meta, 56*(3), 663-692. https://doi.org/10.7202/1008339ar

Biber, D. (1988). *Variation across speech and writing*. Cambridge University

Press.

Biber, D., Douglas, B., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Boden, M. A. (2004). *The creative mind: Myths and mechanisms* (2nd ed). Routledge.

Du, S., Arenas, A. G., Toral, A., Gerrits, K., & Borillo, J. M. (2025). *Optimising ChatGPT for creativity in literary translation: A case study from English into Dutch, Chinese, Catalan and Spanish*. arXiv. https://doi.org/10.48550/arXiv.2504.18221

Gao, Y., Wang, R., & Hou, F. (2023). *How to design translation prompts for ChatGPT: An empirical study*. arXiv. http://arxiv.org/abs/2304.02182

Guerberof-Arenas, A., & Toral, A. (2020). The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces, 9*(2), 255–282. https://doi.org/10.1075/ts.20035.gue

Guerberof-Arenas, A., & Toral, A. (2022). Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces, 11*(2), 184–212. https://doi.org/10.1075/ts.21025.gue

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage.

He, S. (2024). Prompting ChatGPT for translation: *A comparative analysis of translation brief and persona prompts*. arXiv. https://arxiv.org/abs/2403.00127

Hu, K., & Li, X. (2023). The creativity and limitations of AI neural machine translation: A corpus-based study of DeepL's English-to-Chinese translation of Shakespeare's plays. *Babel, 69*(4), 546–563. https://doi.org/10.1075/babel.00331.hu

Li, L., Sleem, L., Gentile, N., Nichil, G., & State, R. (2025). *Exploring the impact of temperature on large language models: Hot or cold?* arXiv. https://doi.org/10.48550/arXiv.2506.07295

Liu, H. (2008). Dependency distance as a metric of language comprehension

difficulty. *Journal of Cognitive Science, 9*(2), 159-191. https://doi.org/10.17791/JCS.2008.9.2.159

Manapbayeva, Z., Zaurbekova, G., Ayazbekova, K., Kazezova, A., & Pirmanova, K. (2024). AI in literary translation: ChatGPT-4 vs. professional human translation of Abai's poem "Spring." *Procedia Computer Science, 251*, 526-531. https://doi.org/10.1016/j.procs.2024.11.143

OpenAI. (n.d.). *API reference*. OpenAI. https://platform.openai.com/docs/api-reference/completions/create#completions_create-top_p

O'Sullivan, C. (2013). Creativity. In Y. Gambier & L. Van Doorslaer (Eds.), *Handbook of translation studies* (Vol. 4, pp. 42-46). John Benjamins.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002),* 311-318.

Park, D., & Padó, S. (2024). *Multi-dimensional machine translation evaluation: Model evaluation and resource for Korean*. arXiv. https://doi.org/10.48550/arXiv.2403.12666

Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024). *Is temperature the creativity parameter of large language models?* arXiv. https://doi.org/10.48550/arXiv.2405.00492

Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., & Tao, D. (2023). *Towards making the most of ChatGPT for machine translation* [Preprint]. SSRN. https://doi.org/10.2139/ssrn.4390455

Toral, A., & Way, A. (2014). Is machine translation ready for literature? *Proceedings of Translating and the Computer 36,* 174-176.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned language models are zero-shot learners. *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*, 1-46.

Yi, H. (1936/2023). When the buckwheat blooms (K. Chong-un & B. Fulton,

Trans.) In B. Fulton (Ed.), *The Penguin Book of Korean short stories* (pp. 3-12). Penguin Classics.

Zanina-Seck, A., & Groener, C. U. (2025). The secret power of syntax: Improving ChatGPT translation quality through sentence constituent analysis? In H. Degen & S. Ntoa (Eds.), *Artificial intelligence in HCI* (Vol. 15821, pp. 242–260). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-93418-6_17

Zhang, R., Zhao, W., & Eger, S. (2025). *How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs*. arXiv. https://doi.org/10.48550/arXiv.2410.18697

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore: Evaluating text generation with BERT*. arXiv. https://doi.org/10.48550/arXiv.1904.09675

<부록 1> 번역 생성에 사용된 프롬프트(P1, P2, P3, P4)

종류	내용
학술적 정의	Creative translation is a problem-solving process that generates a solution which is novel (flexible) by departing from the linguistic structure of the source text, yet acceptable by being appropriate for the target audience and situation.
예시	[Example 1: Idiomatic Re-creation] Source: “알 수 있나요. 도무지 듣지를 못했으니까.” Translation: “Beats me. I never heard it mentioned.” Explanation: Uses a natural English idiom instead of a literal rendering to capture the speaker’s blunt and dismissive tone. [Example 2: Contextual Interpretation] Source: 허 생원은 경망하게도 발을 빗디디었다. Translation: And then, distracted, he lost his footing. Explanation: Interprets the adverb ‘경망하게도’ as a psychological state inferred from context rather than translating it literally as ‘carelessly.’ [Example 3: Dynamic Syntactic Restructuring] Source: 앞으로 고꾸라지기가 바쁘게 몸째 풍덩 빠져버렸다. Translation: His body pitched forward, plunging him deep into the stream. Explanation: Restructures the sentence syntactically and employs vivid, active verbs to convey speed, motion, and visual impact.
P1 Baseline	Translate the following Korean text into English. Translate the following Korean text into English.{B_CONSTRAINT}
P2 Instruction	Translate the following Korean text into English creatively based on this definition: {ACADEMIC_DEF}{B_CONSTRAINT}
P3 RGB	Role: You are a professional literary translator. Goal: Perform a creative translation defined as: {ACADEMIC_DEF} Background/Constraint: Follow the strict output rules below.{B_CONSTRAINT}
P4 CARE	Context: We apply the following definition of creative translation: {ACADEMIC_DEF} Action: Translate the Korean text below into English. Role: You are a professional literary translator. Example (Learn from the rationale, but do NOT output explanations) {EXAMPLES_FOR_P4} Instruction: Based on the examples and role above, provide your translation.{B_CONSTRAINT}

<부록 2> 분석 범주에 대한 기술통계

	P1_T0.2	P1_T0.7	P1_T1.2	P2_T0.2	P2_T0.7	P2_T1.2	P3_T0.2	P3_T0.7	P3_T1.2	P4_T0.2	P4_T0.7	P4_T1.2
BLEU_Mean	100.00	62.59	44.93	44.47	35.03	22.87	53.10	44.87	31.55	57.39	48.03	32.82
SD	0.00	-9.32	-10.75	-12.61	-11.24	-10.64	-10.65	-9.56	-10.06	-12.11	-12.60	-11.02
RIBES_Lemma_Mean	1.00	0.87	0.80	0.80	0.75	0.68	0.83	0.80	0.73	0.85	0.82	0.74
SD	0.00	-0.04	-0.06	-0.06	-0.06	-0.07	-0.05	-0.05	-0.06	-0.05	-0.06	-0.07
acl_Mean	1.07	1.06	1.06	1.44	1.20	1.62	1.26	1.15	1.47	1.04	1.13	1.24
SD	-1.05	-1.03	-1.01	-1.20	-0.97	-1.24	-1.32	-1.10	-1.22	-1.12	-1.16	-1.19
acl_relcl_Mean	1.51	1.42	1.31	1.15	1.31	1.26	1.22	1.26	1.35	1.29	1.15	1.15
SD	-1.39	-1.24	-1.25	-1.28	-1.35	-1.14	-1.17	-1.29	-1.28	-1.26	-1.13	-1.24
advcl_Mean	3.73	4.07	4.46	4.15	4.55	5.33	4.20	4.22	4.67	4.07	4.20	4.64
SD	-2.04	-2.26	-2.53	-1.96	-2.55	-2.59	-2.14	-2.32	-2.40	-2.03	-2.37	-2.00
avg_sent_len_Mean	14.93	15.01	15.41	15.01	15.70	16.85	15.11	15.34	15.86	14.91	15.01	15.59
SD	-3.41	-3.36	-3.76	-2.93	-3.12	-3.72	-3.18	-3.54	-2.90	-3.33	-3.28	-3.48
cc_count_Mean	5.09	5.02	4.84	5.24	5.18	4.67	5.36	5.26	4.64	5.31	4.98	5.02
SD	-2.03	-1.84	-2.07	-1.93	-1.87	-1.83	-2.05	-2.09	-1.88	-2.06	-2.01	-1.90
ccomp_Mean	1.47	1.47	1.38	1.22	1.16	1.20	1.36	1.49	1.44	1.35	1.36	1.42
SD	-1.35	-1.40	-1.15	-1.23	-1.32	-1.21	-1.30	-1.39	-1.27	-1.25	-1.19	-1.40
mdd_Mean	2.99	3.02	3.03	2.99	3.03	3.09	2.99	2.99	3.01	2.98	3.01	3.02
SD	-0.25	-0.28	-0.30	-0.26	-0.26	-0.26	-0.25	-0.25	-0.23	-0.25	-0.25	-0.26
num_sentences_Mean	11.86	11.71	11.56	11.33	11.07	10.71	11.71	11.60	11.33	11.60	11.55	11.22
SD	-2.97	-2.87	-2.89	-2.60	-2.52	-2.59	-2.77	-2.77	-2.43	-2.87	-2.85	-2.77
xcomp_Mean	4.06	4.06	3.98	3.93	4.04	3.46	4.09	4.27	4.27	4.11	4.13	4.02
SD	-2.19	-2.42	-2.42	-2.17	-2.27	-1.90	-2.38	-2.54	-2.16	-2.27	-2.57	-2.48
BERTScore_F1_Mean	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
SD	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00
TTR_Mean	0.64	0.65	0.66	0.68	0.68	0.71	0.66	0.67	0.69	0.66	0.67	0.69
SD	-0.04	-0.04	-0.04	-0.05	-0.05	-0.04	-0.05	-0.04	-0.05	-0.05	-0.05	-0.05
MATTR_Mean	0.88	0.88	0.88	0.89	0.90	0.91	0.89	0.89	0.90	0.88	0.89	0.90
SD	-0.03	-0.03	-0.03	-0.02	-0.03	-0.02	-0.02	-0.02	-0.02	-0.03	-0.03	-0.02
Lexical_Density_Mean	0.47	0.47	0.49	0.48	0.48	0.49	0.47	0.48	0.49	0.47	0.48	0.48
SD	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.03	-0.03	-0.04	-0.04	-0.04	-0.03
Content_Func_Ratio_Mean	1.12	1.13	1.18	1.14	1.15	1.20	1.12	1.15	1.21	1.11	1.13	1.16
SD	-0.14	-0.14	-0.15	-0.17	-0.16	-0.18	-0.15	-0.16	-0.19	-0.16	-0.18	-0.15
Content_Word_Count_Mean	69.76	69.56	71.98	67.73	69.89	73.51	69.96	71.15	74.20	67.87	68.82	70.16
SD	-11.29	-11.19	-11.71	-11.04	-11.17	-11.31	-11.81	-11.69	-11.98	-11.99	-12.13	-10.80
Function_Words_Mean	62.82	62.31	61.51	60.71	61.38	61.76	63.33	62.75	62.35	61.76	61.64	61.04
SD	-10.94	-10.34	-10.23	-11.85	-10.04	-9.79	-12.03	-11.05	-11.02	-11.09	-11.28	-10.00
Passive_Ratio_Mean	0.45	0.44	0.38	0.32	0.33	0.46	0.35	0.38	0.37	0.35	0.29	0.32
SD	-0.62	-0.61	-0.50	-0.51	-0.57	-0.55	-0.57	-0.57	-0.53	-0.56	-0.47	-0.51
Nominalization_Ratio_Mean	0.68	0.76	0.91	0.94	1.09	1.53	0.94	0.90	1.24	0.72	0.98	1.23
SD	-0.72	-0.83	-0.94	-0.92	-0.93	-1.04	-0.91	-0.82	-1.09	-0.81	-0.92	-1.00
Pronoun_12_Ratio_Mean	10.26	10.20	10.17	9.99	9.92	9.58	10.12	9.92	9.61	10.02	9.95	9.68
SD	-3.25	-3.19	-3.05	-3.54	-3.02	-2.67	-3.07	-3.39	-3.32	-3.30	-3.23	-3.21
BERTScore_Mean	1.00	0.97	0.95	0.95	0.94	0.92	0.96	0.95	0.93	0.96	0.95	0.93
SD	0.00	-0.01	-0.01	-0.01	-0.02	-0.02	-0.01	-0.01	-0.02	-0.01	-0.02	-0.02
PC1_Mean	1.29	0.53	0.05	0.10	-0.39	-1.45	0.28	0.08	-0.62	0.52	0.13	-0.52
SD	-1.23	-1.43	-1.57	-1.23	-1.34	-1.52	-1.36	-1.42	-1.32	-1.37	-1.29	-1.45
PC2_Mean	-1.32	-0.56	-0.15	0.10	0.33	0.94	-0.17	0.03	0.60	-0.31	-0.04	0.54
SD	-0.99	-0.93	-1.02	-1.00	-1.03	-1.05	-0.91	-0.93	-0.93	-0.96	-1.01	-0.98

Exploring creative variation in LLM literary translation: A multi-dimensional analysis of temperature and prompting effects

Jin Yim
Graduate School of Translation and Interpretation, Ewha Womans University

Abstract

This study explores the potential for inducing creative variation in Large Language Model (LLM) literary translation by examining the interaction between prompting strategies and the temperature parameter. Unlike traditional machine translation, LLMs offer a stochastic framework that allows users to modulate stylistic nuances. Focusing on Yi Sang's modernist masterpiece *The Wings*, this research investigates how combinations of creativity-oriented instructions (P1–P4) and temperature settings (T0.2–T1.2) affect linguistic outputs. To ensure methodological rigor, the study utilized an API-based environment to isolate session histories and fix sampling variances, thereby ensuring reproducibility. The generated translations were analyzed using Principal Component Analysis (PCA) and Linear Mixed-Effects Models (LMM). Results indicate that higher temperatures, when combined with explicit creativity prompts, significantly increase lexical diversity and structural complexity, successfully leading to outputs that move beyond conventional baselines. Notably, while direct instructions (P2) produced the highest degree of novelty, few-shot examples (P4) acted as an anchor that limited creative divergence. Qualitative analysis further confirms that these statistical shifts translate into sophisticated literary nuances without compromising semantic acceptability. These results provide a methodological foundation for maintaining academic validity in the emerging discourse on AI-generated creative translation.

Keywords: LLM translation; literary translation; creativity; temperature; prompt engineering; Principal Component Analysis (PCA); Linear Mixed-Effects Model (LMM)

키워드: LLM 번역, 문학번역, 창의성, 온도, 프롬프트 엔지니어링, 주성분분석, 선형혼합효과모형

임진(https://orcid.org/0009-0005-4335-3329)
이화여자대학교 통역번역대학원 강사
jin.yim@ewha.ac.kr

논문 투고일: 2026년 2월 15일
1차 심사 완료일: 2026년 3월 2일
2차 심사 완료일: 2026년 3월 8일
게재 확정일: 2026년 3월 16일