번역학연구, 27(1), 193-224 (2026년 3월 31일)
https://doi.org/10.15749/jts.2026.27.1.006

기계번역의 문학번역 적용에 관한 예비 연구: 2025년 챗GPT와 NMT 결과물을 중심으로*

이준호(중앙대학교)

1. 서론

기계번역은 눈부신 기술적 발전을 거듭해 왔다. 특히 2016년 신경망 기계번역(Neural Machine Translation, NMT)의 도입과 2022년 거대 언어모델(Large Language Model, LLM)의 등장은 번역 품질 인식에 많은 변화를 가져왔다. 이러한 변화 속에서 기계번역의 적용 범위는 정보 전달 중심 텍스트를 넘어 점차 문학번역 영역으로까지 확장되고 있다(Castaldo et al., 2025).

기계번역의 문학번역 적용을 논의하기 위한 첫 단계는 문학번역이란 행위를 정의하는 것이다. 하지만 저명한 번역학자 허먼스(Hermans, 2007)가 지적했듯 문학번역을 명확히 정의하는 일은 매우 어렵다. 그럼에도 불구하고 문학으로 읽히기 위해 번역된 글이라 정의한다면 이견이 많지 않을 것이다. 이러한 정의에 기반한다면 문학번역은 하나의 언어를 다른 언어로 전환하는 단순한 행위가 아니라, 문화적 요소를 옮기는 과정이자 언어를 넘어선 문화 간의 소통 활동이며 창의적인 활동으로 보아야 한다.

달리 말해 문학번역은 특정 문제를 인식하고 이해한 뒤 원문으로부터

* 본 논문은 유영번역상 심포지엄에서 필자가 강의한 내용을 발전시킨 원고임을 밝힌다.

다소 벗어나더라도 다양한 대안을 생성하고, 그중 목표 텍스트와 문화에 가장 적합한 해법을 선택하는 과정이다(Bayer-Hohenwarter & Kussmaul, 2020; Rojo, 2017). 여기에 더해, 르페브르(Lefevere, 2016)가 주장했듯 번역은 번역가의 주체적 개입이 이루어지는 일종의 다시쓰기(rewriting) 활동이라는 점까지 고려한다면, 문학번역은 그 속성이 훨씬 복잡해진다. 또한, 이러한 복잡한 인지적 과정을 수행하면서 번역가는 자신의 주체적 판단에 근거한 다양한 전략을 사용한다. 대표적인 예로 베누티(Venuti, 2017)가 제시한 독자가 자연스럽게 이해할 수 있도록 원문의 이질성을 줄이고 목표 문화에 맞추는 방식인 자국화(domestication)와 독자가 원문의 타자성과 낯섦을 느낄 수 있도록 원문 문화의 특수성을 드러내는 방식인 이국화(foreignization) 전략 등이 있을 것이다.

하지만 기계번역은 언어를 처리할 뿐 복잡한 번역 과정에서 주체성을 가진 전략을 구사하기는 어렵다. 따라서 기계번역을 문학번역에 활용하겠다는 시도는 신중한 검토를 요한다. 문학작품을 다른 언어권의 문학으로 새롭게 탄생시키는 것은 언어적 전환 행위가 아닌 전략적 창작 행위에 더욱 가깝기 때문이다.

그러나 기계번역을 비롯한 번역 기술은 빠르게 발전하고 있으며, 현대사회는 생산성 중심의 발전 일로에 있다(Wang, 2024). 이제는 누구나 자동화된 인공지능 기술을 활용하여 끊임없이 새로운 결과물을 생산해 내는 시대가 되었다. 이러한 흐름 속에서 문학번역의 근본적 속성을 충분히 고려하지 않은 채 기계번역의 문학번역 적용 논의는 이미 오래전에 시작되었다(김용출, 2023). 여기에 더해 문학을 단순히 출판 가능한 상품으로 보는 상업화의 시각과, 주어진 기술은 당연히 사용해야 한다고 여기는 기술 편의주의 시각도 존재한다. 이러한 기술 발전의 비이성적 파고 속에서 기계번역의 허와 실을 진단하는 일은 학술적·실무적 가치를 지닌다.

이에 본 연구는 단순하지만 명확한 연구 질문을 설정하였다. 문학번역 수행에 있어, 현재 기계번역은 과거 대비 얼마나 발전했는가? LLM은 NMT 대비 어느 정도 우수함을 보이는가? LLM 기반 기계번역으로 구현 가능한 품질 수준은 어느 정도인가? 상기 연구 질문에 대해 답을 구하고 이를 토대로 문학번역에서의 기계번역 활용 가능성과 한계를 논의하고자 한다.

이 논의를 객관적으로 전개하기 위해서는 기계번역의 발전 상황을 살펴볼 필요가 있다. 이에 2장에서는 국내 및 해외의 기존 연구를 검토한다. 이를 통해 기계번역에 대해 과거부터 제기된 주된 문제점이 무엇이었고, 또 어떻게 문제를 해결했는지 살펴본다. 분석 단계에서는 국내에서 많은 독자가 이미 번역을 통해 읽은『노인과 바다』및『오만과 편견』의 인간번역본, NMT, LLM 기반 번역본을 분석하여 기계번역의 문학번역 적용에 대한 가능성을 추가로 살펴본다. 이어서 NMT의 문학번역 적용의 한계를 비판적으로 고찰한 마승혜(2018)와 이준호(2019)의 논문에서 언급된 기계번역의 문제가 얼마나 해결되었는지 현시점의 데이터를 통해 살펴본다. 이를 통해 2025년 기준 기계번역의 문학번역 적용에 대한 현주소를 파악하고, 미래 연구 주제를 제안하고자 한다.

2. 이론적 배경 및 선행연구

2.1 번역 기술의 발전과 문학번역에의 함의

번역 기술의 초기 발전 단계는 컴퓨터보조번역(Computer-Assisted Translation, CAT)이 큰 역할을 차지했다. CAT는 번역 수행의 주체를 인간번역가로 전제한 상태에서 번역 과정의 효율성과 일관성을 지원하는 도구 및 기술을 의미한다. CAT 환경에서 번역가는 과거 번역 결과를 데이터베이스 형태로 축적하며, 이후 유사 문장이나 반복 어휘가 등장하면 자동으로 제시된 대안을 활용할 수 있다. 이러한 경험은 일종의 번역 과정 부분 자동화로서 번역가는 결과물을 체계적으로 재활용하고, 번역의 생산성 및 일관성 향상이라는 혜택을 볼 수 있다.

반면 CAT와 달리 기계번역은 번역 수행 자체를 인간이 아닌 기계가 담당한다는 점에서 근본적 차이를 가진다. 기계번역은 초기의 규칙 기반을 거쳐 통계 기반으로 발전했으며, 2016년 이후 NMT 도입을 계기로 품질 면에서 획기적 개선을 이루었다. 품질 측면에서 NMT는 문장 단위의 유창성과 자연스러움이 크게 향상된 것으로 평가받는다(Karabayeva & Kalizhanova,

2024; Toral & Way, 2018). 하지만 누락(Yang et al., 2019), 장거리 의존 관계 처리(Pouget-Abadie et al., 2014), 담화 수준의 맥락 파악(Läubli et al., 2018)은 난제로 남아있었고 NMT의 문학번역 적용에는 한계가 있었다.

일례로 문학번역에서 기계번역의 적용 관련 자주 인용되는 토랄과 웨이(Toral & Way, 2018)의 연구는 NMT가 과거 통계 기반 기계번역보다 문학번역에서 더 우수한 품질을 보인다고 주장한다. 다만 인간번역과의 직접 비교에서는 일부 데이터에서만 인간과 유사한 수준을 달성했을 뿐, 전반적으로는 인간번역의 수준에 도달하지 못했다.

코파스 페스터와 노리에가 산티아네스(Corpas Pastor & Noriega-Santiáñez, 2024)는 영어-스페인어에서 문학 텍스트에서 나타나는 창의적 어휘 결합, 특히 변형된 관용구 처리에 있어 기계번역이 여전히 어려움을 겪고 있음을 제시하였다. 이러한 맥락에서 이들은 창의성은 본질적으로 인간 고유의 특성이며 새로운 맥락 속에서 이를 재현하는 일은 NMT에서 매우 어렵다는 점을 강조했다. 즉 NMT의 성능 향상은 확인되었지만, 문학번역의 전면적 자동화를 정당화하지는 못한다는 결론이 또다시 관찰되었다.

하지만 2022년 말 등장한 거대언어모델 기반 시스템인 챗GPT는 기존 NMT와는 다른 방식으로 기계번역의 가능성을 확장한 것으로 평가된다(이창수, 2024; 황지연 등, 2024; Jiang & Zhang, 2024; Peng et al., 2023). 물론, LLM은 번역 전용 모델은 아니며 높은 연산 비용과 응답 속도 문제로 인해 대량 번역 작업에 최적화된 상용 번역 도구라고 보기는 어렵다(Wang et al., 2024). 그럼에도 불구하고 LLM 기반 번역이 주목받는 이유는 번역 결과가 사용자의 명시적 지시, 즉 프롬프트에 따라 달라질 수 있기 때문이다(박수정과 최은실, 2023; Yu & Yao, 2026).

사용자가 특정 작가의 문체적 특성을 설명하거나 번역의 목적을 사전에 제시하면 LLM은 이를 반영한 번역 결과 생성이 가능하다. 이는 번역 결과가 사용자가 통제할 수 없는 모델 내부의 연산에만 의존하던 NMT에서는 불가능한 기능이기에 번역 결과물 생성에 근본적 변화를 가져왔다.

물론 이러한 변화가 항상 문학적으로 완성도 높은 번역을 보장하는 것은 아니지만 번역 과정에서 사용자가 번역의 방향성과 조건을 설정할 수 있게 되었다는 점은 주목할 필요가 있다. 특히 번역이 목적 지향적 행위라는

관점에서 볼 때 LLM 기반 번역은 기존 기계번역이 넘지 못했던 경계에 접근할 가능성이 있다. 바로 이 지점에서 과거에는 본질적으로 인간번역가의 영역으로 간주되었던 문학번역에 대해서도 잘 설계된 프롬프트만 있다면 기계번역이 적용 가능할 수 있다는 논의가 이뤄지고 있다.

2.2 번역 기술 발전에 대한 해외 학계의 논의

최근 문학번역 연구는 LLM의 등장이 문학번역의 난제들을 실질적으로 완화했는가를 중심으로 재편되고 있다. 그러나 다수의 연구는 '개선은 있으나, 격차는 여전히 크다'라는 결론을 공유한다.

은유는 기계번역 연구에서 반복적으로 지목되는 난제다. 카라칸타 등(Karakanta et al., 2025)은 LLM이 영어-네덜란드어의 문학번역, 특히 은유의 번역에서 기존 NMT 시스템보다 우수하다는 명확한 근거를 찾지 못했다고 보고한다. 장 등(Zhang et al., 2025) 또한 영어-중국어-독일어에서 LLM 기반 문학번역이 진전을 보였음에도 불구하고 과도한 직역의 문제를 충분히 해결하지 못했으며 인간번역과의 품질 격차가 여전히 상당하다는 점을 지적한다.

더 나아가 두 등(Du et al., 2025)은 영어-네덜란드어-중국어-스페인어에서 챗GPT의 문학번역 성능을 검토하면서 창의성을 최대한 끌어내는 최적 설정을 모색했으나 결과는 만족스럽지 못했음을 지적하였다. 모든 설정에서 챗GPT의 번역은 인간번역보다 창의성 점수가 현저히 낮고 오류는 더 많았으며, 가장 창의적 설정조차 네 개 언어 중 세 개에서 인간번역 수준에 미치지 못했다.

형식의 재구성이라는 측면에서 보면, 가오 등(Gao et al., 2024)은 챗GPT, 구글번역, 딥엘을 대상으로 중국 고전시의 영어 번역성능을 비교하기 위해 충실성, 유창성, 언어 스타일을 평가하였다. 그 결과 챗GPT는 다른 두 시스템보다 전반적으로 우수한 성능을 보였고, 구글번역과 딥엘 간에는 큰 차이가 없었다. 하지만 정형적 운율을 완벽하게 구현하는 것은 여전히 어려운 과제로 드러났다.

이처럼 기계번역 품질이 꾸준히 향상되고 있음은 분명하지만 문학번역이 제시하는 다양한 난제를 해결하기에는 여전히 충분하지 않다는 지적이 다수이다. 특히 문학번역이 원문이 지닌 고유한 형태와 구조를 보존해야 하

며 동시에 끊임없는 고민과 창의적 선택이 필수적이라는 점을 고려하면 기계번역이 일부 한계를 맞이하는 것은 당연한 결과일 수 있다.

하지만 부정적 평가만 있는 것은 아니다. 최근 연구 중 일부는 공학적 설계 및 워크플로우 통합을 통해 긍정적 결과를 보고했다. 카스탈도 등(Casataldo et al., 2025)은 영어-이탈리아어 LLM 기반 후편집이 문학번역에서 생산성과 창의성의 균형을 달성할 수 있음을 제시한다. 해당 연구는 지원이 잘 이루어지는 언어 쌍에서는 LLM이 생성한 번역을 후편집할 경우 인간번역보다 작업 시간이 크게 단축되면서도 창의성 수준은 유사함을 보고했다.

하지만 기술의 발전 여부와 상관없이 주목해야 할 점은 구베로프 아레나스와 토랄(Guerberof-Arenas & Toral, 2022)은 기계번역을 문학번역의 제약조건으로 규정했다는 것이다. 이 연구는 기계 산출물 후편집하면 번역가의 자유도는 제한되고 창의적 전환을 감소시키는 효과를 낳는다고 보고했다. 기계번역은 문장 작성의 '틀'을 제시하고 번역가는 그 안에서 움직일 수밖에 없으므로 창의성의 폭이 제한될 수밖에 없다는 것이다. 결국, 이러한 문제는 기계번역의 품질이 아무리 향상되더라도 문학번역에서는 걸림돌로 작용할 가능성이 있다.

상기 살펴본 해외 번역학 연구는 기계번역 발전을 긍정적으로 평가하고 이를 문학번역 적용의 근거로 제시하는 연구가 존재하는 반면, LLM 역시 문학번역에 적용하기에 부족함이 많다고 주장하는 연구가 더 많아 보인다. 그렇다면 국내의 연구는 어떠한 결과를 제시하고 있을까? 한국어라는 언어의 특성이 해외 연구와 완전히 다른 결과를 보여주었을지 다음 절에서 논의하고자 한다.

2.3 번역 기술 발전에 대한 국내 학계의 논의

국내 연구 역시 전반적으로 기계번역이 문학번역에서 드러내는 언어적 기술적 한계를 다각도로 지적한다. 마승혜(2018)는『채식주의자』번역 사례를 바탕으로 문학작품 기계번역의 한계를 상세히 고찰하였다. 특히 인간번역과 기계번역의 복합적 인식 및 해석 능력에서의 차이, 해석에 기반을 둔 선택 능력에서의 차이, 재현 및 창조 능력에서의 차이, 독자와 소통 능력에서의 차이에 기반하여 인간번역가가 어떻게 언어적 소통을 넘어서 문학번

역을 수행하는지 설명하며 기계번역이 어떠한 한계를 보이는지 분석하였다.

이준호(2019)는 더욱 기술적인 측면에서 기계번역의 적용 한계를 고찰하였다. 해당 연구에서는 NMT의 대표적 문제로 누락과 문맥 파악의 한계를 지적하였다. 또한 학습되지 않은 미등록어 처리 문제를 중요 제약으로 제시하였다. 더불어 장문 처리의 한계도 언급하며, NMT가 짧은 문장에 비해 긴 문장을 번역할 때 품질이 저하되는 경향이 있고, 이를 보완하기 위해 원문 단순화 또는 임의 분할전략이 시도되어 문학작품의 원문 형태를 해한다고 주장하였다.

여기에 더해 이창수(2019)는 문학번역에서 NMT와 인간번역사의 문체에 차이가 있음을 지적하였으며, 이창수(2021)는 어휘 사용 패턴 측면에서 기계번역과 인간번역 사이에 현저한 차이가 존재한다고 주장하였다. 이러한 연구는 문학번역에 기계번역 적용이 제한적일 수 있는 원인을 재조명한다고 하겠다.

한국어-영어 이외의 언어쌍에서 곽순례(2022)는 아-한 기계번역의 가용성과 한계 연구에서 NMT 결과물을 분석하며, 정성적 분석에서 기계번역이 내용은 이해할 수 있더라도 부자연스럽고, 인간번역과 비교해 가독성이 떨어지며, 문맥에 적합하지 않은 의미 선택이 빈번하게 나타났음을 확인하였다. 이현주(2022)는 문학작품의 중-한 기계번역 결속구조를 인간번역과 비교하여 기계번역이 담화 및 결속 장치를 충분히 구현하지 못하는 문제를 지적하였다. 노금송과 왕원(2024) 역시 한국어 문화소에 관한 기계번역과 인간번역의 비교 분석을 통해 문화적 함의와 문학적 표현에서 기계번역의 한계를 구체화하였다. 즉 국내 연구는 영어-한국어뿐 아니라 다양한 언어 조합에서도 문학번역에서 기계번역의 한계를 반복적으로 보여주었다.

한편 LLM의 확산 이후의 국내 연구는 활용 가능성과 함께 새로운 쟁점도 제기한다. 남철진(2025)은 챗GPT 중국소설 번역의 문제를 오역, 누락, 표현의 세 영역으로 구분하고 예상외로 많은 문제가 발견되었다고 보고하였다. 또한, 임진(2025)은 성중립 표현의 AI 기반 기계번역 양상을 분석하여 모델별 성별 중립성 유지에서 유의미한 차이를 발견했다고 보고하였다. 특히 GPT와 DeepL이 원문에 없는 성별 표지를 추가하여 성중립성을 훼손하는 경향을 보였으며, 이는 기존 연구(Ghosh & Caliskan, 2023)와도 일치한다

는 주장을 제시하였다.

하지만 가능성에 더욱 중점이 있는 연구도 존재한다. 마승혜(2024)는 사례 분석을 통해 챗GPT가 제시하는 번역의 한계를 지적하며, 인간번역 제목과 비교할 때 지시적 기능 또는 호소적 기능 등 어느 한 기능은 수행하지만 여러 기능을 동시에 아우르는 복합적 기능 수행은 충분히 구현하지 못했다고 보고하였다. 다만 인간번역가 및 출판 관계자가 제목을 결정하는 과정에서 챗GPT의 제안을 다양한 가능성 중 하나로 참고할 수 있다는 점에서 제한적 활용 가능성을 제기하였다.

여기에 더해 마승혜(2025)는 정교하게 설계된 프롬프트를 통해 AI에 번역가로서의 역할을 부여하고, 구체적인 번역 원칙과 전략을 입력하면 AI는 창의적 번역을 수행하며 인간번역가의 보조 수단으로 활용될 수 있음을 주장하였다.

이상의 논의를 요약하자면 국내 논의는 기계번역이 문학번역에서 일정한 보조적 역할은 수행할 수 있으나, 인간번역가의 해석적 창의적 개입 없이는 본질적 제약을 극복하기 어렵다는 논의가 아직까지는 크다고 하겠다. 또한, 최근에는 LLM을 포함한 기술의 활용 가능성을 탐색하면서도, 창의적 기능 수행의 한계와 함께 편향 문제 등 문학번역 관련 연구 범위 확장의 움직임이 관찰된다.

이상의 번역 기술 발전 및 해외와 국내의 선행연구를 종합하면, 문학번역에서 기계번역은 NMT 도입 이후 품질이 개선되었고, LLM의 등장으로 프롬프트 기반 제어 가능성이 확대되었으나, 창의성, 은유, 정형성, 담화 결속, 문화소 처리 등 여전히 한계가 반복적으로 확인된다. 물론 후편집을 위한 충분한 기계번역 품질이 확보된다면 생산성 향상이 가능할 수 있지만, 기계번역의 활용으로 인해 고정된 결과물이 생성된다면 인간 창의성의 공간이 줄어들 가능성을 배제하기도 어렵다.

이러한 긍정 및 비판적 시각이 공존하는 상황에서 문학번역에서의 기계번역 활용은 더 체계적이고 심층적인 논의를 요구한다. 이에 본고는 영어-한국어 언어쌍을 대상으로, 기존 연구에서 충분히 다루어지지 않았던 문학번역에서의 NMT와 챗GPT의 적용 가능성과 한계를 검토함으로써 관련 논의의 확장을 시도하고자 한다.

3. 연구 방법

본고는 문학번역에서 LLM 기반 번역이 NMT 대비 우수한 성능을 보이며 문학번역 과업을 수행할 수 있는지를 검토하기 위한 초기 연구이다. NMT와 LLM의 성능을 직관적으로 비교하기 위해, 동일한 원문 텍스트를 대상으로 동일 시점에서 생성된 NMT 및 LLM 기반 번역 결과물을 비교 및 분석하도록 연구를 설계하였다.

이를 위해 국내에서 널리 알려진 영어 문학작품인 『오만과 편견』의 초반부 약 5,300단어와 『노인과 바다』에서 4,949단어를 각각 추출하였다. 이후 번역 과정에서 충분한 맥락이 반영되도록 보증하기 위해, 파파고와 챗GPT-4를 활용하여 2025년 10월 동일 시점에 원문을 200단어 단위로 나누어 입력하고 번역 결과물을 생성하였다. 또한, 학습으로 인한 편향성을 방지하기 위하여 챗GPT가 학습을 하지 않는 설정으로 실험이 시행되었다.

다만 챗GPT-4에서 사용한 프롬프트는 복잡한 설계 과정을 거치지 않았다. 첫째, 과도하게 정교한 프롬프팅이 연구자의 인위적 개입을 증가시켜 모델의 기본적인 번역 성능 평가를 저해할 수 있기 때문이다. 둘째, 실제 번역 환경에서 일반 사용자들이 매우 복잡한 프롬프트를 사용하는 경우는 드물다. 셋째, 프롬프트가 지나치게 복잡할 경우 연구의 재현 가능성 역시 낮아질 수 있다. 이에 본고에서는 '한국 독자가 이해할 수 있도록 영어 텍스트를 한국어로 번역해 달라'는 단순한 프롬프트를 사용하였다.

연구의 타당성을 확보하기 위해 본고에서는 자동 평가와 수동 평가를 병행하였다. 수동 평가에서는 제삼자의 정성적 평가와 연구자의 텍스트 분석을 통한 오류 정량화를 텍스트 전반부와 후반부에 각각 실시하였다.

먼저 자동 평가에는 BLEU 점수를 활용하였다. BLEU 점수가 번역 품질을 절대적으로 판단하는 기준이라고 할 수는 없으나, 인간번역과의 유사성을 정량적으로 측정할 수 있다는 점에서 기계번역 품질 예측을 위해 널리 사용되는 지표이다. 본 연구에서는 인간번역본과 NMT 및 LLM 기반 번역 결과물을 정량적으로 비교하기 위해 BLEU 점수를 적용하였다. 평가의 기준이 되는 인간번역본으로는 『오만과 편견』의 경우 민음사에서 출간한 윤지관 전승희 공역본을, 『노인과 바다』의 경우 교육문화연구회에서 출간한 이종한

의 번역본을 사용하였다.

다음으로 자동 평가의 한계를 보완하기 위해 간략한 수동 평가를 실시하였다. 우선 두 번역 결과물 간의 품질 차이가 직관적으로도 분명하게 관찰되는지 확인하고자, 두 번역을 비교하여 어느 쪽이 상대적으로 우수한지만을 판단하는 평가 과업을 설정하였다.

텍스트 전반부에 대한 평가는 기계번역 및 문학번역 관련 연구 경험을 보유한 번역학 박사학위 소지자 4인에게 의뢰하였다. 평가자는 『오만과 편견』과 『노인과 바다』에 각각 2인씩 배정되었으며, 각 평가자는 하나의 번역이 파파고의 결과물이고 다른 하나가 챗GPT에 의해 생성되었다는 정보만을 제공받은 상태에서 30분간 두 번역문을 맹검 방식으로 비교하고 평가하였다. 평가 기준으로는 내용 전달의 정확성과 한국어 사용의 자연스러움 및 전반적 맥락의 흐름을 반영하는 유창성을 제시하였으며, 세부 점수화 없이 두 번역 중 상대적으로 우수하다고 판단되는 결과물만을 선택하도록 요청하였다. 마지막으로 평가자는 특정 번역을 더 우수하다고 판단하게 된 결정적인 문장을 선정하고, 그 판단의 근거를 간략히 기술하도록 요청받았다.

상기 기술된 직관적 수동 평가를 보완하기 위해, 텍스트 후반부를 대상으로 연구자가 정량적 분석을 추가로 수행하였다. 이를 위해 각 텍스트에서 후반부 50문장을 선정하고, 파파고와 챗GPT의 번역 오류를 정량적으로 비교한 이창수(2024)의 연구에서 제시한 '내용을 잘못 번역한(content)' 오류와 '황당한 수준의 직역식(odd_literal) 오류에 따라 발생 빈도를 분석하였다.

본고의 연구 질문 중 하나는 기계번역 기술이 과거에 비해 어느 정도 발전하였는지 살펴보고, 이를 통해 문학번역에의 적용 가능성을 검토하는 데 있다. 이를 위해서는 과거 기계번역의 문학번역 적용을 분석한 선행연구를 검토하고, 해당 연구에서 지적된 번역상의 문제점들이 현시점에서 어느 정도 개선되었는지를 평가하는 과정이 필요하다. 이에 본고에서는 마승혜(2018)와 이준호(2019)가 분석한 『채식주의자』, 『빈처』, 『운수 좋은 날』의 기계번역 결과물에서 제기된 문제점들이 2025년 기준의 NMT 및 챗GPT 기반 번역에서 개선되거나 해소되었는지를 분석하였다. 특히 어휘 다양성과 관련해서는 코퍼스 기반의 정량적 분석을 시행하였고, 일부 필요에 의해 신조어 데이터 세트를 만들어 추가 분석을 진행하였다.

4. 분석 결과

본 장에서는 NMT와 챗GPT 기반 번역을 직접 비교 및 평가하고, 과거 대비 기계번역 기술이 어느 정도의 발전을 보였는지를 구체적인 사례를 통해 논의한다. 먼저 NMT와 챗GPT 기반 번역의 품질 비교에 대한 종합적 평가를 제시한 뒤, 여전히 남아있는 기계번역의 주요 단점을 논의한다.

4.1 총평

다소 자명하게 보일 수 있으나 2018년과 2019년의 NMT에서 관찰되던 오류가 2025년의 NMT와 챗GPT 번역에서는 상당 부분 해소되었음을 관찰하였다. 이는 번역 기술 전반의 발전을 반영하는 결과로 이해할 수 있다. 여러 가지 개선의 사례를 4.2 이후의 논의에서 단점과 함께 논의할 것이지만, 한 가지 예시만 들자면 다음과 같다.

<예시 1> 운수 좋은 날
원문: **개똥이**가 물었던 젖을 빼어놓고 운다.
NMT1 2019: The **dog** cries and cries while it sucks.
NMT2 2019: **He** cries after taking out the breast that dog bit.
구글번역 2025: **Gaedongi** cries after taking out the milk he had been drinking.
챗GPT: **Kaettong**, having let go of the breast he was sucking, begins to cry.
<이하 모든 예시의 강조는 연구자가 입력>

위의 예시는 일반명사이면서 동시에 고유명사로도 사용될 수 있는 어휘의 처리 그리고 문맥 파악까지도 개선되었음을 보여준다. 같은 단어가 다수의 의미를 지니는 어휘의 처리와 문맥의 파악은 기계번역의 고질적 문제로 지적됐는데, 이를 NMT와 챗GPT의 번역 결과물 모두에서 해결된 사례를 보여주었음은 의미가 있다. 그 외에도 챗GPT가 NMT인 파파고의 품질을 압도하는 사례는 다수 관찰되었고, 이는 이미 발표된 여러 논문의 결과와 그 맥을 같이 한다(Jiang & Zhang, 2024; OpenAI, 2023; Manakhimova et al., 2024).

본 연구에서 사용한 데이터를 활용하여, 자동평가와 수동평가를 실행한 결과 챗GPT의 번역 결과물이 파파고보다 우수함을 지지하는 결과가 나왔다. BLEU 점수 산출 결과, 『오만과 편견』에서는 챗GPT는 0.0651을 기록했지만, 파파고는 0.0408로 상대적으로 낮은 점수를 보였다. 또한, 『노인과 바다』에서도 챗GPT는 0.0644, 파파고는 0.0424로 동일한 경향이 확인되었다. 따라서 챗GPT의 번역이 인간 번역가의 해석과 선택이 반영된 결과물과 더 높은 유사성을 보였다.

다음으로 실시한 전문가 평가에서도 챗GPT 번역의 상대적 우위를 확인할 수 있었다. 『오만과 편견』의 번역본을 검토한 평가자 1과 2는 모두 챗GPT가 수행한 번역을 정확히 식별하였으며, 해당 번역이 더 우수하다고 평가하였다. 『노인과 바다』의 번역본을 검토한 평가자 3과 4 역시 같은 판단을 내렸다.

> 평가자 1: B는 오역, C는 원문의 구조와 다르게 한 문장을 두 문장으로 나누고 자연스럽게 번역 (B는 파파고, C가 gpt 결과물로 생각됩니다. B는 원문구조를 그대로 따라 직역한 번역투가 많고, C는 맥락에 따라 원문 표현을 유지하기 보다는 자연스러운 번역 생성. 또한 대화 상황에서 인물간 관계 설정(부부, 부모-자녀)에 따른 존댓말 사용, 말투 등에 차이를 두고 보다 자연스러운 구어체로 번역한 점이 B보다 C가 자연스러움) [중략]

> 평가자 2: B열에서 나타나는 번역 오류의 양상이 전형적인 신경망 기계번역(NMT)의 한계를 극명하게 보여주기 때문입니다.
> 1. 다의어 처리 및 맥락 파악 능력의 차이: 첫 행부터 'fortune'의 중의적 의미(재산/행운)를 정확히 포착하지 못한 점으로 보아, B열은 학습 데이터의 한계나 넓은 맥락 파악 능력이 부족한 파파고일 가능성이 높습니다. [중략]

물론 정확성과 유창성이라는 다소 단순한 기준으로 평가가 이뤄졌으며, 4인의 제한적 평가이기에 신뢰성이 매우 높지는 않을 수 있다. 하지만 무작위 배정된 번역에서 4인의 평가자가 모두 챗GPT의 번역을 식별하고 상대적으로 우수하다고 일관되게 평가했기에, 전문가 판단에 기반한 일관된 차이

가 존재할 가능성이 있다 하겠다.

여기에 더해 후반부의 오류 정량화 분석에서도『오만과 편견』은 파파고에서 심각할 수준의 직역과 오역은 20개, 챗GPT에서는 2개의 오류만 관찰되었다. 상대적으로 상세 묘사와 은유가 많았던『노인과 바다』는 파파고에서 심각할 수준의 직역과 오역은 34개, GPT에서는 18개의 오류가 관찰되었다.

요약하자면 기술의 발전에 따라 최근에 개발 및 출시된 챗GPT가 NMT보다 더 나은 번역 결과물을 산출하는 것은 자연스러운 결과로 볼 수 있다. 그럼에도 불구하고 영어-한국어 문학번역의 언어쌍에서도 이러한 격차가 정량적 및 정성적으로 확인되었다는 점은 주목할 필요성이 있다.

다만 챗GPT가 상대적으로 우수한 번역을 수행한 것으로 관찰되었으나, 그 품질이 매우 우수한 문학번역으로 평가하기에는 한계도 분명했다. 또한, 문학번역 수행에서 인간 번역가를 대체할 수준의 성과를 보였다고 보기는 더더욱 어려웠다. 이러한 맥락에서 다음 절에서는 챗GPT 기반 문학번역에 여전히 남아있는 문제점을 중심으로 논의를 이어가고자 한다.

4.2 누락 및 추가

단어, 문장 단위의 다양한 누락은 NMT의 문제점으로 지적되어 왔다. 그렇다면 챗GPT에서 누락의 문제는 어느 정도 해결되었을까? 최근 국내 연구를 보더라고 챗GPT가 NMT 대비 누락이 줄어들었다고 주장하기는 어렵다(홍서연, 2025). 또한, 해외 연구에서도 LLM이 여전히 누락의 문제를 완전히 극복하지 못했음을 보고하고 있다(Zhang et al., 2024). 하지만 본 연구의 분석 데이터에서는 데이터의 양이 부족한 탓인지, 챗GPT가 파파고 대비 누락이 많은지 적은지를 명확하게 시사하는 지점은 관찰되지는 않았다. 물론 누락의 예시를 찾는 것은 어렵지 않았다.

<예시 2> 오만과 편견
원문: And his answering immediately to the last question, ‘Oh, the eldest Miss Bennet, beyond a doubt: there cannot be two opinions on that point.’
챗GPT: 아, 장녀 베넷 양이죠. 의심할 여지도 없습니다. 그 점에서는

이견이 있을 수 없죠.’라고요.

이처럼 LLM을 번역에서 사용한다고 하더라도 누락의 가능성은 여전히 존재한다. 더욱 흥미로운 것은 번역 과업에서 지시하지 않은 내용을 추가하는 경향성이 관찰되었다는 것이다.

<예시 3> 운수 좋은 날
원문: 정거장까지 끌어다주고 그 깜짝 놀란 일 원 오십 전을 정말 제 손에쥠에, 제 말마따나 십 리나 되는 길을 비를 맞아 가며 질퍽거리고 온 생각은 아니하고, 거저나 얻은 듯이 고마웠다.
챗GPT: Having pulled me all the way to the station and, to my great surprise, actually placing the one yen and fifty sen into my hand—without considering at all that, as I said, it was a ten-li journey trudged through the rain and mud—I simply felt thankful, as though I had received it for nothing.

챗GPT가 추가한 설명:

“원 오십 전”은 옛 화폐 단위라 그대로 one yen and fifty sen으로 표기했습니다. (sen은 일본/조선 시기 1/100 엔 단위)
“십 리”는 한국의 옛 거리 단위로, 약 4km 정도에 해당하지만 여기서는 뉘앙스를 살려 ten-li라고 그대로 두었습니다.
문체는 원문의 약간의 아이러니와 담담한 회고적 톤을 살리려 했습니다.

연구자의 개입을 방지하기 위해 번역 수행을 위한 매우 기본적인 프롬프트를 제시했음에도 불구하고, 챗GPT가 번역 결과와 함께 자신의 번역 의도와 화폐 단위에 대한 설명을 추가적으로 제시했다는 점에 주목할 필요가 있다. 이러한 추가 정보를 제공하는 것이 번역 결과물 생성에 도움이 되는지는 논쟁의 소지가 있기 때문이다.

여기에 더해 LLM이 사실과 다르거나 원문과 직접적인 관련이 없는 결과물을 생성하는 이른바 환각(hallucination) 현상은 이미 잘 알려져 있으며, 다중언어 번역모델에서도 이러한 문제가 지속적으로 보고됐다. 특히 목표

텍스트에서 벗어난 번역(off-target translations)이나 불필요한 정보의 과도한 생성(over-generation)과 같은 현상은 선행연구에서도 지적된 바 있다 (Guerreiro, 2023). 이러한 특성을 고려할 때, LLM 기반 번역에서는 의도하지 않은 결과물이 생성될 가능성을 완전히 배제하기 어렵다.

물론 이러한 문제는 '원문 그대로 누락이나 추가 없이 충실하게 직역해라' 등의 프롬프트를 통해 일정 부분 통제될 가능성이 있으며, 위의 예시에서도 일부 통제가 가능했다. 그러나 2025년 기준 챗GPT를 활용한 단순 번역 수행에서는 누락의 가능성은 여전히 존재하며, 과생성과 같은 부작용 또한 배제할 수 없다는 점에 유의가 필요하다.

4.3 다의어 처리 및 문맥파악

다음으로 NMT의 한계로 자주 지적되어 온 문제는 문장 내부는 물론 문장 단위를 넘어서는 문맥을 충분히 파악하지 못한다는 점이다. 반면 LLM은 구조적 특성상 NMT보다 훨씬 더 많은 토큰을 입력으로 처리하면서 번역 수행이 가능하며, 더욱 넓은 범위의 문맥 정보를 고려할 수 있는 잠재력을 지닌다(Appicharla et al., 2025). 이를 쉽게 설명하면 LLM은 앞뒤 문장 혹은 그 이상을 포함한 보다 넓은 텍스트를 입력값으로 처리할 수 있어서 단일 문장을 기준으로 번역할 때 발생하는 문맥파악 부족의 문제가 상대적으로 완화될 가능성이 높다. 또한, LLM은 대규모 데이터로 학습되었기 때문에 일반적으로 더 풍부한 배경지식을 활용할 수 있다. 본 연구의 분석은 이러한 가정이 일부 사례에서는 적용 가능했지만 그렇지 않은 사례 역시 확인하였다.

<예시 4> 운수 좋은 날
원문: 그 학생을 태우고 나선 김 첨지의 **다리**는 이상하게 거뿐하였다
NMT 2019: The **bridge** of Kim Chu-ji who picked up the student was strange.
구글번역 2025: Kim Cheomji's **legs**, carrying the student, were strangely light.
챗GPT: After taking the student aboard, Kim Chŏm-ji's **legs** felt strangely light.

위의 예시는 2025년 현재의 구글번역과 챗GPT 모두 2019년의 NMT에서 발생한 문장 내의 맥락 파악과 관련된 다의어(polysemy)의 문제를 잘 극복하고 있음을 보여준다. 하지만 다음 예시는 그렇지 않은 사례를 보여준다.

<예시 5> 채식주의자
원문: “아버지!” **처형**이 외치며 **장인**의 팔을 잡았다. **장인**은 아직 흥분이 가시지 않은 듯 입술을 실룩거리고 있었다
구글번역 2018: "Father!" The **executioner** shouted and caught the **craftsman**'s arm. The **craftsman** was luring his lips as if he had not been excited yet.
구글번역 2025: "Father!" the **executioner** shouted, grabbing **his father-in-law**'s arm. His lips twitched, his excitement still lingering.
챗GPT: “Father!” the **executioner** shouted, grabbing the **craftsman**’s arm. The **craftsman** still seemed not to have calmed down, his lips twitching.
좀 더 문학적으로 다듬은 번역도 가능합니다:
“Father!” the **executioner** cried as she seized the **craftsman**’s arm. He still looked keyed up, his lips quivering.

상기 문제를 관찰한 이후 챗GPT에게 “처형”과 “장인”의 의미에 대해서 필자가 정보를 주어 학습을 시키고, 동일한 번역을 진행했지만 모든 문제가 완전히 사라지지는 않았다.

<예시 6> 챗GPT 수정본
“Father!” his sister-in-law cried out, grabbing her father’s arm. He still seemed unable to contain his agitation, his lips twitching slightly.

문맥 파악에 있어 다음 문제는 생략된 요소의 파악이다. 대표적인 예시가 한국어의 특성상 주어가 없이도 문장의 구성이 가능하고, 문학작품에서 이러한 주어 생략은 빈번하게 발생한다는 점이다. 실제로 기계번역에서 맥락을 완전히 파악하여 올바르게 주어를 복원하는 것은 여전히 어려운 일로 남아있다(Koh, 2023). 이러한 문제는 2025년 현재 NMT와 챗GPT에서도 빈번하게 관찰되었다.

<예시 7> 빈처
원문: 아내가 그것을 찾아 무엇을 하려는 것을 앎이라
NMT 2019: <주어 없이 시작> Know what your wife is looking for and doing.
구글번역 2025: **I** know what my wife is trying to do when she finds it.
챗GPT: **It** was the awareness that **his wife** was seeking it in order to do something with it.

위의 예시는 주어진 단어에 대한 정보를 처리하는 것은 가능했지만, 인물 간의 관계 맥락을 파악하여 의미를 복원하는 작업에는 어려움이 있음을 보여준다. 구글번역은 2025년 현재 "I"라는 명확한 주어 복원에 성공하였다. 하지만 챗GPT는 잘못된 주어를 설정하였고, 아내 역시 "his wife"라고 표기하여 등장인물의 관계를 오역하였다.

<예시 8> 채식주의자
원문: 장인은 탕수육을 **아내의** 얼굴에 들이밀었다.
구글번역 2018: The craftsman poured sweet and sour pork into **his wife** 's face.
구글번역 2025: The father-in-law shoved the sweet and sour pork into **his wife**'s face.
챗GPT: The craftsman thrust the sweet-and-sour pork up to **his wife**'s face.
톤을 조금 바꾼 선택지도 있습니다:
"The craftsman shoved the sweet-and-sour pork toward **his wife**'s face." (거칠게)
"The craftsman held the sweet-and-sour pork right up to **his wife**'s face." (중립)

유사한 사례로 생략된 소유격에 대해서도 문제가 관찰되었다. 원문에서 "아내"는 주인공인 '나'의 아내를 의미한다. 하지만 이를 파악하여 'my wife'라고 복원에 성공한 경우는 없었다. 이처럼 문장에 직접적으로 드러나지 않고 숨겨진 요소의 맥락을 파악하여 처리하는 것은 여전히 난제라고 할 수 있다.

맥락 파악의 범위를 조금 더 확장하자면, 문학작품에서 중요한 등장인물

의 관계를 기계번역이 파악하고 문화적 배경을 번역의 결과물에 반영하는 것은 더욱 어려워 보인다. 다음의 예시는 『오만과 편견』에 등장하는 부부간의 대화이다. 인간번역은 두 주인공 간의 관계를 살려서 번역했지만, 기계번역은 "베넷 씨"라고 번역하여 부부관계의 파악에 실패하였다. 추가적 확인을 위해 텍스트를 500단어 단위로 두 번을 넣어, 챗GPT가 더 큰 맥락을 파악하도록 시도해 보았으나 결과는 달라지지 않았다.

<예시 9> 오만과 편견
원문: Impossible, **Mr. Bennet**, impossible, when I am not acquainted with him myself; how can you be so teasing?" "
인간번역: 안될 말씀이에요, **여보**. 저 자신도 모르는 사이인데. 왜 사람을 자꾸 놀리시기만 하세요?"
파파고 2025: **베넷 씨**, 제가 그와 직접 알지 못할 때는 불가능합니다. 어떻게 그렇게 놀릴 수 있나요?
챗GPT: 그럴 수 없어요, **베넷 씨**. 제가 그 사람을 전혀 모르는데, 어떻게 그런 괴롭히는 말을 하실 수 있죠?

다음의 예시에서도 맥락 파악을 통한 주인공 간의 관계 설정과 문화적 배경의 반영은 여전히 기계번역에게 어려워 보인다. 아래 예시는 『노인과 바다』의 주인공인 노인 산티아고와 어린 소년과의 대화이다.

<예시 10> 노인과 바다
원문:"**Santiago**," the boy said to him as they climbed the bank from where the skiff was hauled up.
인간번역: "**산티아고**, **할아버지**." 소년은 조각배를 끌어올려 놓은 둑으로 올라가면서 노인에게 말했다.
구글번역 2025: "**산티아고**." 소년이 그에게 말했다. 그들은 작은 배를 끌어올린 곳에서 가고 있었습니다.
챗GPT: "**산티아고**," 소년이 말했다. 그들은 조각배가 끌어올려져 있던 곳에서 둑을 올라가고 있었다.

부자연스러운 번역 결과물은 차치하더라도, 기계번역은 노년의 주인공과 소년 사이의 관계 설정에 실패했음을 확인할 수 있다. 이 경우도 500단

어 입력을 통해 추가 검증을 시도해 보았으나 결과는 달라지지 않았다. 이처럼 영한 번역에서 등장인물 간의 관계에 따라 어투와 어역을 적절히 조정하는 일은 고도의 맥락 파악 능력을 요구하는 작업이다. 더 나아가, 이렇게 전략적으로 조정된 어투와 어역을 작품 전체의 번역에서 일관되게 유지하는 것은 더욱 어려운 과제라 할 수 있다. 실제로 본 연구의 데이터에서도 아버지가 딸에게 반말을 사용하다가 다시 존댓말로 전환하는 등 등장인물 간 관계 설정에 실패한 사례를 확인할 수 있었다. 따라서 글의 포괄적 맥락 파악과 여기에 상응하는 번역 전략의 수립은 여전히 기계번역에게는 난제라 하겠다.

4.4 어휘 처리

문학 표현의 풍부함을 살리기 위한 핵심적 요소 중 하나는 어휘 다양성이다. 따라서 번역에 의한 어휘 다양성 감소는 주목할 가치가 있는 주제이다. 과거에 NMT 번역은 인간번역 대비 어휘 다양성이 부족하다는 보고가 있었다(Ploeger et al., 2024). 하지만 LLM과 인간의 글쓰기의 어휘 다양성에 대해서는 여전히 논쟁의 여지가 아직 남아있다(Huang et al., 2025; Kendro et al., 2025; Reviriego et al., 2024). 또한, 번역에서 LLM 기반 번역이 NMT보다 높은 어휘 다양성을 보인다는 연구는 소수에 불과하다(Kong & Macken, 2025).

본 연구에서 사용된 데이터를 사용하여 어휘 다양성을 나타내는 지표인 STTR을 워드스미스 도구를 사용하여 산출한 결과, 어휘 다양성은 NMT< 챗GPT< 인간번역 순으로 나타났다. NMT 대비 챗GPT가 어휘 다양성은 높으나, 아직 인간번역 수준의 다양성은 확보하지 못했음을 부분적으로 보여주는 데이터라 하겠다.

하지만 더욱 큰 문제는 LLM이 NMT 대비 어휘 다양성이 증가했다고 해서 무조건 번역에 긍정적인 영향을 주는 것은 아니라는 점이다. 설계의 특성상 LLM은 결정론적 성격이 NMT 대비 부족하다(Jiang et al., 2024). 즉 'A'라는 단어를 항상 '가'라는 단어로 번역을 하지 않을 가능성이 존재한다는 뜻이다. 일례로『오만과 편견』의 번역 결과물에서 가장 핵심 단어인 "pride"와 "proud"를 "교만"과 "자존심"으로 지속해서 다르게 번역한 것이 관찰되었

다.

<예시 11> 오만과 편견
원문: “Pride,” observed Mary, who piqued herself upon the solidity of her reflections, “is a very common failing, I believe.
챗GPT: “교만은,” 메리가 말했다. 그녀는 늘 자기 사색이 탄탄하다고 자부하고 했다.“아주 흔한 결점이라고 생각해요.”

여기에 더해 어휘 처리와 관련된 기계번역의 대표적인 문제점 중 하나로 OOV(out-of-vocabulary)가 지적되어 왔다(Sennrich et al., 2016). 이는 기계번역 엔진이 학습 단계에서 충분히 접하지 못했거나 전혀 노출되지 않은 어휘가 번역 과정에 등장할 경우, 해당 어휘를 적절히 처리하지 못하는 현상을 의미한다.

하지만 LLM은 NMT 대비 학습 데이터의 양이 많고, 학습 데이터에 포함이 되어 있지 않더라도 RAG(Retrieval-Augmented Generation)를 일부 활용하는 것으로 알려진 만큼 고유명사 혹은 신조어의 처리에 있어 NMT보다 유리하다. 하지만 본 연구의 대상이 된 텍스트에는 신조어가 없었기에 옥스퍼드 사전에 올해 등록된 단어를 찾고, 한국어도 2025년 유행하는 신조어를 구글에서 검색하였다. 이후 구글 제미나이로 해당 단어가 포함된 텍스트를 생성하여 파파고와 챗GPT의 번역 결과물 평가를 시행하였다.

먼저 영어 단어의 경우 예상대로 챗GPT가 파파고 대비 18대 14의 점수로 우월한 모습을 보이기는 했으나 ‘자기도 몰래 강제 퇴장을 당한다’라는 ‘shadow ban’의 의미를 다 살리지는 못하는 등 완전하지 못한 모습을 보였다.

<예시 12> 영어 신조어 번역
원문: Digital silence: the psychological impact of being shadow banned on mental health and self-perception
파파고 2025: 디지털 침묵: 그림자 금지가 정신 건강과 자기 인식에 미치는 심리적 영향
챗GPT: 디지털 침묵: 섀도우 밴으로 인한 정신 건강과 자기 인식에 대한 심리적 영향

한국어 신조어의 영어 번역에서도 결과는 마찬가지였다. 챗GPT가 파파고 대비 18:7로 압도적 우위를 보였는데, 그 이유 중 하나는 챗GPT는 해당 문장을 번역만 하는 것이 아니라 해당 단어에 대한 영어 의미까지 괄호로 추가로 설명하는 접근법을 취했기 때문이다. 하지만 챗GPT 역시 특정 단어에 대해서는 의미를 전달하지 못했다.

<예시 13> 한국어 신조어 번역
원문: 그 사람은 손절미가 넘친다.
파파고 2025: He is full of love.
챗GPT:You're good at cutting people off.

이상의 어휘 처리에 대한 논의를 종합하면 다음과 같다. 챗GPT의 어휘 다양성은 번역 과정에서 장점이자 동시에 한계로 작용할 수 있다. 먼저 어휘의 일관성을 유지하기 위해서는 사전에 단어집을 제공하거나, 특정 어휘 선택을 유도하는 별도의 프롬프트 설계가 요구된다. 실제 단어집을 제공하여 챗GPT를 학습시킨 이후에는 실험을 계속한 결과 어휘 일관성의 문제는 해결되었다.

또한 신조어와 같은 OOV 처리에서는 NMT보다 LLM이 상대적으로 유리한 측면을 보이는 것으로 판단된다. 또한, 기술 발전에 따라 학습 데이터의 규모가 확대되고 학습 주기가 단축됨에 따라 이러한 문제는 점차 완화될 가능성이 있으며 지속적인 연구를 통한 모니터링이 필요한 부분이다. 그럼에도 불구하고 현 단계에서는 기계번역 어휘 처리 결과를 절대적으로 신뢰하기는 어렵고, 인간 번역가의 검토와 개입이 여전히 필요하다고 하겠다.

4.5 형식의 유지와 프롬프트 작성 연구의 필요성

문학번역은 언어의 처리가 아니라 작품의 형식에서 나오는 미적 요소를 간직하며 의미를 전달하는 일이다. 따라서 문장의 호흡을 유지하는 것은 매우 중요하며, 시 번역에 있어서 형식의 유지 역시 매우 중요한 문제이다.

이준호(2019)는 NMT가 원문을 분할하여 처리하는 경향이 있다고 주장했는데, 본 연구의 챗GPT 데이터 역시 원문 분할에서 자유롭지는 않았다.

달리 말해 문학번역에서 형식이 가지는 미적 요소를 보존하기 위해서는 ‘원문의 형식을 그대로 유지하면서 번역하라’라는 추가적 프롬프트를 설정한다면 일부 문제가 해결될 수 있을 것이다. 여기서 한 단계 더 나아가 특정 형식을 유지하는 것이 어떠한 의미를 지니는지 학습을 시키고 결과를 생성하는 과정이 필요할 것이다. 하지만 형식의 유지에만 집중하면 번역의 결과물에 어떠한 변화가 있는지 여부는 추가적 연구가 필요한 영역이라 하겠다.

또 하나 유의할 점은 작업자가 의도한 번역 결과를 도출할 수 있는 프롬프트를 설계하는 것이 생각보다 쉽지 않다는 점이다. 예를 들어 본 연구에서도 ‘헤밍웨이의 문체로 번역해 달라’ 등 다양한 제로 샷 프롬프트를 시도했지만 연구자가 기대하는 짧고 단순한 문체를 안정적으로 생성하기는 어려웠다. 이를 보완하기 위해서는 마승혜(2025)의 연구에서 제시된 바와 같이 구체적인 번역 예시의 제공, 학습 데이터에 대한 명시적 지시, 혹은 매우 정교한 프롬프트 설계가 요구될 수 있다. 경우에 따라서는 온도(temperature)와 같은 생성 관련 하이퍼 파라미터의 조정이 필요할 것이다. 달리 말해 문학번역에 적합한 프롬프트 구현 방식은 아직까지도 활발한 논의가 진행 중인 단계라고 할 수 있다.

두 번째 문제는 프롬프트를 비교적 잘 설계하였다고 하더라도, 별도의 제약 조건을 설정하지 않는 한 프롬프트 과적합 문제가 발생할 수 있다는 점이다. 본 연구에서도 연구자가 설계한 프롬프트가 의외의 결과물로 이어진 사례가 있었다. 예를 들어 ‘이 소설은 민족주의적 성격을 지니므로 해당 측면을 강조하여 번역해 달라’고 요청할 경우, 민족주의와 직접적인 관련이 없는 문장에서도 민족주의적 요소가 과도하게 반영된 번역이 생성될 수 있다. 이러한 경우 문맥이나 문장 단위로 프롬프트를 다시 세분화하여 설정할 필요성이 있으며, 문장 단위의 정교한 프롬프트 설계는 그 자체로 매우 복잡한 작업이 될 수 있다.

따라서 단순한 프롬프트 설계를 통해서는 번역상의 모든 문제를 손쉽게 해결할 수 있다고 보기는 어렵다. 더 나아가 과도하게 복잡한 프롬프트 구성은 기계번역 활용의 핵심적 장점인 생산성 향상에 오히려 반하는 결과를 초래할 수 있다는 점 역시 유의할 필요가 있으며 추가적 연구가 필요한 영역이라 하겠다.

5. 결론

번역 기술은 지속이고 비약적 발전을 이뤄왔다. 이러한 맥락에서 최근 등장한 LLM은 기존의 NMT에 비해 문맥 이해 및 표현의 유연성 측면 등에서 상대적으로 우수한 번역 결과물을 생성한다고 평가받고 있다. 또한, 기존의 연구에서는 LLM을 활용한 규칙 설정 및 조건 부여 방식 그리고 방대한 언어 데이터 학습량 덕분에 LLM을 통해 문학번역이 부분적으로 가능할 것이며, 번역 실무자에게 실질적인 보조 수단으로 작용할 가능성을 제시하였다.

그러나 새로운 기술의 합리적이고 안전한 활용을 위해서는 기술의 긍정적 발전을 과도하게 일반화하기보다는 신중한 분석으로 그 영향과 한계의 균형을 잡는 접근법이 필요하다. 특히 문학번역은 의미 전달을 넘어 미묘한 정서, 상징, 언어 고유의 리듬과 미학을 다루는 섬세한 행위라는 점에서 기술 적용에 따른 여파는 면밀한 검토를 요한다. 따라서 효율성의 향상이라는 혜택뿐 아니라, 번역 오류의 영향도 및 번역 행위의 본질적 가치가 훼손될 가능성에 대해서도 충분한 논의와 비판적 성찰이 필요하다.

이에 본고는 과거 NMT에서 생성된 데이터를 현재의 NMT 및 챗GPT 번역 결과물과 비교하고, 현시점의 NMT와 챗GPT 번역 결과물을 비교 평가하는 접근을 취했다. 분석 결과 챗GPT가 NMT 대비 더 우수한 문학번역 결과물을 생성함을 관찰했고, LLM의 문학번역 적용 가능성을 일부 긍정적으로 평가하였다.

하지만 그 과정에서 아직 해결되지 않은 기술적 난제도 관찰되었다. 특히 여전히 사라지지 않은 누락과 언제 등장할지 예측하기 어려운 결과물의 과생성과 오생성에 본고는 주목하였다. 또한, 어휘 측면에서 표현의 풍부함을 보증할 다양성, 독자의 읽기 경험과 직결되는 일관성, 바로 오역으로 이어질 수 있는 미학습 어휘의 처리 등에 대해 다루었다. 무엇보다 맥락을 파악하여 문장 내 그리고 문장 간의 숨은 의미 파악에 있어 여전히 존재하는 한계점, 원문에 생략된 주어 등의 복원에 대한 문제, 전체 맥락을 파악하여 인물 간의 관계 설정과 어역 및 어투 조정의 문제를 논의하였다. 마지막으로 형식의 유지를 위한 거시적 번역 전략 설계의 부재와 프롬프트 설정의

과적합 문제에 대해 우려를 제기하였다.

이상의 분석을 통해 기계번역은 매우 큰 기술적 발전에도 불구하고 여전히 일차적 결과물을 제시하는 도구이며, 포스트에디팅을 수행한다면 주의 깊게 진행이 필요하다는 것이 현시점의 진단이다. LLM 역시 입력된 데이터와 사전 학습된 언어 패턴을 토대로 가장 개연성 높은 출력을 생성할 뿐, 작품 전체를 관통하는 미학적 의도나 작가의 개별적 세계관을 능동적으로 해석하고 재구성하는 주체로 기능하지 못한다. 오히려 기계가 결과물 생성의 기본적인 틀을 제공하는 과정에서, 번역가의 창조적 개입과 해석의 여지를 제한할 가능성도 내포한다.

따라서 기계번역의 활용은 문학번역의 주체를 대체하는 방향이 아니라, 번역가의 비판적 판단과 창조적 결정을 보조하는 도구로서 제한적인 방향으로 설정해야 할 것이다. 기계번역이 보조적 도구로 사용되는 경우와 번역 행위의 중심 주체로 기능하는 경우는 근본적으로 다른 '번역 활동'이 이뤄지기 때문이다. 따라서 번역 기술을 꼭 사용해야 한다면, CAT 도구 내에서 번역가가 기계번역 결과물을 일부 참조하며 번역하는 것이 중립적 방안일 수 있음을 제안한다.

다만 본문에서도 언급하였듯이, 앞서 제기된 기계번역의 한계 중 일부는 프롬프트 엔지니어링, 파라미터 조정, 학습 데이터의 변화, 지도 학습 방식의 개선 등을 통해 비교적 단기간에 부분적으로 개선될 가능성이 있다. 물론 일부 한계는 장기적인 기술 발전과 데이터 축적을 필요로 할 것이다. 이러한 변화 속에서 번역학자는 특정 시점에서 지적된 한계를 어떻게 해소할 것인가를 탐구하고, 어떻게 문제가 해결되어 가는지를 지속적으로 관찰해야 할 것이다. 이러한 탐구를 통해 그 결과를 연구, 실무, 교육 차원에서 환류시키는 역할을 수행해야 할 것이다.

본고는 새로운 번역 기술의 흐름 속에 제기되고 있는 문학번역에 번역 기술 적용 가능성과 한계를 동시에 고찰한 예비 연구로서 의의가 있다. 하지만 소규모 데이터에 기반한 해석이기 때문에 본고의 결과를 결정론적으로 받아들이기에는 무리가 있다. 따라서 대규모 텍스트를 활용하고 더욱 체계적인 품질 평가 매트릭스를 활용하여 차기 연구를 수행하여 평가의 객관성을 높일 필요가 있다. 이러한 단점에도 불구하고 본 연구가 문학번역에서

기계번역을 고려하는 모든 이에게 잠시나마 문학번역의 본질과 기술의 현황을 다시 한번 생각해볼 계기를 마련했고, 다른 연구자들에게 문학번역의 기계번역 적용에 대한 탐구를 본격화하는 시발점을 제공했다면 본고는 그 목적을 다했다 하겠다.

마지막으로 문학번역 과정에서 기계번역의 사용을 허용하는 것이 과연 문학번역의 발전에 기여할 수 있는가라는 보다 근본적인 질문을 제기하고자 한다. 만약 기술적 편의가 번역가의 해석 능력과 창의성을 대체하거나 약화시키는 방향으로 작용한다면, 장기적으로는 문학번역의 질적 성장이 오히려 저해될 가능성도 배제할 수 없다. 따라서 기계번역의 활용은 효율성 제고라는 측면뿐 아니라, 문학번역의 본질적 가치와의 관계 속에서 신중하게 검토해야 할 것이다.

참고문헌

<1차 자료>

Austen, J. (2003). 오만과 편견 (윤지관, 전승희, 공역). 민음사. (Original work published 1813)

Hemingway, E. (1994). 노인과 바다 (이종한, 역). 교육문화연구회. (Original work published 1952)

한강. (2007). 채식주의자. 창비.

현진건. (1921/2012). 빈처. 보물창고

현진건. (1924/2012). 운수 좋은 날. 보물창고

<2차 자료>

곽순례. (2022). 문학텍스트 아-한 기계번역의 가용성과 한계 연구. 한국이슬람학회 논총, 32(3), 249-276.

김용출. (2023. 5. 20.). 하루 만에 소설 '뚝딱' 편집·번역 '척척'… 출판계 뒤흔드는 AI. 세계일보. https://www.segye.com/newsView/20230519516862

남철진. (2025). 중국 소설의 GPT 번역 문제점-원문에 충실한 번역을 중심으

로. 중어중문학, 101, 351-382.
노금송, 왕원. (2024). 한국어 문화소에 관한 기계번역과 인간번역의 비교 분석 연구. 한국학연구, 73, 299-318.
마승혜. (2018). 문학작품 기계번역의 한계에 대한 상세 고찰. 통번역학연구, 22(3), 65-88.
마승혜. (2024). AI 문학번역, 어디까지 가능한가 — 챗 GPT 가 번역한 한국 문학작품 제목의 기능 분석을 중심으로. 번역학연구, 25(3), 57-85.
마승혜. (2025). AI 페르소나가 문학 번역에 미치는 영향-챗 GPT 프롬프트에 따른 번역 사례 비교· 분석. 통번역학연구, 29(4), 1-35.
박수정, 최은실. (2023). 챗GPT의 아이러니 번역 활용 가능성 고찰. 번역학연구, 24(2), 131-160.
이준호. (2019). 문학번역 적용을 위한 기계번역의 현주소. 통번역학연구, 23(1), 143-167.
이창수. (2019). 문학번역에서의 기계번역과 인간번역 문체에 대한 전산문체학적 비교 연구. 번역학연구, 20(2), 111-130.
이창수. (2021). 기계학습 알고리즘을 활용한 문학번역에서의 기계 번역과 인간 번역 결과물 분류 연구. 번역학연구, 22(1), 199-217.
이창수. (2024). 챗GPT, 파파고, 인간 번역가 간의 한영 문학번역 차이점 연구. 번역학연구, 25(2), 11-37.
이현주. (2022). 문학작품의 중-한 기계번역 결과의 결속구조 분석-인간번역과의 비교를 중심으로. 중어중문학, 87, 259-282.
임진. (2025). 성중립 표현의 AI 기반 기계번역 양상: GPT-4o, Gemini, DeepL 문학 번역을 중심으로. 통번역학연구, 29(3), 61-84.
홍서연. (2025). AI 번역기의 한러 번역성능 비교-파파고, 구글, 챗 GPT를 중심으로. 통번역학연구, 29(1), 207-233.
황지연, 이미령, 원다인. (2024). 표현적 텍스트의 기계 번역 활용 가능성 고찰-K-pop 그룹 뉴진스 노래 가사 번역을 중심으로. 통번역학연구, 28(1), 177-207.
Appicharla, R., Gain, B., Pal, S., & Ekbal, A. (2025). *Beyond the sentence: A survey on context-aware machine translation with large language models*.

arXiv. https://doi.org/10.48550/arXiv.2506.07583

Bayer-Hohenwarter, G., & Kussmaul, P. (2020). Translation, creativity and cognition. In F. Alves & A. Jakobsen (Eds.), *The Routledge handbook of translation and cognition* (pp. 310-325). Routledge.

Castaldo, A., Castilho, S., Moorkens, J., & Monti, J. (2025). *Extending CREAMT: Leveraging large language models for literary translation post-editing*. arXiv. https://doi.org/10.48550/arXiv.2504.03045

Corpas Pastor, G., & Noriega-Santiáñez, L. (2024). Human versus neural machine translation creativity: A study on manipulated MWEs in literature. *Information*, *15*(9), 530.

Du, S., Guerberof-Arenas, A., Toral, A., Gerrits, K., & Borillo, J. M. (2025, June). Optimising ChatGPT for creativity in literary translation: A case study from English into Dutch, Chinese, Catalan and Spanish. *Proceedings of Machine Translation Summit XX: Volume 1*, 578-591.

Gao, R., Lin, Y., Zhao, N., & Cai, Z. G. (2024). Machine translation of Chinese classical poetry: a comparison among ChatGPT, Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications*, *11*(1), 1-10.

Ghosh, S., & Caliskan, A. (2023). ChatGPT perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across Bengali and five other low-resource languages. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 901–912.

Guerberof-Arenas, A., & Toral, A. (2022). Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*, *11*(2), 184-212.

Guerreiro, N. M., Alves, D. M., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., & Martins, A. F. (2023). Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, *11*, 1500-1517.

Hermans, T. (2007). Literary translation. In P. Kuhiwczak, K. Littau, S. Bassnett

& E. Gentzler (Eds.), *A companion to translation studies* (pp. 77-91). Multilingual Matters & Channel View Publications.

Huang, Y., Li, D., & Cheung, A. K. (2025). Evaluating the linguistic complexity of machine translation and LLMs for EFL/ESL applications: An entropy weight method. *Research Methods in Applied Linguistics*, *4*(3), 100229.

Jia, Y., Carl, M., & Wang, X. (2019). How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. *The Journal of Specialised Translation*, *31*(1), 60-86.

Jiang, Z., & Zhang, Z. (2024). *Can ChatGPT rival neural machine translation? A comparative study*. arXiv. https://arxiv.org/html/2401.05176v1

Jiang, Z., Lv, Q., Zhang, Z., & Lei, L. (2024). *Convergences and divergences between automatic assessment and human evaluation: Insights from comparing ChatGPT-generated translation and neural machine translation*. arXiv. https://arxiv.org/pdf/2401.05176

Karabayeva, I., & Kalizhanova, A. (2024). Evaluating machine translation of literature through rhetorical analysis. *Journal of Translation and Language Studies*, *5*(1), 1-9.

Karakanta, A., Nas, M., & Dorst, A. G. (2025). Metaphors in literary machine translation: Close but no cigar? *Proceedings of Machine Translation Summit XX: Volume 1*, 276-286.

Kendro, K., Maloney, J., & Jarvis, S. (2025). *Do LLMs produce texts with" human-like" lexical diversity?*. arXiv. https://doi.org/10.48550/arXiv.2508.00086

Koh, S. (2023). An analysis of ChatGPT's language translation based on the Korean film Minari. *Journal of English Teaching through Movies and Media*, *24*(4), 1-14.

Kong, D., & Macken, L. (2025). *Decoding machine translationese in English-Chinese news analyst: LLMs vs. NMTs*. arXiv. https://doi.org/10.48550/arXiv.2506.22050

Läubli, S., Sennrich, R., & Volk, M. (2018). *Has machine translation achieved*

human parity? A case for document-level evaluation. arXiv. https://doi.org/10.48550/arXiv.1808.07048

Lefevere, A. (2016). *Translation, rewriting, and the manipulation of literary fame*. Routledge.

Manakhimova, S., Avramidis, E., Macketanz, V., Lapshinova-Koltunski, E., Bagdasarov, S., & Möller, S. (2023). Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can ChatGPT Outperform NMT?. *Proceedings of the Eighth Conference on Machine Translation*, 224-245.

OpenAI. 2023. GPT-4 technical report. CoRR, abs/2303.08774.

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*. arXiv. https://doi.org/10.48550/arXiv.2303.13780

Ploeger, E., Lai, H., Van Noord, R., & Toral, A. (2024). *Towards tailored recovery of lexical diversity in literary machine translation*. arXiv. https://doi.org/10.48550/arXiv.2408.17308

Pouget-Abadie, J., Bahdanau, D., Van Merrienboer, B., Cho, K., & Bengio, Y. (2014). *Overcoming the curse of sentence length for neural machine translation using automatic segmentation*. arXiv. https://doi.org/10.48550/arXiv.1409.1257

Reviriego, P., Conde, J., Merino-Gómez, E., Martínez, G., & Hernández, J. A. (2024). Playing with words: Comparing the vocabulary and lexical diversity of ChatGPT and humans. *Machine Learning with Applications*, *18*, 100602.

Rojo, A. (2017). The role of emotions. In J. Schwieter & A. Ferreira (Eds.), *The handbook of translation and cognition* (pp. 369-385). John Wiley & Sons.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers)*,

1715-1725.

Toral, A., & Way, A. (2018). What level of quality can neural machine translation attain on literary text?. In J. Moorkens, S. Castilho, F. Gaspari & S. Doherty (Eds.), *Translation quality assessment: From principles to practice* (pp. 263-287). Springer International Publishing.

Venuti, L. (2017). *The translator's invisibility: A history of translation*. Routledge.

Wang, Y. (2024). The impact of technology on human translators and translation quality: a study on machine translation and computer-assisted translation tools. *English Linguistics Research*, *13*(1), 1-19.

Wang, M., Vu, T., Zhao, J., Shiri, F., Shareghi, E., & Haffari, G. (2024). Simultaneous machine translation with large language models. *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, 89-103.

Yang, Z., Cheng, Y., Liu, Y., & Sun, M. (2019). Reducing word omission errors in neural machine translation: A contrastive learning approach. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6192-6196.

Yu, J. S., & Yao, Y. (2026). Reconstructing Translation Process via LLM Prompt Engineering. In J. S. Yu & Y. Yao (Eds.), *Intelligent language services: Theory and practice with large language models* (pp. 143-197). Springer Nature Singapore.

Zhang, H., Chen, K., Bai, X., Xiang, Y., & Zhang, M. (2024). *Paying more attention to source context: Mitigating unfaithful translations from large language model*. arXiv. https://doi.org/10.48550/arXiv.2406.07036

Zhang, R., Zhao, W., & Eger, S. (2025). How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 10961-10988.

A preliminary study on the application of machine translation to literary translation: Focusing on ChatGPT and NMT outputs

Junho Lee
Advanced Interpretation & Translation Program, Chung-Ang University

Abstract

Recent advances in machine translation have expanded its perceived applicability beyond informational texts to the domain of literary translation, particularly through the use of large language models with prompt-based interaction. However, literary translation differs fundamentally from general translation in that it requires the recreation of aesthetic form, cultural context, and narrative voice rather than the mere transfer of propositional meaning. Accordingly, this study investigates the extent to which neural machine translation (NMT) and ChatGPT-based translation have developed in comparison with earlier machine translation outputs. In addition, it examines whether ChatGPT-based translation outperforms NMT in literary translation tasks through automatic evaluation, expert assessment, and quantitative error analysis. The results indicate that ChatGPT demonstrates overall improvements in translation quality and stronger contextual understanding, suggesting its potential usefulness as a supportive tool in literary translation. Nevertheless, persistent limitations remain, including omissions, overgeneration, lexical inconsistency, inadequate handling of out-of-vocabulary items, and difficulties in conveying implicit meanings and maintaining global narrative coherence. These findings indicate that, despite recent technological advances, machine translation should be regarded as a supplementary aid rather than a substitute for human literary translators.

Keywords: Literary translation; machine translation; NMT; LLM; ChatGPT
키워드: 문학번역, 기계번역, 신경망번역, 거대언어모델, 챗GPT

이준호(https://orcid.org/0000-0003-0397-6829)
중앙대학교 국제대학원 전문통번역학과 조교수
brandnon4tni@cau.ac.kr

논문 투고일: 2026년 2월 11일
1차 심사 완료일: 2026년 3월 2일
2차 심사 완료일: 2026년 3월 8일
게재 확정일: 2026년 3월 16일