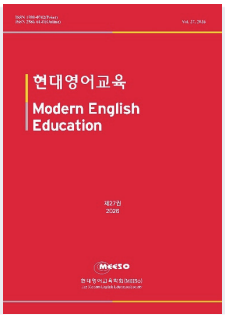
Modern English Education, Vol. 27, (2026) pp. 396-411
doi: https://doi.org/10.18095/meeso.2026.27.1.396



Modern English Education

ISSN: 1598-0782 / e-ISSN: 2586-6141
http://journal.meeso.or.kr

텍스트 유형별 기계번역 산출물에 대한 학습자와 훈련된 평가자 평가 및 기계번역 활용 읽기에 대한 학습자 인식

김성연 (한양대학교)

Received: 4 August 2026
Revised: 31 August 2026
Accepted: 15 September 2026

Kim, Sung-Yeon. (2026). Learner and trained rater evaluation of machine translation (MT) across text types and learner perceptions of MT-assisted reading. *Modern English Education*, *27*, 396-411.

Keywords
Machine translation, translated output, reading text, text type, machine translation adequacy evaluation
기계번역, 번역 산출물, 읽기 지문, 텍스트 유형, 기계번역 적합성 평가

Sung-Yeon Kim
Professor
Department of English Education
Hanyang University
sungkim@hanyang.ac.kr
ISNI: 0000 0004 6489 2060

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea(NRF-2022S1A5A2A01047853).

Abstract

This study investigated whether and to what extent EFL college learners and a trained bilingual rater differed in evaluating the adequacy of machine translation (MT) across four text types: English for Specific Purposes (ESP), argumentative (Opinion), expository (Information), and literary (Literature). A mixed-methods design was adopted to examine translation adequacy ratings and learner perceptions of MT-assisted reading. At the macro level, the two groups showed close agreement: both rated Opinion and ESP as highly adequate (Rater $M = 4.63$, 4.55; Student $M = 4.29$, 4.41) and Literature as the most vulnerable domain (Rater $M = 3.64$; Student $M = 4.06$). Despite this shared rank order, a notable disparity emerged at the micro level. The rater scored Literature more critically owing to its stylistic complexity, whereas the learners rated it more leniently, a pattern that may reflect the surface-level readability of the text rather than the quality of its translation. Survey results also revealed a mismatch between perceived accuracy and actual preference: although ChatGPT was rated as the most accurate tool by 62% of participants, more learners preferred Papago (57.1%) over ChatGPT (42.9%) due to its L1-optimized interface. Furthermore, significant affective trade-offs emerged as the learners viewed MT as a helpful scaffold for reading, while simultaneously holding mixed feelings of trust and skepticism toward MT output. The study concludes with pedagogical implications for developing critical MT literacy in EFL curricula.

서론

최근 생성형 인공지능(Generative Artificial Intelligence, AI)와 대형언어모델(Large Language Model, LLM) 기술의 폭발적인 성장은 외국어 교육 패러다임 전반에 실질적인 대변혁을 이끌어내고 있다. 이러한 진화는 언어교육 분야에도 고스란히 투영되어, 인공지능과 기계학습(Machine Learning) 알고리즘을 독해 및 작문 지도에 통합하려는 실증적 시도로 이어지고 있다. 구글은 2006년에 온라인 번역기 Google Translate(GT)를 출시했는데, 초창기의 GT는 구절 기반 알고리즘으로 방대한 디지털 데이터에서 단어 쌍을 분석하는 프로그램이었다. 2016년에는 인공신경망

구조에 기반한 구글 신경 기계번역(Neural Machine Translation, NMT) 시스템이 도입되어, 기존 GT보다 향상된 정확성을 보이게 되었다(Alkhofi, 2025; Sun, 2017). GT 기술의 발전은 모국어(L1)와 목표어(L2) 간 장벽을 낮추며 L2 학습자들이 원문 텍스트에 접근할 수 있는 새로운 가능성을 열어주었다(Karnal & Pereira, 2015; Stapleton & Leung, 2019). 이후 GT, 파파고, ChatGPT와 같은 기계번역(Machine Translation, MT) 프로그램들은 한국인 EFL 학습자의 영어 읽기 및 학습 과정에서 일상적인 도구로 널리 사용되고 있다.

그러나 MT 기술의 비약적인 발전에도 불구하고 국내외 선행연구의 상당수는 L2 쓰기 산출물의 품질 평가나 쓰기 학습 과정에서의 MT 활용 효과에 편중되어 있는 실정이다. 상대적으로 L2 읽기 과정에서 MT가 어떤 도움을 줄 수 있는지 집중적으로 조명한 연구는 매우 드물다. 구체적으로 학습자가 독해 과업 수행 시 MT를 어떻게 활용하는지, MT 사용 경험이 그 신뢰도와 유용성에 대한 인식에 어떠한 영향을 미치는지, 그리고 무엇보다 읽기 자료의 유형에 따라 MT에 대한 신뢰도가 어떻게 달라지는지를 실증적으로 분석한 연구는 부족하다(E. S. Chung, 2023). 이러한 공백은 대부분의 선행연구가 학습자나 교사를 대상으로 한 일회성 설문 기반의 자기보고(self-report) 데이터에 의존해 왔다는 방법론적 한계와도 맞닿아 있다. 그 결과 학습자가 MT의 언어적 유창성 이면에 존재하는 심층적 오역이나 텍스트 유형에 따른 질적 차이를 실제로 얼마나 정확하게 지각하고 있는지는 별도로 검증되지 못한 채 남아 있다.

또한 텍스트는 유형에 따라 번역 알고리즘의 처리 난이도와 산출물의 질적 수준에서 체계적인 차이를 보일 수밖에 없다. 특히 문학이나 창작 텍스트는 함축적 의미, 문체적 효과, 상호 텍스트성에 크게 의존하기 때문에 통계 기반 및 신경망 기반 번역 시스템이 다루기 어려운 유형의 글로 보고되어 왔다(Cespedosa & Mitkov, 2023; Toral & Way, 2018). 이에 반해 논설문이나 설명문, 특수목적 영어(ESP)와 같이 상대적으로 정형화된 구조와 명시적 어휘를 사용하는 텍스트는 번역 알고리즘이 처리하기에 비교적 용이할 것으로 예상되지만, 이러한 격식성, 비유적 표현의 정도에 따른 번역 난이도의 차이가 학습자의 주관적 인식 수준에서도 동일하게 감지되는지, 그리고 그것이 훈련된 평가자가 판정한 실제 번역 품질과 일치하는지는 아직 실증적으로 검증된 바가 없다.

이에 본 연구는 일회적 인식 조사의 한계를 극복하고 연구의 타당도를 제고하기 위해 다원적(triangulated) 설계를 통해 다양한 자료를 수집, 분석하고자 한다. 구체적으로 논설문(opinion), 문학(literature), 설명문(information), 특수목적 영어(ESP) 등 텍스트 유형별 산출물에 대한 학습자 평가 데이터와 훈련된 이중언어 평가자의 산출물 평가 결과를 비교하고, MT 활용 읽기에 대한 학습자의 관점이 과업 수행 경험 전후로 어떻게 달라지는지를 함께 분석하고자 한다. 연구목적을 연구 질문으로 구체화하면 다음과 같다.

1) 한국인 대학생 학습자가 지각하는 기계번역 산출물의 적합성은 읽기 텍스트의 유형에 따라 차이가 있는가?
2) 기계번역의 신뢰도에 대한 학습자 인식은 읽기 텍스트의 MT 산출물 평가 과업 수행 전후에 차이가 있는가?
3) 텍스트 유형별 번역 산출물에 대한 훈련된 평가자의 평가는 학습자 집단의 평가와 차이가 있는가?

선행연구

기계번역과 언어교육

최근 언어교육 환경에서 MT 테크놀로지를 교수학습 과정에 활용할 수 있는 가능성에 대한 관심이 급증함에 따라 관련 연구들이 다각도로 전개되어 왔다. 선행연구들은 EFL 학습자들이 시간을 절약하고 L2 수행 능력의 부족을 보완하기 위해 MT를 활용한다고 그 동기를 설명한다. Niño(2009)는 번역 도구의 높은 이용 가능성, 번역의 즉각성, 지원 언어의 다양성, 그리고 어휘나 단순한 통사 구조의 글 번역 시 보이는 높은 정확성을 MT의 교육적 장점으로 보고하였다. Bahri와 Mahadi(2016)는 17명의 말레이시아 학생을 대상으로 한 조사에서 학습자들이 MT를 언어 학습의 유용한 보조 도구로 인식하며 수용하는 경향이 있음을 발견하였고, Briggs(2018) 역시 80명의 대학생 대상 설문에서 약 49%가 온라인 번역을 언어학습 도구로서 충분히 가치 있다고 평가하며 빈번하게 사용한다고 보고하였다.

그러나 이러한 실용적 이점과 학습자들의 높은 요구에도 불구하고, 전통적인 학교 현장에서는 기계번역의 정확성과 신뢰도에 대한 우려 때문에 학술적 활용에 대해 회의적인 관점을 유지해 왔다(Clifford et al., 2013; Groves & Mundt, 2015; Jolley & Maimone, 2015; Somers et al., 2006). 교강사의 인식은 연구에 따라 엇갈리는데, GT의 활용이

학습자의 언어 능력 향상에 도움이 될 수 있다고 긍정적으로 평가한 Stapleton과 Leung(2019)의 연구와 달리, Clifford 등(2013)은 교강사들이 MT가 학습에 미치는 긍정적 영향에는 회의적인 반응을 보인다고 보고하였다. 초기 연구들은 MT 출력물이 여전히 일부 어휘와 구조 면에서 근본적인 모호함을 내포하고 있다고 지적하였으며(Somers, 2011), 숙달도가 낮은 학습자가 번역기에 과잉 의존할 경우 인지적 자율성이 저해되고 궁극적으로 L2 회피 행동으로 이어질 수 있다고 경고하였다(Ducar & Schocket, 2018). 나아가 윤리적 측면에서 문단이나 지문 전체를 무비판적으로 기계 번역하여 제출하는 행위는 학술적 정직성을 훼손하고 표절을 조장하기 쉽다는 우려도 제기되었다. Somers 등(2006)은 직접 작성한 글과 기계번역 산출물을 단순히 비교하는 것 자체가 불공평하다고 비판하였고, Jolley와 Maimone(2015) 역시 지문 전체를 통째로 기계 번역하는 것은 비윤리적이라는 교사들의 부정적 평가를 보고하였다. Groves와 Mundt(2015)는 학생들이 모국어로 작성한 글과 GT로 기계 번역한 글을 비교한 결과, MT 산출물이 담화공동체(discourse community)가 요구하는 기준에 미달하며 오류가 심각하다고 보고하면서, 번역기로 영역한 글의 작성자를 진정한 원저자(original author)로 규정할 수 있는지에 대해 의문을 제기하였다.

그러나 최근 NMT 및 생성형 AI 기술의 발전은 이러한 기존의 비판적 평가에 반하는 결과를 다수 제시하며 인식의 전환을 요구하고 있다. 중국의 성인 EFL 학습자 124명을 대상으로 한 연구에서 GT를 통해 산출된 글이 단어 수, 철자 및 문법 오류, 단어별 오류 등 여러 지표에서 학생이 직접 쓴 글보다 우수한 것으로 나타났으며(Tsai, 2019), O'Neill(2016)의 실험에서도 온라인 번역기를 사용한 학생들이 이해도, 내용, 문법, 철자 등에서 대조군보다 높은 점수를 얻었다. Stapleton과 Leung(2019) 역시 홍콩 초등학생을 대상으로 한 연구에서 GT를 사용해 작성한 글이 직접 영작한 글보다 문법적 정확성 면에서 우수함을 확인하였다. 한국인 학습자를 대상으로 기계번역 활용 여부에 따른 작문의 유창성, 정확성, 복잡성 차이를 분석한 연구에서는 MT가 정확성 향상에는 기여하지만 통사적, 어휘적 복잡성에는 뚜렷한 도움을 주지 못하며, 유창성에는 유의한 변화가 없음을 확인하였다(E. Chung & S. Ahn, 2022). 아울러 그 효과는 학습자 수준에 따라 상이하게 나타났는데, 수준이 낮은 학습자는 통사적 복잡성 측면에서, 수준이 높은 학습자는 어휘적 복잡성 측면에서 상대적으로 더 큰 도움을 받는 것으로 나타났다. 학습자 수준에 따른 효과 차이는 학습자가 기계번역을 사용하는 목적 자체가 다른 데서 비롯되었을 가능성을 시사한다. 실제로 숙달도에 따른 기계번역 활용 양상을 살펴본 연구(S. Jang & S. Jwa, 2024)에서도 수준이 높은 학습자는 논리적 흐름과 담화적 연결을 점검하기 위한 목적으로, 수준이 낮은 학습자는 문법과 어휘를 확인하기 위한 목적으로 기계번역을 활용하는 경향이 있음을 발견하였다.

최근에는 LLM 기반 도구와 기계번역을 비교하는 시도들이 많이 이루어지고 있는데, 한국어-영어 대화체 번역 오류 분석 연구에서 ChatGPT는 표본 문장의 약 58%에서 정확한 번역을 산출한 반면 파파고는 약 38%에 그쳤으며, 어휘 오류 역시 ChatGPT(190건)에 비해 파파고(301건)가 약 1.6배 더 많이 발생하였다(J. Lee, 2025). 경찰 대화문을 대상으로 한 별도의 AI 번역기 평가에서도 ChatGPT가 구글번역, 파파고보다 훨씬 높은 평가를 받았다(S. Hong, 2025). 이처럼 MT 도구는 접근성과 편의성에 힘입어 학생들의 필수 학습 도구로 자리 잡는 추세이며 번역 기술이 신경망 기반에서 LLM 기반으로 전환됨에 따라 도구 간 성능 격차 역시 계속 갱신되고 있다(Alhaisoni & Alhaysony, 2017; Alkhofi, 2025).

기계번역과 L2 읽기

이상의 연구들을 통해 기존 MT 연구의 대부분이 교사와 학습자의 일회적 인식을 파악하는 설문조사에 국한되거나, 쓰기 과업을 일회적으로 수행하게 한 후 산출물의 언어적 요소를 분석하는 데 집중되어 있음을 알 수 있다. 상대적으로 읽기 지도 맥락에서 MT의 활용 방안을 다룬 연구는 극히 제한적이다(E. S. Chung, 2023). Chung(2023)은 L2 독해 과정에서의 MT 활용을 다룬 선행연구를 검토하며, 구글 번역 사용 여부에 따라 읽기 전략 사용의 차이를 비교한 Karnal과 Pereira(2015)의 연구가 "the only study that has specifically examined this topic"이라고 밝혀, 읽기 이해 맥락에서 MT를 정면으로 다룬 연구가 사실상 전무하다시피 함을 확인하였다(p. 2).

다만 최근에는 MT가 L2 독해 과정에서 수행하는 역할이 점차 부각되고 있다. 학습자들은 독해 상황에서 MT를 텍스트 이해를 돕는 도구로 활용하고 있으나, 그 효과에 대한 인식은 연구에 따라 차이가 있다. Chung(2023)은 한국인 상급 영어 학습자 30명을 대상으로 문학 텍스트 독해 시 원문(source text, ST)과 파파고의 비편집 번역(raw MT, RMT) 산출물의 제시 순서를 달리하여 그 효과를 검증하였다. 그 결과, RMT를 ST보다 먼저 제시받은 학습자들이 ST보다 RMT에 유의하게 더 많이 집중했으며, 이는 번역 오류에 대한 높은 수용도와 세부 내용 회상(recall)의 정확도 저하로 이어지는 것으로 나타났다. 이에 근거하여 Chung(2023)은 ST를 우선적으로 읽고 기계번역은 이후 확인용 보조

자료로만 활용할 것을 제안하였다. 다만 이 연구는 문학 장르 한 편, 상급 학습자 집단만을 대상으로 하였기에, 연구자도 “the results may look different depending on the proficiency level, text difficulty, and text genre”라며 텍스트 유형에 따른 일반화 가능성의 한계를 인정하였다(p. 16).

그러나 이러한 긍정적 인식이 모든 연구에서 일관되게 나타나는 것은 아니다. Klimova(2025)는 체코 및 국제 학생 99명을 대상으로 한 설문에서 응답자의 61.6%가 MT 사용 후에도 외국어 학습 동기를 유지했지만 38.4%는 오히려 동기가 저하되었다고 보고하였으며, 학생들은 MT가 문화적 뉘앙스와 맥락을 정확히 전달하지 못한다는 점을 주요 한계로 지적하며 MT 사용에 대한 명확한 제도적 가이드라인의 필요성을 요구하였다고 밝혔다. 영어전용 강의 환경에서 실시한 면담 연구에서도 학생들은 독해력 향상, 어휘 확장, 읽기 속도 증가를 위해 MT를 적극 활용하는 것으로 나타났으나, 같은 연구에서 교강사들은 학생의 번역 산출물에서 나타나는 오역이 원문에 대한 불충분한 몰입을 반영하며 언어 간 표절로 이어질 수 있다는 우려를 제기하였다(High et al., 2025). 학생과 교강사 모두 MT를 유용한 도구로 인식하면서도, 무비판적이고 상시적인 사용이 언어능력 발달을 저해할 수 있다는 양가적 태도를 보인 것이다. 이처럼 MT의 읽기 지원 효과에 대한 인식은 긍정적 평가와 우려가 공존하며 연구마다 상이하게 나타나지만, 그 기저에는 공통된 방법론적 한계가 자리한다. 즉 이러한 주관적 인식이 실제 번역 산출물의 질과 얼마나 부합하는지는 별도로 검증되어야 할 문제로 남아 있는 것이다. 기존 연구들은 대부분 일회성 설문 기반의 자기보고 데이터에 의존하여, 학습자가 MT의 언어적 유창성 이면에 존재하는 심층적 오역이나 텍스트 유형에 따른 질적 차이를 실제로 검증하지 못했다.

이러한 방법론적 한계는 MT 텍스트 단순화 효과를 다룬 최근 연구에서도 확인된다. Shih(2022)는 통제언어(controlled language)를 활용한 사전 편집(pre-editing)이 신경망 기계번역 산출물의 가독성에 미치는 효과를 텍스트 분석과 설문조사를 통해 실증적으로 확인하였다. 그 결과 어휘, 구문 측면에서 단순화된 통제 텍스트의 번역물에 90점 이상을 부여한 응답 비율이 45%로 비통제 텍스트의 10%를 크게 상회하였고, 어려운 어휘와 문법 오류를 이해 장애 요인으로 지목한 비율 역시 통제 텍스트에서 뚜렷하게 낮게 나타났다. 그러나 이 결과는 응답자의 주관적 평가에 기반한 기술통계일 뿐이며, 원문 코퍼스의 객관적 가독성 지표나 훈련된 평가자에 의한 문장 단위 번역 품질 평가와의 교차 검증은 이루어지지 않았다는 한계를 지닌다. 즉 학습자가 이해하기 쉽다고 느낀 텍스트와 실제로 정확하게 번역된 텍스트가 반드시 일치하는지는 확인되지 않은 채 남아 있는 것이다.

이러한 인식-품질 간 괴리는 학습자의 자기보고에서뿐 아니라 자동평가지표 자체의 한계에서도 비롯된다. Scarton과 Specia(2016)는 이중언어 대역평가(Bilingual Evaluation Understudy, BLEU)와 같은 전통적 자동평가지표가 동일한 MT 시스템이 생성한 문서들에 대해 서로 비슷한 점수를 부여하는 경향이 있어 문서 단위의 절대적 품질을 변별하는 데 신뢰하기 어렵다는 문제의식에서 출발하였다. 이들은 독일어 원문을 여러 MT 시스템(Google Translate, Bing, Systran, 통계기반 시스템 Moses)으로 영어 번역한 뒤, 영어에 유창한 화자들이 해당 번역문에 대한 독해 문제에 답한 결과를 문서 수준의 품질 점수로 전환하는 방법을 제안하였다. 실제로 이들이 산출한 BLEU 점수는 동일 시스템 내에서 표준편차가 매우 낮게 나타나, BLEU 점수만으로는 같은 시스템이 생성한 문서들 간 질적 차이를 변별하기 어렵다는 점이 확인되었다. 이는 자동화된 지표든 학습자의 자기보고든, 단일한 평가 방식만으로는 번역 산출물의 실제 품질을 온전히 포착하기 어려움을 시사한다.

이러한 인식-품질 간 괴리는 평가자 집단에서도 유사하게 나타난다. Alkhofi(2025)는 사우디 대학의 번역 전공 학생과 GT가 각각 산출한 영어-아랍어 번역물을 22명의 교수가 평가하도록 한 결과, GT 번역물이 학생 번역물보다 일관되게 높은 평가를 받았음에도 불구하고 평가자들이 우수한 번역을 학생의 것으로, 열등한 번역을 GT의 것으로 잘못 귀속시키는 지각적 편향(perceptual bias)을 보였다고 보고하였다. 이는 AI 번역에 대한 평가가 실제 품질과 무관하게 왜곡될 수 있음을 시사한다. 이와 같이 MT 평가 결과가 평가 방법이나 평가자에 따라 일관되지 않게 나타나는 근본적인 이유 중 하나로, 번역 품질 평가 자체의 개념과 측정 방법이 아직 표준화되어 있지 않다는 점을 들 수 있다. Castilho 등(2018)은 인간 및 기계번역 품질 평가에 대한 기존 접근법들을 개관하며, 여러 학문 분야에 걸쳐 다양한 평가 방식이 혼재해 있음에도 근본적이고 광범위한 쟁점들이 여전히 해결되지 않은 채 남아 있다고 지적하였다. 결국 학습자의 주관적 인식, 자동평가지표, 독해 수행 등 서로 다른 층위의 지표들이 상호 검증되지 않은 채 개별적으로 사용되어 온 것이 MT 평가 연구의 공통된 한계라 할 수 있다.

요약하면 선행연구들은 읽기보다 쓰기 중심, 방법론적으로 자기 보고에 의존, 대체로 단일 텍스트 유형에 국한되었다는 점에서 제한적이다. 이러한 평가 방법의 다층적 한계에 더하여, 기계번역의 산출 품질 자체도 원문의 언어(Shadiev et al., 2019), 텍스트 유형(van der Wees et al., 2018) 등에 의해 결정적인 영향을 받는 것으로 알려져 있다. 특히 문학 텍스트는 함축적 의미, 문체적 효과, 상호텍스트성에 크게 의존한다는 점에서 MT가 다루기 가장

까다로운 도메인으로 통념상 인식되어 왔다. 그러나 이러한 통념을 직접 검증한 연구들은 대체로 문학 장르 내부의 비교에 국한되어 있어, 문학과 비문학 텍스트 간 상대적 난이도 차이를 명시적으로 규명하지는 못하고 있다. 예를 들어 Cespedosa와 Mitkov(2023)는 소설과 시라는 두 문학 장르에 대한 MT 성능을 BLEU로 비교하였고, Toral과 Way(2018)는 소설이라는 단일 장르 내에서 신경망 기반과 통계 기반 번역 시스템의 성능을 비교하여 NMT가 유의미하게 더 우수함을 확인하였다. 이 연구는 최신 NMT 기술이 문학 번역에서도 상당한 개선을 이루었음을 보여주지만, 동시에 두 연구 모두 문학과 논설문, 설명문, ESP 등 비문학 장르 간 직접적 비교는 다루지 않았으며, 평가 방식 또한 BLEU나 일부 인간 평가에 국한되어 있어 학습자의 주관적 인식과 훈련된 평가자의 평가를 함께 고려한 것은 아니다.

즉 문학 텍스트가 다른 장르에 비해 실제로 더 어려운지, 그리고 이러한 난이도 차이가 학습자의 주관적 인식과 훈련된 평가자의 평가 수준에서도 일관되게 나타나는지를 검증한 연구는 여전히 부재하다. 이에 본 연구는 다원적(triangulated) 설계를 통해 텍스트 유형별(논설문, 문학, 설명문, ESP) 학습자 평가 데이터와 훈련된 평가자의 산출물 평가 결과를 비교하고, MT 활용 읽기에 대한 학습자의 관점이 과업 수행 전후로 어떻게 달라지는지 함께 분석하고자 한다.

연구 방법

연구 대상

본 연구는 다양한 유형의 영어 텍스트에 대해 신경망 기반 기계번역 알고리즘이 생성한 한국어 번역 산출물의 적합성(adequacy)을 다각도로 검증하기 위해 설계되었다. 이를 위해 훈련된 이중언어 평가자 1인과 학습자 사용자 집단을 평가자로 구성하여 산출물의 질을 평가하였다. 주 채점자는 외국어고등학교를 졸업한 영어교육 전공자로서 한국어와 영어에 능통한 이중언어 사용자였으며 영어 능력 수준은 유럽공통언어 표준 등급에 따라 C1 수준이었다. 이 평가자는 사전 모의 채점 및 채점기준 조정 세션을 거쳐 채점 기준에 대한 이해를 확립한 후 본 채점을 수행하였다(자료 수집 절차 참고).

한편, 대학생 학습자 집단은 한국어를 모어로 하는 21명의 학생으로 구성되었는데, 이들은 모두 영어교육을 주전공 또는 다중전공으로 하는 학생들이었다. 한 명의 학생(3학년)을 제외하고 모두 1학년생들이었으며 57%(n = 12)가 여학생이었고 9명이 남학생이었다. 이들 중 57%(n = 12)의 학생들은 영어 독해력 자기 평가에서 읽기 능력을 우수하다고 평가한 반면 9.5%(n = 2)는 부족하다고 평가하였다. 한편, 33%의 학생들은 영어 읽기 능력을 보통 수준으로 평가하고 있었다.

연구 도구

본 연구에 사용된 연구 도구는 크게 영어 텍스트, 번역 산출물 평가를 위한 워크시트, 평가지와 기계번역 활용 읽기에 대한 관점을 조사하기 위한 설문검사지이다. 우선, 영어 읽기 텍스트는 영어 원서 데이터베이스에서 가독성 지수와 학년 수준(Flesch-Kincaid Grade Level, F-K GL)을 고려하여 다음 4가지 유형의 글로 구성되었다: 특수목적 영어(ESP), 논설문(opinion), 설명문(information), 문학(literature). 첫째, 특수목적 영어(ESP) 텍스트는 전공서의 행동주의 이론 부분에서 추출되었다. 가독성 검증 결과 Flesch Reading Ease 점수는 34.1로 어려운 수준이었으며, Flesch-Kincaid Grade Level은 13.7로 대학생 수준(college level)의 가장 높은 학술적 난이도를 나타냈다. 논설문 텍스트는 공감(empathy)을 주제로 다룬 글로 가독성 점수는 58.9로 다소 어려운(fairly difficult) 수준이었으며 F-K GL은 9학년 수준으로 평정되었다. 셋째, 정보성 텍스트인 설명문은 광고를 주제로 한 글로 가독성 점수는 61.7로 보통 수준이었으며 F-K GL은 8학년 수준으로 확인되었다. 마지막으로 문학(literature)은 Cather(1915)의 소설, *The Song of the Lark* 중 일부를 발췌한 것으로 Flesch Reading Ease 점수는 71.9로 비교적 읽기 쉬운 수준(fairly easy)이었으며, F-K GL은 8학년 수준이었다. 요약하면 대학교 전공 서적에서 발췌한 ESP 텍스트를 제외한 다른 텍스트들은 8~9학년 수준이었으며 문학(소설) 텍스트의 난도가 상대적으로 낮았다. 다만 각 텍스트 유형이 지문 한 개로 구성되어 가독성 수준에도 상당한 편차가 존재하므로, 이후 확인되는 텍스트 유형 간 차이는 유형 자체의

속성 뿐 아니라 개별 지문의 특성이 부분적으로 반영된 결과일 수 있음에 유의할 필요가 있다.

네 가지 유형의 텍스트를 선정한 후에는 각각의 텍스트를 번역했는데 이 때 맥락 손실을 최소화하기 위해 전체 지문을 한 번에 입력하였다. 번역 도구로는 선행연구(Groves & Mundt, 2015; Scarton & Specia, 2016; Stapleton & Leung, 2019; Tsai, 2019)에서 가장 폭넓게 검증되어 온 구글 번역을 사용하였다. 구글 번역 산출물을 문장 단위로 분할하여 원문과 번역 산출물을 병렬 방식으로 제시하는 방식으로 번역 적합성 평가 과업을 위한 워크시트를 만들고 원문의 의미에 얼마나 가까운지 정도를 1~5점(원문 의미를 가장 정확하고 적합하게 표현) 척도로 평가하게 하였다.

또한, 기계번역 활용 읽기에 대한 학습자 반응을 조사하기 위해 설문검사지를 구성했는데 학생들의 평소 기계번역 활용 경험, MT 도구 사용 현황, 읽기 능력에 대한 자기평가 등을 묻는 질문들 외에도 읽기 번역 산출물의 적합성 평가 전과 후 MT의 유용성에 대한 인식을 조사하기 위한 5점(1점: 전혀 그렇지 않다 ~ 5점: 매우 그렇다) 리커트 척도형 문항이 포함되어 있었다. 그 밖에도 MT 프로그램을 사용하는 이유, 도움되는 요인 등을 묻는 선택형 문항과 MT 사용 후 언어적, 정의적 변화 등을 기술하는 개방형 문항들이 포함되었다.

자료 수집

본 연구의 자료수집은 영어교육과에서 영어문법 수업을 수강하는 영어과 예비교사들을 대상으로 이루어졌다. 총 4개의 텍스트에 대한 번역 적합성 평가는 4주에 걸쳐 진행되었으며 매주 수업 시간에 30분 정도 할애하여 학생들이 다른 유형의 텍스트를 읽고 문장 단위로 적합성을 평가하였다. 아울러 학생들은 문장 단위 평가를 마친 후, 해당 회차에서 다룬 텍스트 유형의 번역 산출물에 대한 소감을 자유롭게 서술하도록 안내 받았다. 학생들은 번역 산출물 평가 및 소감 작성 후 제공된 온라인 설문 링크를 방문하여 설문에 응답하였다. 연구자는 자료 수집에 앞서 연구의 목적을 설명하고 익명성과 보안 유지에 대해 안내하였다. 또한 연구 참여를 희망하지 않는 학생들에게는 대체 활동을 제공하였다. 이를 통해 학생들의 독해능력 자기평가 점수, MT 도구(예: ChatGPT, 구글 번역, 파파고 등) 선호도, MT의 신뢰도 및 유용성에 대한 인식, 텍스트 유형별 번역 적합성 평가 결과 등을 자료로 수집하였다.

이와 함께 훈련된 평가자가 번역 산출물을 평가하였다. 본채점에 앞서 채점 기준의 신뢰도와 타당성을 확보하기 위해, 모어와 목표어(영어)에 모두 능통한 이중언어 사용자 3명(이 중 1인은 이후 본채점을 담당할 주 채점자)과 연구자를 포함한 총 4인의 평가자가 47개 문장의 번역 산출물에 대한 모의 채점(pilot rating)을 독립적으로 실시하였다. 문장들을 1–5점 척도로 채점한 결과, 평가자 간 신뢰도는 급내 상관계수(Intraclass Correlation Coefficient) ICC(2,4) = .73(95% CI [.67, .87])으로 양호한 수준이었다. 이후 연구자와 주 채점자는 평가자 간 불일치가 두드러진 문항을 중심으로 2회 이상의 채점기준 조정 회의(norming session)를 진행하여 채점 기준에 대한 해석 차이를 논의하고 조정함으로써 평가자 간 판단 기준을 일치시켰다. 이러한 과정을 통해 채점 기준에 대한 공통된 이해가 확립되었다고 판단하여 이후 본채점은 주 채점자가 수행하였다.

결과 분석

본 연구에서 수집된 자료의 분석을 위해 통계 프로그램 SPSS 26을 사용하였다. 연구 질문별로 구체적인 분석 절차는 다음과 같다. 우선, 한국인 대학생 학습자(n = 21)가 지각한 텍스트 유형별(ESP, opinion, information, literature) MT 산출물의 적합성 점수에 차이가 있는지 검증하기 위해, 텍스트 유형(4수준)을 요인으로 하는 반복측정 일원분산분석(one-way repeated-measures ANOVA)을 실시하였다. Mauchly의 구형성(sphericity) 검정을 통해 사전 가정을 확인하였으며, 가정이 충족되지 않아 Greenhouse-Geisser 교정값을 적용하여 보고하였다. 주효과가 유의미하게 나타난 경우, 텍스트 유형 간 구체적인 차이를 식별하기 위해 Bonferroni 교정을 적용한 사후 대응비교(pairwise comparison)를 실시하였다.

이와 함께 읽기 텍스트에 대한 번역 산출물 평가 과업과 관련하여 실시한 설문조사 데이터를 분석하였다. 설문조사는 텍스트 유형별 평가 과업을 수행하기 전 학습자가 평소 지니고 있던 MT의 신뢰도에 대한 전반적인 인식을 조사한 섹션과, 텍스트 유형별 평가 과업을 수행한 후 MT의 신뢰도에 대한 의견을 다시 물은 섹션으로 구성되었으며, 분석은 이 두 섹션의 응답값 비교에 초점을 두었다. 이 두 시점(과업 전후)에 따라 MT의 신뢰도에 대한 학습자 인식에 변화가 있는지 검증하기 위해 시점을 요인으로 하는 반복측정 일원분산분석을 실시하였으며, 설문조사의 각 문항에 대해서는 기술통계 분석을 함께 실시하였다.

훈련된 평가자 데이터의 경우, 문장 단위 번역 산출물에 대한 평가값이 텍스트 유형별로 차이를 보이는지 검증하기 위해 비모수 통계 검정인 Kruskal-Wallis 검정을 실시하였다. 텍스트 유형 간 문장 수가 상이하고 일대일 대응이 불가능하여, 대응표본 검정 대신 전체 문장을 집단으로 간주하는 Kruskal-Wallis 검정을 사용하였다. 이때 효과 크기(epsilon-squared, ε^2)와 함께 두 집단 간 순위 일치도, Figure 3의 분포 패턴을 함께 검토하여 결과를 다각도로 해석하였다. 주효과가 유의미하게 나타난 경우, 텍스트 유형 간 구체적인 차이를 식별하기 위해 Dunn(1964)의 방법에 기반한 쌍별비교(pairwise comparison)에 Bonferroni 교정을 적용하여 사후 검정을 실시하였다.

아울러 학생들이 매 회차 작성한 텍스트 유형별 소감문(n = 82; 결측 2건 제외)에 대해서는 연구자가 질적 내용 분석을 위해 개방 코딩을 실시하였다. 82개의 소감문을 반복적으로 정독한 결과 21개의 개방 코드를 도출했는데 이들 중 관련성 있는 것들은 연결하여 6개 상위 범주(어휘, 구문, 담화, 긍정적 평가, 메타인지, 장르 인식)로 통합하였다.

연구 결과

텍스트 유형별 번역 적합성에 대한 학습자 평가

한국인 대학생 학습자(n = 21) 관점에서 텍스트 유형별 MT 산출물의 적합성과 산출물 평가 과업 수행 전후 MT의 유용성 및 신뢰도에 변화가 있는지 검증하기 위해 반복측정 일원분산분석(repeated-measures ANOVA)을 실시하였다. 텍스트 유형별 번역의 적합성에 대한 학습자 평가에서 점수는 ESP(M = 4.41, SD = .40), opinion(M = 4.29, SD = .40), information(M = 4.27, SD = .35), literature(M = 4.06, SD = .47) 순으로 나타났다(Table 1 참조). 텍스트 유형별로 평가 점수에 차이가 있는지 검증하기 위해 Mauchly의 구형성 검정 결과를 확인했는데 구형성 가정이 기각되어(W = .51, p = .026), Greenhouse-Geisser 교정치(ε = .69)를 적용하였다.

분석 결과, 텍스트 유형에 따른 번역 적절성 인식은 통계적으로 유의미한 차이가 있었다[F(2.06, 41.20) = 7.12, p = .002, ηp^2 = .26]. 사후 다중비교를 통해 텍스트 유형 간 차이를 비교한 결과 문학과 다른 유형의 텍스트(ESP, opinion, information) 간 차이가 모든 쌍에서 통계적으로 유의한 것으로 나타났다. 흥미롭게도 원문 자체의 가독성이 상대적으로 낮았던(즉 난도가 높았던) ESP와 논설문 뿐 아니라, 가독성이 보통 수준이었던 설명문에서 가독성이 가장 높았던 문학 텍스트보다 적합성 평가 점수가 통계적으로 유의미하게 높게 나타났다.

TABLE 1
Descriptive Statistics for Different Text Types: Student Rating

Text Types	Mean (M)	Standard Deviation (SD)
ESP	4.41	0.40
Opinion	4.29	0.40
Information	4.27	0.35
Literature	4.06	0.47

학생들이 글감 유형별로 번역 산출물의 적합성을 평가한 결과를 도식화하면 Figure 1과 같다. 그래프 y축의 수치는 5점 평가 척도의 점수값을 의미하며, 하단의 1, 2, 3, 4는 각각 전공 텍스트(ESP), 논설문(opinion), 설명문(information), 문학(literature)을 의미하며 텍스트 유형별로 표시된 수직 구간은 점수 값의 범위를 보여준다.

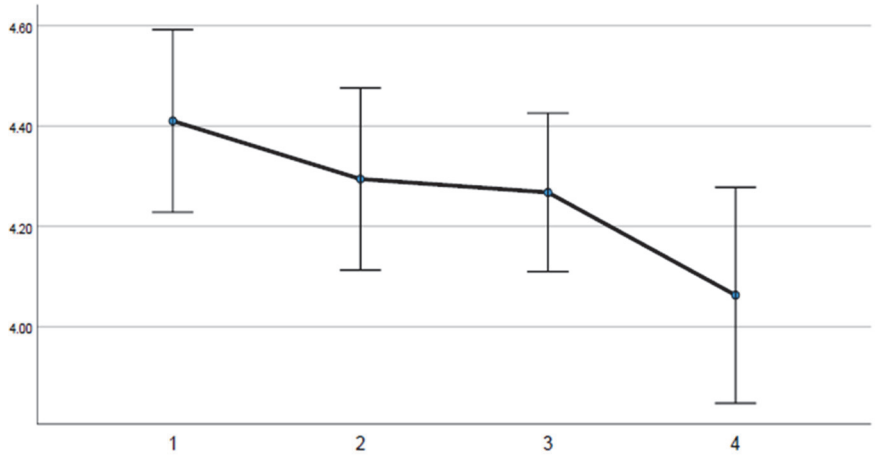
이러한 정량적 평가 결과를 뒷받침하는 질적 근거를 확인하기 위해, 학생들이 매 회차 작성한 텍스트 유형별 소감문(n = 82)을 개방 코딩했는데, 그 결과 6개 상위 범주(어휘, 구문, 담화, 긍정적 평가, 메타인지, 장르 인식)로 통합된 21개의 개방 코드가 도출되었다(범주별 대표 발췌문은 Appendix 참고). 범주별 출현 빈도를 텍스트 유형에 따라 비교한 결과(Table 3 참조), 문학 텍스트에서 구문 관련 코드(n = 25)가 다른 세 유형(각 9~16건)보다 뚜렷하게 많이 보고된 반면, 긍정적 평가(n = 3)는 가장 적어 Table 2의 정량적 결과와 일관된 패턴을 보였다.

TABLE 2

Post-hoc Comparisons Between Different Text Types

(P) Text	(Q) Text	Mean Difference (P-Q)	Standard Error	Significance	Lower Bound	Upper Bound
1	2	0.12	0.09	1.00	-0.14	0.37
	3	0.14	0.09	0.88	-0.13	0.42
	4	0.35*	0.10	0.01	0.07	0.63
2	1	-0.12	0.09	1.00	-0.37	0.14
	3	0.03	0.05	1.00	-0.12	0.18
	4	0.23*	0.05	0.00	0.08	0.39
3	1	-0.14	0.09	0.88	-0.42	0.13
	2	-0.03	0.05	1.00	-0.18	0.12
	4	0.21*	0.06	0.02	0.02	0.39
4	1	-0.35*	0.10	0.01	-0.63	-0.07
	2	-0.23*	0.05	0.00	-0.39	-0.08
	3	-0.21*	0.06	0.02	-0.39	-0.02

Notes. 1: ESP, 2: Opinion, 3: Information, 4: Literature; Significance: *p* adjusted by *Bonferroni; *: p < .05*



Notes. 1: ESP, 2: Opinion, 3: Information, 4: Literature

FIGURE 1

Learner Evaluation of Machine Translation Adequacy Across Different Text Types

TABLE 3

Frequency of Open-coded Categories in Students' Reflections Across Text Types

Category	Information	Opinion	Literature	ESP	Total
Lexical level	7	11	3	5	26
Syntactic level	10	16	25	9	60
Discourse level	7	6	14	12	39
Positive evaluation	14	15	3	16	48
Metacognition	14	4	6	3	27
Text type awareness	1	0	1	2	4

특히 담화 범주에 속하는 고유명사, 호칭, 존비어 처리 오류는 문학 텍스트에서 4회 보고된 반면, 설명문(0회), 논설문(0회), ESP(1회) 등에서는 문제로 많이 언급되지 않았다. 한 학생(참여자 7)은 "여자-남자 형제의 호칭을 '형'이라고 번역한 것을 보고 번역기의 문화적 맥락을 번역할 수 있는 능력이 탁월하지 않다는 것을 느꼈다"고 기술하였다. 반면 ESP와 논설문에서는 긍정적 평가가 가장 높게 나타났으며(각 n = 16, 15), 학생들은 "학술적인 글이었고 대체적으로 번역의 완성도가 좋았다"(참여자 3), "학술 관련 내용이라 정확도가 높았다"(참여자 10)와 같이 정형화된 문체가 기계번역에 유리하다는 점을 스스로 인지하고 있었다.

MT 활용 읽기에 대한 학습자 인식

학생들은 번역 산출물을 평가한 후에 설문조사를 통해 과업 수행 전후에 기계번역의 정확성 및 신뢰도에 변화가 있는지 응답했다. 구체적으로 기계번역 프로그램이 영문을 국어로 정확하게 해석한다, 기계번역의 오류가 많다(역코딩), 영어 읽기를 할 때 기계번역 프로그램을 신뢰하지 않는다(역코딩) 등의 항목에 대한 응답값의 차이를 반복 측정 일원분산분석을 통해 분석하였다.

TABLE 4
Learner Perception: Trustworthiness of Machine Translation

Time	Mean (*M*)	Standard Deviation(*SD*)
Pre-view	2.84	0.74
Post-view	3.16	0.74

Table 4에 제시된 바와 같이 과업 수행 후 학습자 관점에서 MT의 신뢰도 값이 근소하게 상승했지만 통계적으로 유의한 차이는 아니었다[$F(1, 20) = 1.77$, $p = .198$, $\eta p^2 = .08$]. 이를 도식화하면 Figure 2와 같다. 그림에서와 같이 Time 1과 비교했을 때 Time 2에 MT의 신뢰도 값이 0.3점 올라갔지만 그 차이가 매우 근소하여 통계적으로 유의하지는 않았다.

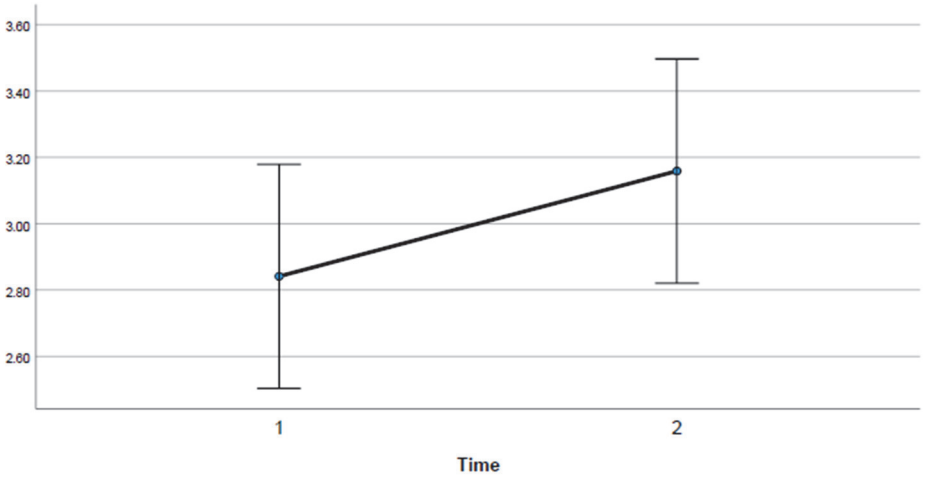


FIGURE 2
Learner Perception of MT Trustworthiness before (Time 1) and after (Time 2) MT Adequacy Evaluation

실제로 기계번역의 정확성을 평가하는 문항에 대해 21명의 학생 중 12명(57%)만 '그렇다'는 긍정 응답을 했고 평균값은 3.57로 나타났다. 한편, 영어 읽기 상황에서 활용되는 주요 번역 프로그램(ChatGPT, 파파고, 구글 번역)의 정확성에 대한 학습자들의 평가(1, 2, 3 순위 배정) 결과를 분석한 결과 도구 간 순위가 뚜렷한 것이 확인되었다. 우선, 전체 응답자 중 13명(62%)이 ChatGPT를 가장 정확한 번역 도구(1위)로 선택하여 평균 순위 1.62를 기록하여 독보적인 신뢰도를 확보하고 있는 것으로 나타났다. 한편, 파파고는 7명(33.3%)의 선택을 받아 중위 서열(2위)을 형성하였다. 구글 번역의 경우에는 주로 2위(52.4%)나 3위(42.9%)로 밀려나 평균 순위 2.38로 가장 낮은 정확성

평가를 받았다. 이를 통해 구글 번역이 웹 접근성이 좋아도 학습자의 정확도 평가가 낮았던데 반해, 대형언어모델(LLM) 기반의 번역 산출물에 대한 평가가 높았음을 알 수 있다.

그러나 학생들이 영어 읽기를 할 때 자주 사용하는 번역 프로그램은 MT 프로그램의 정확성 평가 결과와는 차이가 있었다. 분석 결과, 학습자들의 선호도는 파파고(n = 12, 57.1%)와 ChatGPT(n = 9, 42.9%)로 크게 양분되는 양상을 보였다. 파파고를 선호하는 이유와 관련해서는 모국어(L1)에 최적화된 번역이라는 기대와 어휘 학습의 편의성 때문인 것으로 나타났다. 학습자들은 파파고가 국내 기업에서 개발한 서비스라는 점에서 해외에서 개발된 번역 프로그램에 비해 한국어 표현이 가장 자연스러울 것이라는 높은 신뢰를 갖고 있었다. 또한 번역문 산출 시 단어 뜻을 동시에 제공하는 인터페이스의 편의성과 용이한 접근성을 주요 선호 이유로 기술하였다. 반면 ChatGPT를 선호하는 학습자들은 담화 수준의 맥락 파악 능력과 대화 기반의 상호작용성을 핵심 요인으로 지목하였다. 학습자들은 GPT가 글의 전체 맥락과 단어 고유의 어감을 고도로 파악하여 훨씬 매끄러운 텍스트를 도출한다고 인식하고 있었다. 특히 전통적인 번역기와 차별화되는 ChatGPT만의 독창적 유용성으로 "앞선 내용을 기억하여 오타를 능동적으로 수정하는 기능"과 "이해가 되지 않는 부분을 재질문하여 번역 오차를 줄여가는 상호작용 역량"을 언급하였다.

한편, 학습자들이 실제 영어 학습 생태계에서 MT를 사용하며 직면했던 한계와 오류 경험을 묻는 개방형 문항에 대한 응답을 범주화한 결과, 문법 오류 및 번역의 질 저하(n = 7, 33%)가 가장 빈번하게 언급되었다. 학습자들은 어순이 불완전하거나 문장이 명사형으로 종결되는 현상, 반말과 존댓말의 상황적 맥락을 구분하지 못하는 존비어 처리의 한계, 그리고 특정 문장을 통째로 건너뛰는 누락 현상 등을 MT 산출물과 관련된 황당한 경험으로 지목하였다. 특히, 한 학습자는 MT가 문학 텍스트 번역에서 심각한 기능적 열세를 보인다고 보고하여 앞서 검증된 텍스트 유형별 산출물의 편차와 맥을 같이 하였다.

두 번째로 높은 빈도를 차지한 항목은 맥락 미반영 및 다의어 오역 등(n = 5, 24%)이었다. 학습자들은 텍스트의 중심 주제를 인지하지 못한 채 단어를 평면적으로 치환하는 번역 엔진의 전형적인 한계를 경험했는데 구체적인 예로 글의 의미로 사용된 지문을 'fingerprint(지문, 指紋)'로 오역하거나, 여성 권리 신장의 의미인 '여권(女權)'을 'passport(여권, 旅券)'로 오역하는 등 동음이의어 처리에 취약함을 문제로 인식하고 있었다. 마지막으로 고유명사, 호칭 및 관용 표현의 자의적 직역(19.0%, 4명)도 한계로 지적되었는데 구체적으로 고유한 인명을 일반 명사로 해석하여 문장을 왜곡하거나, 한국어 고유 관용구인 "누워서 떡 먹기"를 문화적 등가 어휘가 아닌 말 그대로 직역("eating rice cake while lying down")하여 의미 전달을 완전히 상실한 사례들이 보고되었다. 아울러 본 연구의 독해 과업 과정에서 산출된 "Mr. Kronborg"의 "씨. Kronborg" 번역 투와 같은 호칭 결함을 예리하게 식별해 낸 학생도 있었다.

기계번역을 사용하는 이유와 관련해서는 학생들은 여러 개의 선택지 중 영어 실력이 부족해서(n = 9, 43%), 시간 절약을 위해(n = 9, 43%)를 가장 많이 선택했으며, 학교 과제 성적을 잘 받기 위해(n = 6, 29%)를 선택한 빈도도 높았다. 한편, '과제와 관계 없이 혼자 공부할 때 도구로 사용한다'를 선택한 학생들도 적지 않았다(n = 6, 29%). 기계번역의 어떤 면이 도움이 되느냐는 질문과 관련해서는 문법의 정확성 향상(33%), 어휘력 신장(24%), 원어민 같은 언어 산출(24%), 자신감 증가(14%), 영어교과 성적 향상(14%), 과제 시간 절약(14%) 등을 선택하였다.

한편, 기계번역 활용 읽기 과업을 수행하는 과정에서 한국인 대학생 학습자들이 체감한 언어적, 감정적 차원의 변화를 다각도로 분석했는데 그 결과, 학습자들은 MT를 단순한 편리 도구로 수용하는 것을 넘어, 긍정적인 반응과 부정적인 반응이 팽팽하게 교차하는 복합적인 심리적 역동을 경험하는 것으로 확인되었다. 긍정적인 반응을 보인 학생들은(n = 8, 38%) MT가 영어 독해 불안감을 낮춰주는 인지적 스캐폴딩(scaffolding)을 하기에 난해한 구문을 마주했을 때 정서적 안도감이 생기며, 번역기가 제공하는 풍부한 언어 입력을 통해 자연스러운 표현을 습득하여 자신감이 상승한다고 밝혔다. 이와 대조적으로 정서적 불안 및 회의감(n = 7, 33%)을 표현한 학생들은 MT의 산출물을 무조건 신뢰하지 못하고 오역 가능성에 대한 끊임없는 불안을 표출하며, 타 플랫폼을 통해 2차, 3차로 교차 재검증을 수행하는 인지적 부담을 겪고 있었다. 흥미롭게도 MT에 의존하는 과정에서 학습자 본인의 언어 숙달도가 퇴화, 정체될 수 있다는 우려와 성찰이 "허무함," "의존에 대한 찝찝함" 등의 부정적 감정 어휘로 표현되었다는 점이다. 이를 통해 대학생 학습자들이 기계번역을 사용하며 편리함과 시간 단축이라는 이익을 누리지만 동시에 MT 의존에 대한 성찰적 불안과 신뢰성에 대한 인지적 회의감을 느끼고 있음을 알 수 있다.

텍스트 유형별 번역 적합성에 대한 훈련된 평가자 평가

한편, 텍스트 유형별로 번역 적합성을 평가한 훈련된 평가자의 평정 자료(n = 173) 전체를 대상으로 비모수 검정인

Kruskal-Wallis 검정을 실시하였다. 분석 결과, 텍스트 유형에 따른 평가자의 점수값은 통계적으로 유의미한 차이가 있는 것으로 나타났다[$H(3) = 22.88$, $p < .001$, $\varepsilon^2 = .13$]. Bonferroni 교정을 적용한 사후 비교 결과, 평균 순위(mean rank)가 가장 낮았던 문학(literature, 63.03)은 ESP(100.95, $p = .001$) 및 논설문(opinion, 101.37, $p < .001$)과는 통계적으로 유의한 수준에서 낮은 것으로 나타났으나, 설명문(information, 78.67)과는 통계적으로 유의미한 차이를 보이지 않았다($p = .586$). 한편 ESP와 논설문 간에는 유의미한 차이가 없었으며, 설명문과 논설문 간 차이는 경계 수준에 머물렀다($p = .050$). 이를 통해 기계번역 산출물의 질이 텍스트 유형에 따라 차이가 있으며 그 양상이 단순히 문학 대 비문학의 이분법이 아니라 텍스트 유형 간 상대적인 스펙트럼을 보이고 있음을 알 수 있었다.

기술통계 결과를 구체적으로 살펴보면 Table 5와 같다. 논설문(opinion)이 평균 4.63점($SD = .87$)으로 가장 높았고, 전공 텍스트(ESP)가 평균 4.55점($SD = 1.08$)으로 그 뒤를 이었다. 두 유형 모두 중앙값(median)이 5점 만점을 기록하였으며, 왜도(skewness)가 각각 -2.58과 -2.50, 첨도(kurtosis)가 각각 6.62와 5.39로 나타나 뾰족하고 왼쪽으로 긴 꼬리를 가진 분포를 보였다. 이는 대다수의 문장이 만점 근처에 밀집한 가운데, 소수의 문장만이 낮은 점수를 받아 이례적인 이상치를 형성하고 있음을 시사한다(Figure 3 참조). 설명문(Information)은 평균 4.19점($SD = .98$, 왜도 = -.82, 첨도 = -.61)으로 그 뒤를 이었으며, 문학 텍스트는 평균 3.64점($SD = 1.40$, 왜도 = -.64, 첨도 = -.85)이라는 가장 낮은 점수를 받았다. 설명문과 문학 텍스트는 왜도와 첨도 모두 완만한 수준을 보여, 앞의 두 유형과 달리 점수가 비교적 고르게 분포되어 있음을 알 수 있었다.

Figure 3은 텍스트 유형별 평가자 평정 점수의 분포를 응답 비율로 나타낸 것이다. 그림에서 확인할 수 있듯이, ESP와 논설문은 5점 응답 비율이 각각 82%와 80%에 달해 응답의 대부분이 최고점에 집중되어 있었으며, 4점 이하의 응답은 20% 미만에 그쳤다. 이는 앞서 보고한 강한 음의 왜도 및 큰 양의 첨도와 일치하는 양상으로, 두 텍스트 유형에서 훈련된 평가자가 대다수의 문장을 만점에 가깝게 평가하되 소수의 문장에서만 낮은 점수를 부여했음을 시각적으로 보여준다. 반면 설명문과 문학은 5점 응답 비율이 각각 52%와 39%로 상대적으로 낮았고, 3~4점 구간의 응답이 고르게 분포되어 있었다. 특히 문학은 1점과 2점 응답이 각각 11%에 달해 네 유형 중 낮은 점수의 비중이 가장 컸으며, 이는 문학 텍스트가 앞서 확인된 바와 같이 가장 낮은 평균($M = 3.64$)과 가장 큰 표준편차($SD = 1.40$)를 보인 것과 일치한다. 이러한 분포 패턴은 텍스트 유형에 따른 번역 적합성의 차이가 단순히 평균값의 차이에 그치지 않고, 응답의 분산과 극단값의 존재 양상에서도 뚜렷하게 나타남을 보여준다.

TABLE 5

Descriptive Statistics for Different Text Types: A Comparison of Trained Rater (TR) and Student Ratings

Text Type	TR *N*	TR *M*	TR *SD*	TR Median	Skewness (*SE*)	Kurtosis (*SE*)	TR Rank	S *M*	S *SD*	S Rank
E	38	4.55	1.08	5.00	-2.50 (.38)	5.39 (.75)	2	4.41	0.40	1
O	51	4.63	0.87	5.00	-2.58 (.33)	6.62 (.66)	1	4.29	0.40	2
I	48	4.19	0.98	5.00	-0.82 (.34)	-0.61 (.67)	3	4.27	0.35	3
L	36	3.64	1.40	4.00	-0.64 (.39)	-0.85 (.77)	4	4.06	0.47	4

Note. TR (Trained rater) M: sentence-level analysis based on the full dataset ($n = 173$); *S (Student) M*: individual learner averages ($n = 21$); Rank: descending order of mean rating within each group (1 = highest, 4 = lowest); E: ESP, O: Opinion, I: Information, L: Literature

흥미롭게도 번역 적합성에 대한 훈련된 평가자와 학생의 평가 패턴은 유사한 것으로 나타났다. 훈련된 평가자 평정과 학습자 평정은 분석 단위(문장 단위 대 참여자 단위)와 척도 분포가 상이하므로, 원점수 평균값을 직접 비교하는 것은 통계적으로 적절하지 않다. 이에 원점수 대신 순위(rank)를 제시하여 두 집단 간 상대적인 평가 경향을 비교할 수 있도록 하였다. 훈련된 평가자와 학습자 집단의 순위를 비교한 결과 두 집단은 매우 유사한 평가

경향을 보였다. 훈련된 평가자는 논설문(1위)과 ESP(2위)를 근소한 차이로 가장 높게 평가한 반면, 학습자는 ESP(1위)와 논설문(2위)의 순서만 뒤바뀌었을 뿐 두 유형 모두 상위권으로 평가하였다. 특히 설명문(3위)과 문학(4위, 최하위)에 대한 평가는 훈련된 평가자와 학습자 집단에서 완전히 일치하여, 두 집단 모두 문학 텍스트의 번역 적합성을 가장 낮게 인식하고 있음을 확인할 수 있었다.

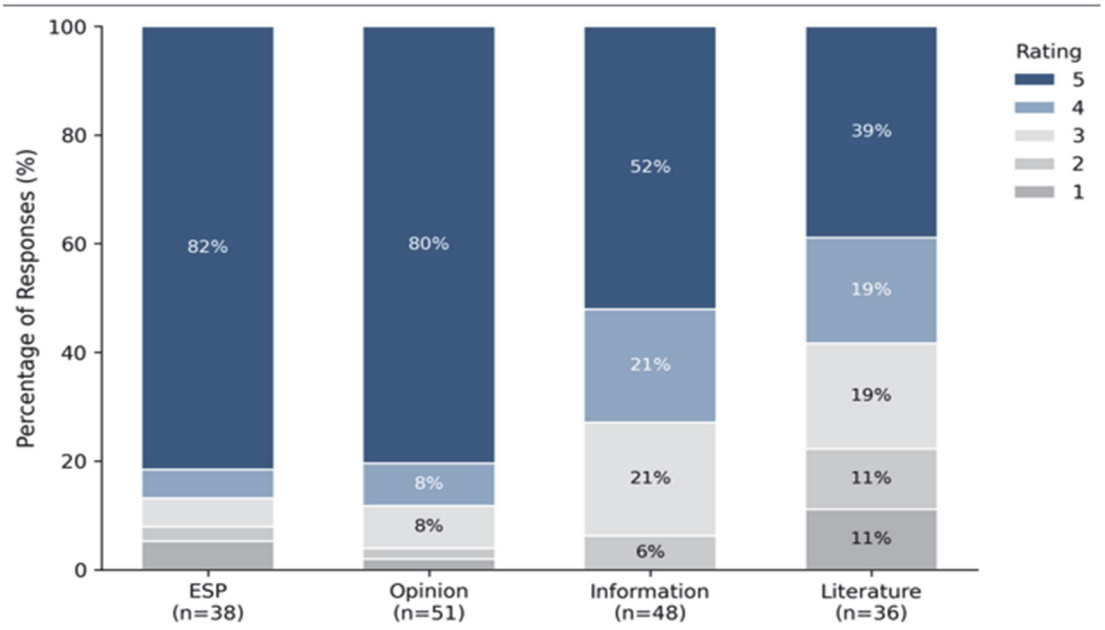


FIGURE 3
Distribution of Trained Rater Evaluation of Translation Adequacy by Text Type (n = 173)

논의

본 연구는 MT 산출물의 적합성에 대한 한국인 EFL 대학생 집단과 훈련된 평가자의 평가 결과를 통해 텍스트 유형별로 적합성 점수에 차이가 있는지 실증적으로 분석하고자 하였다. 이를 통해 텍스트 유형별 기계번역의 적합성과 관련하여 훈련된 평가자와 학습자 간 평가 패턴이 유사한 것으로 확인되었다. Kruskal-Wallis 검정과 반복측정 분산분석 결과를 통해 두 그룹 모두 ESP와 논설문(opinion)의 번역 적합성을 높게 평가한 반면 문학(literature)은 가장 낮게 평가한 것을 알 수 있었다. 이는 수사적 구조가 정형화되고 정보제공을 위한 텍스트에서 최신 LLM 및 NMT 알고리즘이 완벽에 가까운 완성도를 보인다는 선행연구 결과와 맥을 같이 한다(Aeni et al., 2024; Y. B. Yoon, 2023). 훈련된 평가자는 이러한 통사적 정확성을 중심으로 평가하여 만점에 가까운 점수를 부여한 것으로 보인다. 또한 예비교사들로 구성된 학습자 집단 역시 텍스트 유형별 기계번역의 상대적 한계를 식별할 수 있는 기본적인 인지적 역량을 갖추고 있음을 시사한다.

그러나 두 집단 간 평가 경향성이 거시적 측면에서는 유사해보이지만 미시적 측면에서는 인식에 격차가 존재한다는 점을 주목할 필요가 있다. 훈련된 평가자 평가에서 문학 텍스트는 평균 순위 최하위(63.03)와 낮은 점수(M = 3.64)를 기록함과 동시에, 문장별 번역 산출물 평가에서 큰 편차(SD = 1.40)를 보였다. 반면 학습자들은 동일 텍스트에 4.06점이라는 관대한 평정을 부여했는데, 이는 문학 영역의 원문이 Flesch Reading Ease 71.9로 네 가지 텍스트 유형 중 가독성이 가장 높다는 사실과 직결된다. 즉, 학습자들은 실제 번역 산출물에 내재된 심층적인 맥락 오류나 비유 왜곡을 예리하게 분별하지 못한 채, 번역 엔진이 출력한 산출물의 표층적 가독성(surface-level readability) 때문에 과잉 관대화(over-tolerance) 경향을 보인 것으로 해석된다. “Mr. Kronborg”를 “씨. Kronborg”와 같이 번역하는 오류를 식별해 낸 일부 학생도 존재했으나, 대다수는 “원문이 쉬우니 번역도 잘 된 것 같다”고 오판했을 가능성이 크다. 이는 학습자가 지각한 가독성이 실제 번역 품질과 반드시 일치하지 않을 수 있다고 제안한

Shih(2022)의 문제의식과 맥을 같이 하며, 학습자의 주관적 판단이 실제 언어적 정확성을 담보하지 못하므로 별도의 검증이 필수적이라는 본 연구의 문제의식을 뒷받침한다. 이러한 현상은 앞서 기술된 개방형 설문 응답 결과와도 연결되는데, 학습자들은 문학 번역의 맥락적 오역(동음이의어 및 관용구 직역)을 한계로 인지하면서도, 독해 불안을 상쇄해 주는 스캐폴딩 도구로 MT를 수용하는 과정에서 인지적 부조화(cognitive dissonance)를 겪고 있음이 확인되었다. 이와 유사한 양상은 텍스트 유형별 소감문 개방 코딩 분석에서도 나타났는데, 담화 범주(고유명사, 호칭, 존비어 처리 오류, 문화적 맥락 이해 부족 등)에 속하는 코드가 문학 텍스트에서 유독 집중적으로 보고되어(Table 3 참조), 학습자들이 사후 성찰 단계에서는 번역의 미시적 오류를 구체적으로 인지할 수 있는 것으로 나타났다. 그럼에도 이러한 인지가 문장 단위 실시간 평가 점수(4.06점)에는 충분히 반영되지 못했던 것은 텍스트의 가독성이 높았기 때문인 것으로 이해된다. 특히, 통계적으로 최고 정확도를 부여한 도구(ChatGPT)와 최고 선호도를 보인 도구(파파고)가 양분된 현상은, 학습자가 테크놀로지의 심층적 역량(맥락 파악 능력)보다 표층적 편의성(L1 최적화 인터페이스)에 끌려 도구를 선택할 수 있는 가능성을 보여준다. 흥미롭게도 본 연구의 참여자들은 구글 번역을 선호하지 않았는데 이와 같은 결과는 참여자들이 DeepL이나 ChatGPT보다 구글 번역을 선호한다고 보고했던 Klimova(2025)의 결과와 차이가 있다.

한편, MT 사용 결과 인지적, 정의적 변화에 대한 응답을 분석한 결과 기계번역은 단순한 언어적 보조 도구를 넘어 정서적 혜택과 인지적 부담이 교차하는 복합적 매개체임이 확인되었다. 학습자들은 독해 불안 감소와 시간 절약이라는 실용적 이점을 누리면서도, 동시에 MT 산출물에 대한 근본적인 불신으로 인해 타 플랫폼을 통해 2차, 3차 교차 검증을 수행할 필요를 인지하고 있었다. 흥미롭게도 학습자들은 복잡한 텍스트일수록 절대적인 독해 시간을 단축해 주는 기능적 장점을 명확히 인지하면서도 "기계번역을 보며 다시 나만의 해석을 재시도"하는 능동적인 포스트 에디팅 성향의 학습 흐름도 관찰되었다. 그러나 동시에 영어 전공자로서 MT의 고도화가 향후 자신들의 진로에 장애물이 될 것이라는 정서적 위기감과 불안(33%)을 느끼고 있었다. 이와 같은 결과는 학생들이 기계번역의 효율성, 효과성을 인정함과 동시에 문화적, 맥락적 뉘앙스를 표현하지 못하는 한계를 인지하고 있다고 보고한 Klimova(2025)의 결과와 맥을 같이 한다. 요약하면 번역 산출물의 질은 텍스트 유형별로 차이가 있으며 대학생 학습자들은 평소 과제 수행을 위해 MT 도구를 많이 활용하면서도, 텍스트 유형별로 번역 산출물의 질을 엄격하게 평가하고 오류를 가려낼 수 있는 분별력은 충분히 갖추지 못한 것으로 나타났다.

결론

본 연구는 혼합연구방법을 통해 MT의 적합성과 신뢰도를 다각도로 검증했으나 몇 가지 한계점이 있다. 우선, 학습자 집단이 영어교육을 전공하는 소수의 학생들(n = 21)로 구성되어 있어, 일반 EFL 학습자에게 결과를 일반화하기 어렵다. 또한, 텍스트 유형별로 지문이 한 편씩만 포함되어 있고 지문 간 가독성 수준에도 편차가 존재하므로, 본 연구에서 확인된 텍스트 유형 간 차이가 유형 자체의 속성에서 비롯된 것인지 개별 지문의 특성에서 비롯된 것인지 완전히 분리하기는 어렵다. 향후 연구에서는 유형별로 다수의 지문을 포함하여 이러한 가능성을 통제할 필요가 있다. 뿐만 아니라, 모의 평가 단계에서 복수의 채점자를 활용했고 채점자 연수(rater training)를 진행했지만 최종적으로 한 명의 훈련된 이중언어 평가자가 평가를 독립적으로 수행했다는 점에서 제한적이다. 아울러 Kruskal-Wallis 검정에서는 문장 173개를 서로 독립적인 관찰치로 간주하였으나, 실제로는 같은 지문에서 추출된 문장들이 같은 저자, 같은 번역기, 같은 평가자의 판단을 공유한다는 점에서 완전한 독립성을 가정하기는 어렵다. 즉 문장 수(173개)에 비해 실제 독립적인 표본의 단위는 지문 4편에 더 가까우며, 이 때문에 본 연구에서 보고된 유의확률이 실제보다 과대 추정되었을 가능성이 있다. 이러한 이유로 본 연구는 통계적 검정 결과를 확정적 증거가 아닌 탐색적 지표로 간주하였으며, 유의확률과 함께 효과 크기(ε^2 = .13, 중간 정도의 효과 크기)와 순위 패턴(Table 5, Figure 3)을 함께 검토하여 해석의 균형을 도모하였다. 또한 표층적 유창성에 근거한 관대한 평가라는 해석은 본 연구 결과에서 나타난 통계적 패턴에 기반한 사후적(post hoc) 설명으로 한계가 있다. 이를 직접 검증하기 위해서는 학습자의 시선 추적(eye-tracking)이나 사고 구술(think-aloud) 등의 추가적인 자료수집과 증거가 필요하다.

마지막으로, 본 연구의 자료 수집 시점 이후 3년여의 시간이 경과하였으며, 이는 학술 연구의 특성상 자료 수집과 출판 사이에 발생하는 일반적인 시차이나, 그 사이 기계번역 기술, 특히 생성형 AI 기반 번역 도구의 성능이 빠르게 발전하였다는 점은 별도로 고려할 필요가 있다. 따라서 본 연구에서 보고된 도구별, 텍스트 유형별 세부 수치(예: 특정 번역 도구의 정확도 순위)는 현재 시점의 절대적 성능을 대변하지 않을 가능성이 크다. 그럼에도 불구하고

앞서 논의한 텍스트 유형별 평가와 실제 품질 간의 괴리와 학습자의 과대평가 경향성은, 번역 산출물의 표층적 유창성에 기반한 인간의 인지적 경향성에서 비롯된 것으로 볼 수 있다. 이러한 경향은 도구의 성능이 향상될수록 오히려 심화될 가능성이 있어, 본 연구의 시사점은 기술 발전과 무관하게 유효할 가능성이 있다. 향후 연구에서는 최신 생성형 AI 기반 번역 도구를 대상으로 본 연구의 설계를 반복 검증함으로써, 이러한 텍스트 유형별 격차와 학습자의 관대한 평가 경향성이 기술의 발전 속도와 무관하게 지속되는 구조적 현상인지 확인할 필요가 있다.

이와 같은 한계점에도 불구하고 본 연구의 결과는 L2 지도에서 교육적 함의가 많다. 우선, 언어 교육 현장에서 학습자들에게 막연한 불신만을 가르칠 것이 아니라 미시적 오역과 MT의 한계를 스스로 식별할 수 있도록 돕는 비판적 안목을 배양할 수 있도록 MT 사후 교정(post-editing) 교육 및 MT 리터러시 훈련이 필요함을 시사한다. 우선 학습자들이 표층적 유창성 때문에 잘 해석되었다고 믿는 가독성의 덫에 걸리지 않도록 도와야 한다. 이를 위해 번역 산출물의 심층적 맥락 오류와 문화적 오역을 비판적으로 식별, 분석할 수 있도록 교육과정을 설계해야 할 것이다. 교사는 학생들이 MT 출력물을 비판적으로 사용할 수 있도록 과업을 구안하여 이를 교수학습 활동으로 적용해야 할 것이다. 예를 들어 다양한 유형의 텍스트에 대한 MT 산출물과 원문의 의미를 비교 분석하는 읽기 활동을 통해 문학 텍스트의 경우 기계번역 산출물을 맹신하기 어려운 점을 인지할 수 있도록 도와야 할 것이다. 학습자의 자율성과 어휘, 통사 등의 언어적 분별력을 신장할 수 있도록 교육하는 것이 필요한데 MT 훈련(training)은 텍스트 유형에 따라 다르게 설계, 적용될 수 있을 것이다. 예를 들어, 문학 텍스트에서는 학습자가 원문의 은유적 표현이나 어조가 번역문에서 어떻게 처리되었는지 직접 대조하고 대안적 번역을 시도해보는 활동을, ESP 텍스트에서는 전공 분야의 핵심 용어와 관용적 학술 표현이 정확한 대응어로 번역되었는지 용어집 형태로 점검하고 오류를 수정하는 활동을 설계할 수 있으며, 논설문 텍스트에서는 저자의 주장이 의도된 의미로 해석되었는지 등에 집중하여 평가하는 것이 필요하다. 뿐만 아니라, 다양한 유형의 MT 도구를 직접 사용하고 그 산출물의 질을 대조, 비교할 수 있는 기회를 제공해야 할 것이다. 학생들이 단순 어휘 매칭이나 즉각적인 한국어 대체 표현이 필요할 때는 파파고를 선호하는 반면, 텍스트에 대해 심층적인 맥락 파악이 필요한 경우에는 ChatGPT를 선택하는 경향을 감안하여 교실 환경에서 학생들이 텍스트 유형별, 과업별로 목적에 맞는 번역 도구를 활용할 수 있도록 가이드라인을 만들어 제공해야 할 것이다.

기계번역 테크놀로지의 초창기 출현은 학교 영어교육의 필요성과 유용성을 위협할 것이라는 우려와 부정적 평가를 낳기도 했다(Groves & Mundt, 2015; Jolley & Maimone, 2015). 하지만 테크놀로지의 발전을 전향적으로 수용한다면 MT는 한국의 영어 교육 환경에서 읽기 지도의 효율성을 극대화하는 새로운 가능성을 열어줄 수 있다. 디지털 기기에 익숙한 학생들의 효과적인 학습을 위해서는 MT 테크놀로지를 맹목적으로 비판할 것이 아니라, 교수학습 효과를 높이기 위한 전략을 MT 활용 읽기 교수학습 모형에 결합하고 효과적으로 MT 도구를 활용할 수 있는 방안을 적극 모색해야 할 것이다. 2022년 가을 이래 많은 생성형 AI 도구들이 개발되어 확산되고 있지만 웹을 통해 쉽게 접속하여 번역 산출물을 구할 수 있는 MT 도구의 편의성을 감안했을 때 학교 현장에서는 학생들이 생성형 AI와 함께 MT 도구들을 지속적으로 사용하고 비판적으로 활용할 수 있도록 다각도로 지원해야 할 것이다.

References

Aeni, R., Baharuddin, Putera, L. J., & Melani, B. Z. (2024). The accuracy of ChatGPT in translating linguistics text in scientific journals. *Didaktik: Jurnal Ilmiah PGSD STKIP Subang*, *10*(1), 59–68. https://doi.org/10.36989/didaktik.v10i1.2559

Alhaisoni, E., & Alhaysony, M. (2017). An investigation of Saudi EFL university students' attitudes towards the use of Google Translate. *International Journal of English Language Education*, *5*(1), 72–82. https://doi.org/10.5296/ijele.v5i1.10696

Alkhofi, A. (2025). Man vs. machine: Can AI outperform ESL student translations? *Frontiers in Artificial Intelligence*, *8*, Article 1624754. https://doi.org/10.3389/frai.2025.1624754

Bahri, H., & Mahadi, T. S. T. (2016). Google translate as a supplementary tool for learning Malay: A case study at Universiti Sains Malaysia. *Advances in Language and Literary Studies*, *7*(3), 161–167. https://doi.org/10.7575/aiac.alls.v.7n.3p.161

Briggs, N. (2018). Neural machine translation tools in the language learning classroom: Students' use, perceptions, and analyses. *JALT CALL Journal*, *14*(1), 3–24. https://doi.org/10.29140/jaltcall.v14n1.j221

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to human and machine translation quality assessment. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation quality assessment: From principles to practice* (pp. 9–38). Springer. https://doi.org/10.1007/978-3-319-91241-7_2

Cather, W. (1915). *The song of the lark.* Houghton Mifflin.

Cespedosa, A. I., & Mitkov, R. (2023). Machine translation of literary texts: Genres, times, and systems. In R. L. Gutiérrez, M. Escribe, T. Mihajlov, M. Kunilovskaya, & R. Mitkov (Eds.), *Proceedings of the First NLP4TIA Workshop* (pp. 48–53), INCOMA Ltd.

Chung, E. S. (2023). L2 learners' use of raw machine-translated output in reading comprehension. *Computer-Assisted Language*

Learning Electronic Journal, *24*(3), 1–19. https://callej.org/index.php/journal/article/view/44
Chung, E. S., & Ahn, S. (2022). The effect of using machine translation on linguistic features in L2 writing across proficiency levels and text genres. *Computer Assisted Language Learning*, *35*(9), 2239–2264. https://doi.org/10.1080/09588221.2020.1871029
Clifford, J., Merschel, L., & Munné, J. (2013). Surveying the landscape: What is the role of machine translation in language learning? *Revista de Innovación Educativa*, *10*, 108–121. http://doi.org/10.7203/attic.10.2228
Ducar, C., & Schocket, D. H. (2018). Machine translation and the L2 classroom: Pedagogical solutions for making peace with Google translate. *Foreign Language Annals*, *51*(4), 779–795. https://doi.org/10.1111/flan.12366
Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6*(3), 241–252. https://doi.org/10.2307/1266041
Groves, M., & Mundt, K. (2015). Friend or foe? Google translate in language for academic purposes. *English for Specific Purposes*, *37*(1), 112–121. https://doi.org/10.1016/j.esp.2014.09.001
High, M. D., McIntosh, A., Li, S., & Ji, Y. (2025). Student machine translation use in a transnational English-medium instruction university: Navigating development and expedience. *Journal of English-Medium Instruction*, *4*(2), 238–258. https://doi.org/10.1075/jemi.24010.hig
Hong, Seoyeon. (2025). Comparison of Korean-to-Russian translation performance using AI translation engine. *Interpreting and Translation Studies*, *29*(1), 207–233. http://dx.doi.org/10.22844/its.2025.29.1.207
Jang, Sujin, & Jwa, Soomin. (2024). An investigation of machine translator use: A case study with two Korean EFL students at different English proficiency levels. *Modern English Education*, *25*, 328–342. https://doi.org/10.18095/meeso.2024.25.1.328
Jolley, J. R., & Maimone, L. (2015). Free online machine translation: Use and perceptions by Spanish students and instructors. In A. J. Moeller (Ed.), *Learn languages, explore cultures, transform lives* (pp. 181-200). Central States Conference on the Teaching of Foreign Languages.
Karnal, A. R., & Pereira, V. V. (2015). Reading strategies in a L2: A study on machine translation. *The Reading Matrix: An International Online Journal*, *15*(2), 69–79. https://www.readingmatrix.com/files/13-5728575a.pdf
Klimova, B. (2025). Use of machine translation in foreign language education. *Cogent Arts & Humanities*, *12*(1), Article 2491183. https://doi.org/10.1080/23311983.2025.2491183
Lee, Jeong-Hwa. (2025). A comparative study of ChatGPT and Papago Korean-to-English translations of dialogue from the adaptation of Pachinko. *International Journal of Contents*, *21*(3), 138–151. https://doi.org/10.5392/IJoC.2025.21.3.139
Niño, A. (2009). Machine translation in foreign language learning: Language learners' and tutors' perceptions of its advantages and disadvantages. *ReCALL*, *21*(2), 241–258. https://doi.org/10.1017/S0958344009000172
O'Neill, E. M. (2016). Measuring the impact of online translation on FL writing scores. *IALL Journal of Language Learning Technologies*, *46*(2), 1–39. https://doi.org/10.17161/iallt.v46i2.8560
Scarton, C., & Specia, L. (2016). A reading comprehension corpus for machine translation evaluation. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 3652–3658). European Language Resources Association (ELRA). https://aclanthology.org/L16-1579
Shadiev, R., Sun, A., & Huang, Y.-M. (2019). A study of the facilitation of cross-cultural understanding and intercultural sensitivity using speech-enabled language translation technology. *British Journal of Educational Technology*, *50*(3), 1415–1433. https://doi.org/10.1111/bjet.12648
Shih, C. L. (2022). Pre-editing and machine translation: Developing marine English reading materials. *Compilation & Translation Review*, *15*(2), 123–157. http://doi.org/10.29912/CTR.202209_15(2).0004
Somers, H. (2011). Machine translation: History, development, and limitations. In K. Malmkjær & K. Windle (Eds.), *The Oxford handbook of translation studies* (pp. 427–440). Oxford University Press.
Somers, H., Gaspari, F., & Niño, A. (2006). Detecting inappropriate use of free online machine translation by language students—A special case of plagiarism detection. In V. Hansen, & B. Maegaard (Eds.), *Proceedings of the eleventh annual conference of the European association for machine translation* (pp. 41–48), European Association for Machine Translation. https://aclanthology.org/2006.eamt-1.6/
Stapleton, P., & Leung, B. K K. (2019). Assessing the accuracy and teachers' impressions of Google translate: A study of primary L2 writers in Hong Kong. *English for Specific Purposes*, *56*, 18–34. https://doi.org/10.1016/j.esp.2019.07.001
Sun, D. (2017). Application of post-editing in foreign language teaching: Problems and challenges. *Canadian Social Science*, *13*(7), 1–5. https://doi.org/10.3968/9698
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. https://doi.org/10.1007/978-3-319-91241-7_12
Tsai, S. (2019). Using google translate in EFL drafts: A preliminary investigation. *Computer-Assisted Language Learning*, *32*(5-6), 510–526. https://doi.org/10.1080/09588221.2018.1527361
van der Wees, M., Bisazza, A., & Monz, C. (2018). Evaluation of machine translation performance across multiple genres and languages. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (pp. 3822–3827). European Language Resources Association (ELRA). https://aclanthology.org/L18-1604/
Yoon, Yeo Bom. (2023). Analysis of ChatGPT's accuracy on Korean-English translation. *Korean Journal of Elementary Education*, *34*(4), 215–231. https://doi.org/10.20972/kjee.34.4.202312.215

Appendix

Representative Excerpts by Category and Text Type

Categories	Text Type	Excerpts
Lexical level	Literature	Silly 같은 단어를 모두 “어리석은”이라는 의미로 해석해 긍정, 부정 여부를 파악하기 어렵다.
Syntactic level	Opinion	영어에서 동사 앞 do가 강조 표현이란 것을 한국어로 바꿀 때 잘 표현되지 않았다.
Discourse level	Literature	여자-남자 형제의 호칭을 ‘형’이라고 번역한 것을 보고 번역기의 문화적 맥락을 번역할 수 있는 능력이 탁월하지 않다는 것을 느꼈다
Positive evaluation	ESP	학술 관련 내용이라 정확도가 높았다. 여러 번의 활동 중 오늘이 가장 정확했던 것 같다.
Metacognition	Information	기계번역을 종종 사용하는 사람으로서 신중해야겠다고 생각했다. 기계번역은 보조적인 도움으로 생각하고 나 자신이 번역의 주체가 되어야 한다는 생각이 들었다.
Text type awareness	Literature	앞 전의 텍스트와는 다르게 서사가 있는 글의 번역은 수준이 많이 떨어진다.