http://dx.doi.org/10.15749/jts.2022.23.5.005

번역학연구
23권 5호
(2022.12.31.)

풀 포스트에디팅에 대한 고찰
－풀 포스트에디팅 생산성에 영향을 주는 요소를 중심으로

이 준 호 · 김 순 미
(한국외대 · 숙명여대)

1. 서론

기계번역의 사용이 폭발적으로 증가하고 있으나, 그 결과물에는 아직 다양한 문법, 어휘, 맥락적 오류가 존재한다. 이런 배경에서 인간이 기계번역 결과물을 사후 편집하는 포스트에디팅은 학계와 업계 모두의 관심사가 아닐 수 없다. 이러한 상황 속에서 산업계는 증가하는 포스트에디팅 수요에 적극적으로 대처하기 위해 적정한 과금체계를 정립하는 것에, 학계는 인간번역과 기계번역의 특징 및 차이점 분석과 학생들을 위한 포스트에디팅 교육에 많은 관심을 보여 왔다(김순미, 신호섭, 이준호 2019). 그 외에도 포스트에디팅 가이드라인의 정립, 기계번역 품질 등 산학 공통의 관심 주제도 존재한다. 하지만 이 모든 관심과 노력은 포스트에디팅이 인간번역과 비교하여 생산성 우위가 있으며 품질은 최소한 동등해야 한다는 전제가 필요하다. 그렇지 않다면 “포스트에디팅을 실무에서 사용하는 것은 무리”가 될 것이며 “교육을 제공할 필요성 역시 감소할 수밖에 없다”(이준호 2021a: 57).

이와 같은 배경에서 한국번역학회 회원들과 언어산업계(LSP, language

service providers) 종사자들은 <2018년 신경망 분과>를 조직하여 포스트에디팅 생산성에 관한 공동연구를 진행하였다. 총 15명의 참가자가 ISO18587 가이드라인에서 제시한 라이트 포스트에디팅(light post-editing)[1]을 실시했으며, 인간번역과 포스트에디팅의 생산성 차이 비교를 위해 포스트에디팅 작업에 소요되는 '시간적 노력'과 '기술적 노력'을 살펴보았다(김순미, 신호섭, 이준호 2019: 53). 실험을 통해 '분당 단어 처리 수' 기준으로 전체 참가자의 포스트에디팅 생산성은 인간번역 대비 78% 높으며, 수정에 필요한 노력을 보여주는 TER(Translation Edit Rate)[2]은 5% 정도로 매우 낮다는 결과를 도출했다(ibid.: 65). 김순미 외(2019)는 국내 번역학계 최초의 포스트에디팅 생산성 연구로서 인간번역 대비 포스트에디팅의 생산성 우위를 확인하여, 라이트 포스트에디팅 실행의 근거를 마련했다는 의미가 있다. 그러나 라이트 포스트에디팅의 생산성만을 측정했다는 점에서 후속 연구의 필요성을 안고 있었다. 라이트 포스트에디팅은 '최소 수정을 통한 적절한 수준의 번역' 결과물을 목표로 하기에, 인간번역 수준의 결과물 및 출간을 지향하는 풀 포스트에디팅(full post-editing) 생산성은 다른 결과를 나타낼 수 있기 때문이다.

이후 국내 번역학계에서 진행된 풀 포스트에디팅[3] 생산성 연구에서는 풀

---

1) ISO18587 라이트 포스트에디팅 가이드라인
   a. 가능한 기계번역을 있는 그대로 사용.
   b. 누락 혹은 추가된 정보가 없어야 함.
   c. 부적절한 콘텐츠 모두를 수정.
   d. 부정확 혹은 모호한 의미의 경우 문장을 재구성(김순미, 신호섭, 이준호 2019: 54).

2) TER은 기계번역 결과물에 대한 품질평가를 위해 개발된 공식이다(Snover et al. 2006: 33). 하지만 TER을 통해 기계번역 결과물을 작업자가 얼마나 수정했는지 파악하는 것도 가능하다.

3) ISO18587 풀 포스트에디팅 가이드라인
   a. 누락 혹은 추가된 정보가 없어야 함
   b. 부적절한 콘텐츠 모두를 수정
   c. 부정확 혹은 모호한 의미의 경우 문장을 재구성
   d. 문법, 통사, 의미적으로 올바른 컨텐츠를 생산
   e. 고객 및 특정 영역의 언어를 준수
   f. 철자, 구두법 규칙을 적용

포스트에디팅은 '분당 단어 처리 수' 기준으로 인간번역 대비 생산성이 34% 높다고 보고하였다(이준호 2021a). 이는 김순미 외(2019)에서 관찰된 라이트 포스트에디팅 생산성 우위인 78%보다 확연히 낮은 수치로, 풀 포스트에디팅은 라이트 포스트에디팅보다 생산성이 떨어질 수 있음을 보여준 의미가 있다. 또한 이준호(2021a)는 김순미 외(2019)에서 한 발 더 나가 포스트에디팅 결과물에 대한 품질평가를 실시하여, 포스트에디팅을 통해 작업 시간이 단축되었어도 인간번역 대비 '품질'은 저하되지 않았다는 점을 보여주고 있다. 그러나 연구자가 지적하듯 짧은 실험 시간과 소수의 데이터 샘플 때문에 연구의 신뢰도가 높다고 보기는 어려우며, TER과 같은 기술적 노력에 대한 분석도 없다.

여기에 더해 두 연구 모두 포스트에디팅의 기반이 되는 기계번역 결과물(raw machine translation)의 품질에 대한 상세 분석을 시행하지 않았다. 따라서 기계번역 결과물의 품질이 작업자에게 어떠한 영향을 주고, 궁극적으로는 생산성에 어떠한 영향을 미치는지를 간과했다는 단점이 있다.

결국, 국내 번역학계에서 발표된 포스트에디팅 관련 연구는 지난 몇 년간 급증하고 있으나 포스트에디팅의 가장 근본적 연구라 할 수 있는 생산성 관련 연구는 아직 부족한 실정이다. 이에 본 연구는 풀 포스트에디팅에 관한 생산성 연구로서 상기 연구들의 단점을 보완하고 발전시키고자 한다.

첫째, 영-한 풀 포스트에디팅이 인간번역 대비 생산성 우위가 있는지를 재확인하고자 한다. 그리고 이 과정에서 생산성 증가가 품질의 희생을 전제로 이루어지고 있는 것은 아닌지 확인하고자 한다. 둘째, 풀 포스트에디팅의 생산성 논의를 위해 라이트 포스트에디팅의 생산성과 비교 분석을 진행하고, 풀 포스트에디팅의 생산성을 어떻게 이해해야 할지 제안한다. 셋째, 풀 포스트에디팅의 기술적 노력에 대한 논의를 위해 라이트 포스트에디팅의 TER과 비교 분석을 진행하고, 생산성과 기술적 노력의 상관관계를 분석한다. 넷째, 그 간의 국내 포스트에디팅 연구에서 집중적으로 논의하지 않았던 개인의 번역 속도 및 기계번역 결과물의 품질이 포스트에디팅 생산성에 어떠한 영향을 주는지 논의한다.

상기 논의를 위해 다음과 같은 연구 설계를 통해 연구 절차의 객관성 향상

---

g. 텍스트의 유형에 적절한 스타일을 사용하여 고객사가 제시한 스타일 가이드를 준수

h. 형식(formatting) 규칙을 적용(김순미, 신호섭, 이준호 2019: 54).

을 위해 노력하였다. 첫째, 텍스트 선택에 더욱 주의를 기울였다. 이준호(2021a)에서는 풀 포스트에디팅을 실시했으나 김순미 외(2019) 연구와 다른 텍스트를 사용했기에 직접적인 비교가 어려운 점이 있다. 이에 본 연구에서는 김순미 외(2019)에서 사용한 'IT 매뉴얼' 텍스트 일부를 실험에 사용하여 텍스트의 차이가 실험 결과에 영향을 줄 수 있는 변수를 배제하였다. 둘째, 이준호(2021a)의 실험보다 텍스트 양과 참가자 수를 늘려 연구의 신뢰성을 높였다. 셋째, 김순미 외(2019: 66)는 '적정 수준의 수정'에 대한 교육이 없이 전개된 연구로서 '포스트에디팅을 처음 접하는 피험자'를 대상으로 한 연구이기에, 참가자들이 지나치게 오류를 많이 식별하거나 아니면 지나치게 적게 식별하는 문제로 인해 '적절 수준의 결과물'이 도출되었는지 파악이 어렵고 생산성 손실 혹은 이익에 대한 개인차가 컸던 것이 사실이다. 이에 본 연구는 '적정 수준의 수정'이 무엇이며, 포스트에디팅 가이드라인을 사용하는 방법에 대한 교육을 시행 후 실험을 진행하여 포스트에디팅의 실제 작업 현실을 더욱 잘 반영하여 노력하였다.

이상의 준비 과정을 통해 본 연구는 풀 포스트에디팅의 인간번역 대비 생산성 우위를 재확인하고 산학협력에 도움이 되는 정보를 제공하고자 한다. 특히, 라이트 포스트에디팅과 풀 포스트에디팅에 대한 간접 비교를 통해 풀 포스트에디팅의 생산성에 대해서 어떠한 기대치를 가지는 것이 합리적인지 제언하고자 한다.

2. 선행연구

2.1 포스트에디팅 생산성 관련

2016년 11월 구글 신경망 기계번역의 출현 이후 2017년부터 국내에서도 기계번역 및 포스트에디팅 관련 연구가 쏟아져 나오기 시작했는데, 이는 기계번역 품질이 연구 대상이 될 수준으로 올라섰음을 의미한다(최문선 2019: 283). 연구 주제는 기계번역 결과물의 품질과 오류 유형, 기계번역의 발전 현황과 미래, 포스트에디팅 교육 등에 집중되어 있었고, 기계번역의 한계나 엔진 성능 비교, 포스트에디팅 가이드라인 등에 관한 연구도 소수 발표되었다(김순미, 신호

섭, 이준호 2019).

그러나 국내 번역학계에서 본격적인 생산성 연구는 거의 이루어지지 않았다(최문선 2019: 286). 더욱 큰 문제는 생산성 정보가 없다면 번역작업에 기계번역 사용 여부를 결정하기 어려우며, 번역 단가 산정도 어려울 수 있다(신지선 2020: 98). 따라서 포스트에디팅 생산성에 대해서는 학계가 업계의 요구를 따라가지 못하고 있다고 평가할 수 있다. 실제로 현재 국내 번역학계에서 찾을 수 있는 생산성 연구는 김순미 외(2019), 김자경(2022), 이준호(2021a), 이준호(2021b) 정도이며, 논문 일부에서 학생들의 포스트에디팅 시간을 측정하고 활동에 대한 소감을 분석한 연구(박혜경 2018)가 있을 뿐이다.

따라서 국내의 포스트에디팅 생산성 연구는 여전히 기초적인 데이터를 구축하는 단계이며, 생산성에 대한 논의를 고도화하는 단계라고 진단할 수 있다. 구체적인 사례로, 이준호(2021b)는 트랜스로그(Translog-II)를 활용하여 인간번역과 포스트에디팅의 키스트로크(keystroke) 양상과 휴지(pause)를 중심으로 인지적 노력을 측정했고, 시간적 노력 데이터 분석을 통해 포스트에디팅의 생산성을 측정했고, 기술적 노력과 인지적 노력이 시간적 노력과 연관성이 있는지 분석했다. 김자경(2022) 역시 키스트로크 로깅과 스크린 레코딩 프로그램을 활용하여 포스트에디팅 과정에 들어가는 시간, 기술적 노력, 검색 차원의 노력을 인간번역 과정과 비교하여 살펴보았다.

반면 해외에서는 더욱 다양한 연구 영역에서 깊이 있는 연구가 이뤄지고 있다. 해외 포스트에디팅 연구 동향에 관해 연구한 신지선(2020)은 주요 연구 동향을 15가지로 분류하고 있는데 이 중 9개 항목이 직간접적으로 생산성 연구와 관련이 있을 정도로 생산성은 중요한 주제로 연구가 되어왔다.

1) 경제적, 시간적 효율성을 중심으로 보는 기계번역의 생산성 (인간번역과 비교)
2) 포스트에디팅에 들어가는 노력 (소요 시간, 인지적 노력, 키보드나 마우스를 움직이는 등 물리적 동작 측면)
3) 번역 단가 책정 (생산성과의 상관관계)
4) 기계번역의 품질에 영향을 미치는 변수 (텍스트 종류 및 주제, 번역 방향, 언어 조합)
5) 번역평가 (자동평가 방식)

6) 과정 중심 연구 (키스트로크, 아이트래킹, 스크린 레코딩 등 활용)
7) 교육 및 훈련 (적정 훈련 기간, 커리큘럼 제안)
8) 결과물의 품질 (포스트에디팅에 소요된 시간, 인지, 물리적 노력과 결과물 품질 간의 상관관계)
9) 번역 메모리와 기계번역의 비교 (생산성, 품질) (신지선 2020: 92-97, 밑줄은 필자의 것)

연구 주제 1)과 2)는 시간 단축, 에디터의 인지적 노력, 물리적 노력 등 포스트에디팅 수행 시 핵심적인 생산성 요인들을 연구한다. 주제 3)은 시간과 노력으로 측정되는 생산성을 기반으로 번역 단가를 책정하는 연구이다. 주제 4)는 생산성이 텍스트나 언어 쌍, 번역 방향 등 외부 요인의 영향을 받는다는 점을 다룬다. 주제 5)는 객관적 생산성 연구를 위해서 사용하는 BLEU(Bilingual Evaluation Understudy), GTM(General Text Matcher), TER 등 평가방식의 적합성을 다루는 연구이며, 주제 6)은 기술적, 인지적 노력 측정을 위해 트랜스로그, 아이트래킹 등 첨단 소프트웨어 사용을 다룬다. 주제 7)은 포스트에디팅 교육이나 실험 시 학생과 피실험자들을 적절한 방법으로 교육시키는 것이 중요하다는 점을 보여주는 연구이다. 주제 8)은 생산성 산출 후 결과물이 적정 수준의 품질에 도달했는지를 보는 것이고, 9)는 번역 메모리를 사용한 인간번역과 포스트에디팅 결과물 간 품질이 얼마나 차이 나는지 등 에디팅 결과물의 품질을 분석하는 것으로, 모두 생산성 관련 중요한 연구 분야이다.

본 연구는 위의 주제 중 노력과 시간 중심 생산성 측정, 생산성과 품질, 그리고 생산성에 영향을 미치는 기계번역 품질의 수준 등의 영역에 집중하여 기존의 국내 연구에서 미진했던 논의를 보완하고 추가로 연구가 필요한 영역을 식별하고자 한다.

2.2 포스트에디팅 생산성 분석 방법

생산성 연구를 위해서는 우선 생산성의 '기준'을 무엇으로 설정할 것인지가 필수적이다. 기계번역 결과물을 평가하는데 주관성을 낮추고 시간 효율을 증대시키기 위해 인간의 번역문을 참조 기준으로 삼아 어휘 수 비율이나 에디팅 횟수 등을 비교하는 자동평가 방식인 BLEU, GTM, TER, METEOR11(Metric for

Evaluation of Translation with Explicit Ordering) 등을 이용하여 평가한 결과를 분석하는 연구가 많이 진행되고 있다(신지선 2020: 95). 그러나 이들 도구는 자동 계산되는 방식으로 정확성에서 완벽하다 할 수 없으므로 활용 가능성을 검증하는 시도는 계속되고 있다.

예를 들어 오브라이언(O'Brien 2011)은 포스트에디팅 수정률인 TER과 기계번역 결과물과 인간번역 간 일치율인 GTM, 두 가지 기준이 실제 생산성 측정에 효과적인 수단이 될 수 있는지를 연구하였다. 이를 위해 위의 기준들이 아이트래킹을 통한 시선 고정 시간(fixation time)과 포스트에디팅 수행 시간과 어떤 상관관계가 있는지 분석하였다. 결론적으로 GTM이 낮을수록(인간번역과 기계번역 결과물 간 차이가 크게 날수록), TER이 높아질수록(수정을 많이 할수록), 피시험자가 특정 부분에 시선을 고정하는 시간이 증가하고 포스트에디팅 수행 속도가 감소하는 경향이 있음을 실험을 통해 확인하였다(ibid.: 6). 물론 TER은 하나의 단어에 대해 몇 번의 수정을 했어도 최종 수정 행위만 계산에 포함하는 방식으로 작업자가 최종 결과물을 생산하기 위해 텍스트를 읽고, 판단하는 인지적 노력을 반영하지 못하는 단점이 있다(이준호 2021b: 273). 그러나 이러한 단점에도 불구하고 TER을 사용하여 기술적 노력을 평가한 연구는 다수 존재하며(Gaspari et al. 2014; Daems et al. 2019 등), 그 이유는 TER을 통한 기술적 노력의 객관화가 가져다주는 이득이 단점을 압도하기 때문일 것이다. 이에 본 연구에서도 TER을 기술적 노력의 지표로 설정하고, TER과 생산성의 관계를 논의하고자 한다.

본 연구에서 활용한 두 번째 생산성 측정 기준은 해외 많은 연구에서도 활용한 '단어처리량(throughput)'이다(김순미, 신호섭, 이준호 2019; 김자경 2022; 이준호 2021a; Aranberri et al. 2014; Plitt and François 2010). 예를 들어 아란베리 외(Aranberri et al. 2014)는 '단어처리량'을 기준으로 언어전문가와 주제전문가 사이의 포스트에디팅 생산성을 비교하는 연구를 시행했다. 이를 통해 언어전문가들은 17.66%의 생산성 향상, 주제전문가들은 12.43%의 생산성 증가세를 보인다고 보고하였다.

또한 신지선(2020)의 연구 동향 분석에서 볼 수 있듯이 생산성은 단순히 속도의 빠르고 느림에 관한 것만이 아니며, 생산성에 영향을 미치는 요소는 다양하게 존재한다. 아란베리 외(2014) 연구는 생산성이 포스트에디팅에 대한 태도

와 훈련, 원천어 텍스트의 난이도, 기계번역 결과물의 품질에 따라 영향을 받음을 보고하였다. 특히 배경지식과 언어 난이도를 개별적으로 분석했는데, 주제전문가는 언어 난이도가 높은 텍스트를 어렵게 생각했고, 언어전문가는 배경지식이 더 필요한 텍스트를 어렵게 생각하는 것으로 나타났다. 또한, 기계번역 결과물 품질의 경우 언어 학습이 잘 되어 기계번역 결과가 좋을수록 포스트에디팅 결과도 좋았다.

따라서 본 연구에서도 기존의 국내 연구에서 다뤄지지 않았던 개인의 번역 속도와 기계번역 결과물의 품질이 포스트에디팅 생산성에 어떠한 영향을 미치는지 주목하여 포스트에디팅 생산성에 대한 더욱 공고한 이해를 확보하고자 한다.

3. 연구 방법

3.1 연구 텍스트와 참가자

본 연구의 핵심 목표는 김순미 외(2019)에서 미처 논의되지 못했던 풀 포스트에디팅의 생산을 논의하는 것이다. 따라서 연구의 일관성을 위해 김순미 외(2019)에서 사용한 텍스트의 일부를 그대로 사용하여 실험을 설계하였다. 먼저 텍스트의 종류는 정보 전달 성격을 지니는 IT 매뉴얼을 선택하였다. 2018년 연구 설계 당시, 포스트에디팅 업무 실상을 반영하기 위해 영한 IT 매뉴얼이 포스트에디팅 의뢰가 가장 많다는 신경망 분과 기업들의 의견을 수렴하였다. 이후 기업측에서 주술 구조가 매우 간결한 텍스트를 선택하여 제공하였고 이를 활용하여 연구가 진행되었다. 이번 연구에서도 2018년 실험 당시 사용된 영어 원문의 일부를 그대로 사용하여 연구의 연속성을 보장하였다. 하지만 기계번역 엔진의 지속적인 발전으로 인해 2022년 현재 생성되는 기계번역 결과물을 사용할 경우 연구의 연속성을 보장할 수 없었다. 따라서 2018년 하반기 실험 당시 평가가 좋았고, 지금도 널리 사용되는 구글번역의 결과물을 그대로 사용하여 실험 및 분석을 진행하였다.

참여자 모집에 있어서는 포스트에디팅 업무 현장의 실태를 반영하고, 김순

미 외(2019)와의 연구 연속성을 위해 학부생과 통번역대학원생 모두를 포함하였다. 온라인 공지를 통해 참여자를 모집했으며, 모든 참여자는 자발적으로 참여하였다. 참여자들은 연구의 목적에 관해 설명을 들었으며, 데이터 제공에 동의하였고 소정의 사례금을 받았다. 본 연구는 참여자들의 학업에 영향을 최소화하기 위해 학기가 종료된 이후 실시했으며, 통번역대학원 3학기 재학생 8명과 학부 통번역 전공자 4학년 6명, 총 14명이 참여하였다.

본 연구의 초기 분석에서는 학부생과 대학원생의 포스트에디팅 생산성 차이가 있는지를 살펴보았으나, 학부생 그룹의 포스트에디팅이 더 빠른 경우가 1회 있었고, 대학원생 그룹의 포스트에디팅이 더 빠른 경우가 2회 있었다. 결국, 상기 데이터를 통해 대학원생과 학부생의 "번역 능력" 혹은 "경험"에 따른 생산성 차이를 판단하기 어려웠으며, 이는 본 연구의 범위도 아니다. 이에 학부생과 대학원생의 결과물 및 생산성 비교는 추가 데이터를 확보하여 별도의 연구에서 논의되어야 할 대상임을 밝힌다.

마지막으로 김순미 외(2019)의 미래 방향성 제안에 따라 모든 참여자는 포스트에디팅에 대한 기초 교육을 받았다. 구체적으로는 풀 포스트에디팅과 라이트 포스트에디팅의 차이, ISO18587에 근거한 포스트에디팅 가이드라인, 포스트에디팅 수행 절차, 신경망 번역의 전형적인 오류 유형, 기계번역 결과물의 필수 수정과 선호도에 의한 수정 등 수정의 범주 등을 이론적으로 학습했으며, 해당 이론에 기반하여 풀 포스트에디팅과 라이트 포스트에디팅을 실습하고 연구자의 피드백을 받았다. 6명의 학부 참여자는 통 번역 전공 과정에 개설된 정규 과목에서 포스트에디팅을 학습했으며, 8명의 석사과정생은 연구자가 진행한 6회의 포스트에디팅 특강에 참여하였다[4].

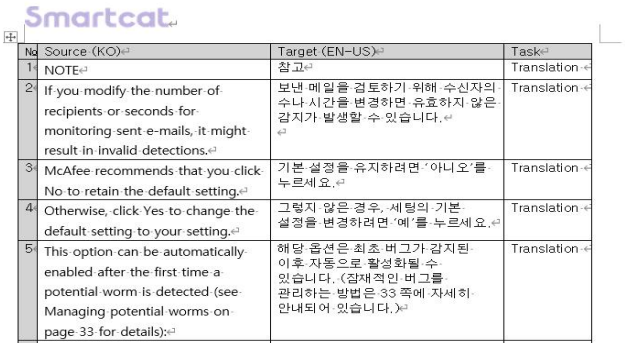
3.2 실험 설계 및 실행

실험은 온라인 환경에서 시간 통제하에 이뤄졌으며 IT 매뉴얼 텍스트를 6개의 실험용 텍스트로 나눠서 실시하였다. 연구 과정에서 번역 보조 도구를 사용할 경우 모든 참여자가 동일한 기계번역 결과물로 작업을 진행하지 않을 가

4) 참여자의 영문 이니셜 별로 참여자 번호를 배정했으며, 참고로 1, 2, 3, 7, 8, 9, 11, 14가 대학원생 그리고 4, 5, 6, 10, 12, 13이 학부생이다.

능성이 있기에, 연구자가 무료 번역 보조 도구인 스마트캣을 사용하여 MS 워드 형식으로 파일을 생성하여 동일한 텍스트를 참여자에게 배포하였다. 인간번역의 경우 출발언어 텍스트만, 포스트에디팅의 경우 김순미 외(2019)에서 사용한 것과 동일한 구글번역 결과물을 출발언어 텍스트와 병렬로 배치하였다.

〈그림 1〉 실험에 사용된 텍스트 캡처


Smartcat

№	Source (KO)	Target (EN-US)	Task
1	NOTE	참고	Translation
2	If you modify the number of recipients or seconds for monitoring sent e-mails, it might result in invalid detections.	보낸 메일을 검토하기 위해 수신자의 수나 시간을 변경하면 유효하지 않은 감지가 발생할 수 있습니다.	Translation
3	McAfee recommends that you click No to retain the default setting.	기본 설정을 유지하려면 '아니오'를 누르세요.	Translation
4	Otherwise, click Yes to change the default setting to your setting.	그렇지 않은 경우, 세팅의 기본 설정을 변경하려면 '예'를 누르세요.	Translation
5	This option can be automatically enabled after the first time a potential worm is detected (see Managing potential worms on page 33 for details):	해당 옵션은 최초 버그가 감지된 이후 자동으로 활성화될 수 있습니다. (잠재적인 버그를 관리하는 방법은 33 쪽에 자세히 안내되어 있습니다.)	Translation

실험 이전에 적절한 작업 시간 설정을 위하여, 산학협력위원회 참여 업체인 한샘EUG에 의뢰하여 사전 테스트를 시행하였다[5]. 본 실험에 사용된 텍스트와 유사한 IT 매뉴얼 텍스트를 현지화 작업자 2인에게 의뢰하여 인간번역과 포스트에디팅 작업 소요 시간을 측정하였다. 그 결과 3년 경력을 가진 작업자는 인간번역(348단어)에 17분, 포스트에디팅(344단어)에 9분이 소요되었으며, 2년 경력을 가진 작업자의 경우 인간번역(348단어)에 25분, 포스트에디팅(344단어)에 15분이 소요되었다. 상기 데이터를 근거로 본 실험에서는 350단어 미만의 텍스트를 선택하고, 작업 시간은 20분으로 설정하였다. 또한, 시간의 압박으로 인해 품질이 낮아지는 것을 방지하기 위해 시간 내에 할 수 있는 분량까지만 완전하게 번역해 달라고 참여자들에게 요청하였다.

다만 동일 주제의 텍스트에 대한 번역을 연속적으로 진행하기 때문에, 번역자들이 텍스트에 점차 익숙해진다면 생산성에 영향을 줄 가능성이 있다. 실험의 설계에 있어 이러한 변수를 완전히 배제할 수는 없을 것이나, 어떠한 번역

5) 자발적 연구 기여로 별도의 보수 없이 진행되었다.

방식을 먼저 수행하느냐에 따라 발생할 수 있는 생산성에 대한 영향도를 줄이고자 노력하였다. 이에 세트 1에서는 인간번역을 먼저 시행하고 포스트에디팅을 실시했으며, 세트 2와 3에서는 포스트에디팅을 먼저 실시하고 인간번역을 나중에 실시하였다.

〈그림 2〉 실험 절차에 대한 도식도

이상의 연구 설계를 통해 총 14명의 참여자가 3번의 세트에서 제출한 42개의 인간번역과 42개의 포스트에디팅 결과물을 확보했으며, 전체 결과물에 대한 시간적 노력을 1분당 처리한 단어 수 기준으로 분석하였다. 또한, 작업자들의 수정량을 파악하기 위해 42개의 포스트에디팅 결과물을 소프트웨어가 인식할 수 있는 형태로 태깅하고 자바스크립트를 사용하여 TER을 산출하였다.

다음으로 포스트에디팅의 생산성 증가가 품질을 희생하며 확보한 것인지 파악하기 위해 각 세트에서 인간번역 대비 포스트에디팅의 생산성이 매우 높은 참여자와 낮은 참여자의 결과물을 추출하여 평가를 진행하였다. 물론 포스트에디팅의 품질평가가 본 연구의 주된 목적은 아니지만, 코엔(Kohen 2009)이 제시한 다양한 평가 방법 중 충분성과 유창성을 기준으로 평가를 진행했으며, 이를 통해 인간번역 대비 포스트에디팅 결과물이 최소한 비열등성을 지니는지를 알아보고자 하였다.

마지막으로, 포스트에디팅 생산성에 영향을 줄 것으로 예상되는 기계번역 결과물의 오류를 모두 분석하였다. 추가로 포스트에디팅 생산성에 영향을 줄 수 있는 요소 탐색을 위해 설문 조사를 통해 참여자의 인적 데이터 및 텍스트에 대한 의견을 수집하였다.

4. 분석 결과

본 장에서는 풀 포스트에디팅 생산성을 보여주는 시간적 노력, 생산성과 품질, 수정량, 기계번역 결과물의 품질 등에 대한 분석 결과를 제시한다. 이를 통해 포스트에디팅 생산성에 대한 더욱 공고한 정보를 제공하고, 추가적 연구 주제를 도출하며, 포스트에디팅 서비스 및 단가 형성에 대한 산학 간 논의에 이바지하고자 한다.

4.1 분당 단어 처리 수 기반 생산성

세트 1에 대한 인간번역과 포스트에디팅의 생산성 결과는 <표 1>과 같다. 생산성 측정 결과 포스트에디팅 분당 단어 처리 수는 평균 23.38로 인간번역 평균인 11.09 대비 두 배 이상 높은 것으로 나타났다.

참여자 간의 개인차가(표준편차 0.7) 존재하는 것은 사실이지만 인간번역이 포스트에디팅보다 빠른 경우는 없었다. 다만 한 가지 예상하지 못한 결과는 세트 1의 포스트에디팅 생산성 우위가 김순미 외(2019)에서 보고한 라이트 포스트에디팅 생산성 우위인 78%보다 더 높다는 점이다. 풀 포스트에디팅을 실시했는데, 라이트 포스트에디팅보다 더 높은 생산성 우위가 나타났다는 것은 의미 있는 결과이며 그 원인과 함의를 결론에서 추가로 논의하고자 한다.

〈표 1〉 세트 1의 생산성 결과

	인간번역 1 (337단어)	포스트에디팅 1 (342단어)	인간번역 대비 생산성 (배수)
참여자 1	12.55	25.02	1.99
참여자 2	11.00	22.31	2.03
참여자 3	11.90	18.66	1.57
참여자 4	8.70	24.07	2.77
참여자 5	12.50	28.50	2.28
참여자 6	8.70	24.96	2.87
참여자 7	11.90	17.72	1.49
참여자 8	12.50	25.33	2.03
참여자 9	9.60	24.14	2.51
참여자 10	8.05	34.20	4.25
참여자 11	18.04	25.77	1.43

참여자 12	6.65	14.20	2.14
참여자 13	9.60	19.00	1.98
참여자 14	13.55	23.47	1.73
평균	**11.09**	**23.38**	**2.22**

세트 2에 대한 인간번역과 포스트에디팅의 생산성 결과는 <표 2>와 같다. 인간번역 대비 포스트에디팅 생산성 결과는 세트 1의 평균인 2.22보다 감소한 1.66으로 포스트에디팅이 인간번역 대비 66% 정도 높은 생산성을 보였다.

〈표 2〉 세트 2의 생산성 결과

	포스트에디팅 2 (334단어)	인간번역 2 (334단어)	인간번역 대비 생산성 (배수)
참여자 1	17.86	12.05	1.48
참여자 2	19.65	14.80	1.33
참여자 3	22.61	9.95	2.27
참여자 4	21.17	16.70	1.27
참여자 5	20.55	17.51	1.17
참여자 6	17.06	13.65	1.25
참여자 7	23.03	15.45	1.49
참여자 8	19.65	16.70	1.18
참여자 9	38.61	10.10	3.82
참여자 10	14.90	14.45	1.03
참여자 11	17.06	14.80	1.15
참여자 12	16.70	7.35	2.27
참여자 13	16.70	11.75	1.42
참여자 14	31.57	15.45	2.04
평균	**21.22**	**13.62**	**1.66**

또한, 세트 3에 대한 인간번역과 포스트에디팅의 생산성 결과는 <표 3>과 같다. 인간번역 대비 포스트에디팅 생산성 평균은 1.21로 인간번역 대비 21% 정도 빠른 생산성을 보였다. 특이한 점은 다른 세트와 달리 세트 3에서는 4명의 참여자가 인간번역이 더 빠른 경우가 관찰되었다. 이 때문인지 세트 3은 다른 세트 대비 가장 낮은 인간번역 대비 생산성을 기록하였다.

〈표 3〉 세트 3의 생산성 결과

	포스트에디팅 3 (334단어)	인간번역 3 (334단어)	인간번역 대비 생산성 (배수)
참여자 1	17.05	14.30	1.19
참여자 2	20.78	15.15	1.37
참여자 3	16.51	9.55	1.73
참여자 4	26.90	21.38	1.26
참여자 5	19.55	17.83	1.10
참여자 6	17.67	18.18	0.97
참여자 7	18.28	15.15	1.21
참여자 8	27.42	17.30	1.58
참여자 9	23.58	15.15	1.56
참여자 10	20.27	21.63	0.94
참여자 11	19.15	17.27	1.11
참여자 12	10.00	14.15	0.71
참여자 13	13.40	17.30	0.77
참여자 14	30.46	20.14	1.51
평균	**20.07**	**16.75**	**1.21**

이상 3개 세트 결과에 따르면 풀 포스트에디팅이 인간번역 대비 69.66% 정도의 생산성 우위를 보였다. 이를 김순미 외(2019)에서 보고한 라이트 포스트에디팅의 인간번역 대비 생산성 우위인 78%와 비교하면 8% 포인트 정도 낮은 수치이다. 상기 데이터는 물론 두 실험의 참여자가 다르다는 점을 고려하더라도, 풀 포스트에디팅을 실시하면 라이트 포스트에디팅보다 생산성이 감소할 수 있음을 실증적으로 보여주었다는 의미가 있다. 하지만 라이트 포스트에디팅 대비 두드러지는 생산성 감소는 아니라는 점에 주목할 필요가 있다.

4.2 생산성과 품질의 관계

포스트에디팅의 생산성이 높을 경우, 품질을 희생하고 높은 생산성을 확보한 것은 아닐지에 유의할 필요가 있다. 이를 검증하기 위해서는 인간번역 대비 포스트에디팅의 생산성이 매우 높은 참여자의 인간번역 및 포스트에디팅 결과물의 품질 비교가 필요하다. 인간번역 대비 포스트에디팅 생산성이 매우 높은 참여자의 결과물 분석 결과 포스트에디팅이 인간번역 대비 품질이 현격히 낮다면 품질을 희생하면서 생산성을 확보한 것이기 때문이다. 또한, 상기 논거를 강

화하기 위해서는 인간번역 대비 포스트에디팅의 생산성이 높지 않았던 참여자들의 인간번역과 포스트에디팅 결과물이 유사한 수준의 품질을 나타내는지 평가해볼 필요가 있다. 하지만 현실적으로 84개의 모든 텍스트에 대해서 품질을 검토하는 것은 어려운 일이다. 따라서 각 데이터 세트에서 인간번역 대비 포스트에디팅의 생산성이 가장 높은 참여자 2명과 인간번역 대비 포스트에디팅의 생산성이 가장 낮은 참여자 2명의 텍스트를 추출하여, 인간번역과 포스트에디팅 결과물 간에 현격한 품질 차이가 있는지 비교하였다.

〈표 4〉 포스트에디팅 생산성 차이 상위/하위 참여자 샘플링

	생산성 차이 낮음	생산성 차이 높음
세트 1	참여자 7과 11	참여자 10과 6
세트 2	참여자 5와 8	참여자 9와 3
세트 3	참여자 12와 13	참여자 14와 2

포스트에디팅 결과물이 인간번역 대비 열등하지 않은 품질을 보였는지를 평가하기 위해 번역학 박사과정을 전공하고 학부에서 번역 전공자를 지도하고 있는 교수자 1명과 산업계의 현지화 전문가 1인에게 하기 충분성과 유창성 기준에 따라 평가를 요청하였다.

1) 충분성: 결과물이 출발언어 텍스트의 의미를 충실하게 전달하고 있다.

10점	출발언어 텍스트의 의미를 매우 충실하게 반영하고 있다.
8점	출발언어 텍스트의 의미를 충실하게 반영하고 있으나, 미미한 오류가 있다.
6점	출발언어 텍스트의 전반적인 의미를 반영하고 있으나, 중대 오류가 있다.
4점	출발언어 텍스트의 의미를 반영하고 있지 못하며, 중대 오류가 많다.
2점	출발언어 텍스트의 의미를 전혀 반영하지 못하고 있다.

(상기 평가 기준은 연구자가 작성)

2) 유창성: 결과물이 자연스러운 한국어이며, 구성 및 문법의 어색함이 없다.

10점	자연스러운 한국어를 구사하며, 문장의 구성 및 문법에 있어 문제가 전혀 없다.
8점	전반적으로 자연스러운 한국어를 구사하나, 문장의 구성 및 문법에 있어 사소한 문제가 있다.

6점	한국어 구사가 다소 부자연스럽고, 문장의 구성 및 문법에 있어 몇 가지 문제가 있다.
4점	한국어 구사가 부자연스럽고, 문장의 구성 및 문법에 중대한 문제가 있다.
2점	매우 부자연스러운 한국어의 구사 및 다수의 중대한 오류가 있다.

(상기 평가 기준은 연구자가 작성)

상기 정량 평가 결과를 요약하자면 총 24번의 유창성과 충분성 평가에서 평가자 1이 인간번역을 우수하게 평가한 경우는 4회뿐 이었다. 반면 포스트에디팅이 우수하다고 평가한 경우는 13회나 되었으며, 동률은 7회였다. 평가자 2 역시 인간번역을 우수하게 평가한 경우는 한 번밖에 없었다. 반면 포스트에디팅이 우수하다고 평가한 경우는 11회나 되었으며, 동률은 13회였다.

따라서 인간번역 대비 포스트에디팅 생산성이 높은 참여자들이 품질을 희생하여 생산성을 확보했다는 근거를 찾기는 어려웠다. 또한, 인간번역 대비 포스트에디팅의 생산성이 '높은' 그룹이 '낮은' 그룹 대비 포스트에디팅 결과물의 품질이 열등하다는 근거를 찾을 수도 없었다.

요약하자면 두 평가자 모두 포스트에디팅의 충분성과 유창성이 높다고 응답한 횟수가 압도적으로 높으며, 평가자 간에 인간번역과 포스트에디팅의 우위에 대한 의견이 엇갈린 경우는 1회뿐이었다. 따라서 4.1장에서 확인한 포스트에디팅의 높은 생산성은 대부분의 경우 품질을 희생하면서 얻은 것이 아니라 할 수 있다.

물론 상기 품질평가는 간소화된 분석이라는 점에 주의할 필요가 있다. 또한 김자경(2022: 20)은 "포스트에디팅에서 시간 절감이 상대적으로 적게 나타난 참가자들에게서 정확성 오류 수정 결과가 더 좋았음"을 주장하고 있다. 따라서 후행 연구에서는 모든 참여자의 포스트에디팅 결과물을 상세하게 분석하여, 작업 시간과 오류 수정의 성공률을 연계해서 분석한다면 생산성에 대한 더욱 공고한 정보 제공에 기여할 수 있을 것이다.

4.3 TER, 기술적 노력

TER은 기계번역을 수정하는 기술적 노력을 나타내는 대표적인 지표 중에 하나로서 김순미 외(2019)에서는 구글, 네이버 파파고, 카카오 등 각 번역 엔진

별로 TER을 계산했으며 구글번역 결과물의 수정률은 약 3%에 불과하다고 보고한 바 있다. 하지만 이번 실험은 풀 포스트에디팅이고, 학습자들이 풀 포스트에디팅의 품질 수준 달성을 위해서 무엇을 수정해야 하는지를 학습한 경험이 있으므로 TER이 적어도 3%보다는 높을 것으로 예상하였다.

〈표 5〉 TER 결과

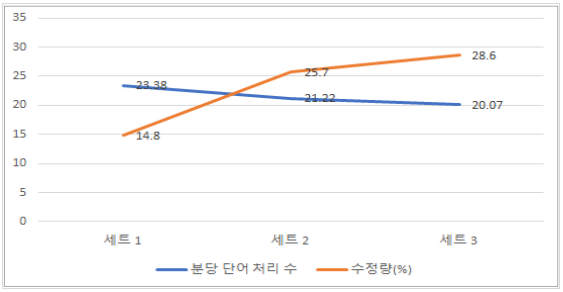
	포스트에디팅 1	포스트에디팅 2	포스트에디팅 3
참여자 1	0.108	0.212	0.277
참여자 2	0.123	0.306	0.273
참여자 3	0.223	0.191	0.233
참여자 4	0.145	0.270	0.245
참여자 5	0.097	0.363	0.373
참여자 6	0.234	0.306	0.418
참여자 7	0.167	0.201	0.225
참여자 8	0.149	0.259	0.289
참여자 9	0.212	0.129	0.269
참여자 10	0.112	0.392	0.446
참여자 11	0.033	0.198	0.225
참여자 12	0.297	0.335	0.333
참여자 13	0.134	0.345	0.329
참여자 14	0.041	0.086	0.068
평균	**0.148**	**0.257**	**0.286**

<표 5>에서 확인할 수 있듯 TER 평균은 0.230으로(23% 수정), 풀 포스트에디팅 모드에서 참여자들이 확실히 라이트 포스트에디팅 대비 많은 수정을 가했다는 것을 알 수 있다. 여기서 주목할 부분은 풀 포스트에디팅에서 라이트 포스트에디팅 대비 7배 이상(3%에서 23%로 증가)의 수정을 했지만, 시간적 노력은 단 8% 포인트만 차이를 보인다는 사실이다. 즉, 많이 수정한다고 반드시 작업 시간이 선형적으로 증가하는 것은 아니며, 수정량과 작업 시간은 정비례하지 않을 수 있다는 가정이 가능하다.

구체적으로 생산성과 작업 시간의 상관관계를 살펴보면, 세트 1의 포스트에디팅 분당 단어 처리 수는 23.38, 세트 2는 21.22, 세트 3은 20.07로 세트 간의 차이가 크지 않다. 하지만 TER 평균은 세트 1이 0.148, 세트 2가 0.257, 세트 3이 0.286으로, 세트 1과 세트 3 사이에는 두 배 가까운 TER 차이가 존재한다.

따라서 생산성과 수정의 양이 정비례하지 않는다.

〈그림 3〉 포스트에디팅 생산성과 TER 평균의 추세선



하지만 통계적으로 모든 참여자의 '분당 포스트에디팅 단어 처리 수(생산성)'와 'TER(수정량)'의 관계를 분석한 결과를 보면 조금은 다른 해석이 가능하다. 피어슨 상관관계 분석은 두 가지 변수 간의 상관관계를 수치상으로 나타낼 수 있다. 상관관계 분석 해석 가이드라인에 따르면 학자마다 수치의 해석에 차이는 있지만, R이 0.1에서 0.3 사이면 낮은 상관관계, 0.3에서 0.5 사이면 중간 정도의 상관관계, 0.5 이상이면 제법 강한 상관관계, 1.0이면 완전한 상관관계가 있다고 볼 수 있으며, P<0.5일 경우 통계의 유의성을 주장할 수 있다(Akoglu 2018).

〈표 6〉 포스트에디팅 단어 처리 수와 TER의 관계

	R value	P value	해석
세트 1	-0.540	0.047	제법 강한 부적 상관관계, 통계적 유의미
세트 2	-0.752	0.002	강한 부적 상관관계, 통계적 유의미
세트 3	-0.488	0.077	통계적으로 유의하지 않음

<표 6>에서 보듯이 세트 1에서는 생산성과 수정량 간에 -0.540의 부적 상관관계를 보였고 이는 통계적으로 유의미했다. 또한, 세트 2에서는 -0.752이란 부적 상관관계를 보였고 이 역시 통계적으로 유의미했다. 반면 세트 3에서는 통계적으로 유의미하지는 않지만, 중간 정도의 부적 상관관계를 관찰할 수 있었다.

따라서 수정량과 생산성은 부적 상관관계, 즉 수정량이 증가하면 생산성이 감소하는 관계가 있다는 것을 확인할 수 있었다. 하지만 모든 데이터에서 통계적으로 유의미한 부적 상관관계가 있는 것은 아닌 것으로 나타났다. 이에 적어도 본 실험 결과, 수정량과 생산성의 감소는 '일부' 상관관계가 있는 것으로 볼 수 있으나 그 상관관계가 절대적이라 주장하기는 어려워 보인다.

또한 김자경(2022), 이준호(2021b) 역시 모두 수정량이 생산성을 충분히 설명하지 못함을 지적했음을 감안하면, 수정량만으로 전체 작업 시간을 설명하는 것은 무리가 있어 보인다. 반면 김순미 외(2019) 대비 수정률이 매우 높지만, 생산성이 많이 줄어들지 않은 이유는 연구 참여자의 철저한 사전 교육 여부에서 찾을 수 있을 것이다. 김순미 외(2019)는 정규 과정에서 포스트에디팅을 학습한 인원이 거의 없었다. 하지만 이번 연구에는 포스트에디팅 가이드라인에 기반하여 기계번역의 전형적 오류, 수행 방법, 수정의 적절성 등을 학습한 인원만 참여하였다. 여기에 더해 상대적으로 번역에 대한 경험이 많은 것으로 볼 수 있는 통번역 전문 석사과정생들이 참여자의 다수라는 점 역시 고려할 필요가 있다. 따라서 참여자들의 학습 수준과 번역 활동에 대한 능숙함이 '단시간에 다수의 수정'을 가능하게 하는지는 추후 연구가 필요하다.

4.4 생산성에 영향을 준 요소에 대한 분석

3개 세트의 실험 결과는 포스트에디팅이 인간번역 대비 전반적으로 생산성이 높지만, 참여자 개인마다 포스트에디팅 생산성에 차이가 있으며, 세트를 진행함에 따라 인간번역 대비 포스트에디팅의 생산성 우위가 감소하는 경향을 보여주었다. 이번 장에서는 그 원인을 데이터에 기반하여 분석해 보고자 한다.

4.4.1 개인별 포스트에디팅 생산성 차이

가장 먼저 생각해볼 수 있는 점은 '개인의 번역 수행의 속도가 포스트에디팅 수행 속도에 영향을 미치는가?'이다. 이를 조금 더 구체적으로 검증하기 위해 인간번역의 생산성과 포스트에디팅의 생산성 간의 상관관계를 통계적으로 분석하였다. 즉, 번역의 속도가 빠른 참여자가 포스트에디팅 속도도 빠른가에 대해서 살펴보았다.

〈표 7〉 개인별 번역과 포스트에디팅 속도 간 상관관계

	R value	P value	결과
세트 1	0.144	0.623	약한 양적 상관관계, 통계적으로 유의미하지 않음
세트 2	-0.087	0.770	약한 부적 상관관계, 통계적으로 유의미하지 않음
세트 3	-0.516	0.059	제법 강한 부적 상관관계, 통계적으로 유의미하지 않음

하지만 위의 표에서 확인할 수 있는 것처럼 R 값이 크지 않으며 두 변수 사이에 통계적 유의성 역시 찾기 어렵다.

또 한 가지 중요한 점은 김자경(2022: 10)에서 지적하듯 포스트에디팅을 통한 작업 시간 절감의 정도에 개인차가 존재했다는 점이다. 또한, 본 연구에서도 세트별로 개인의 번역 속도와 포스트에디팅 속도 간에 일관성을 찾을 수 없었다. 즉, 일정하게 번역 속도가 빠르거나 포스트에디팅 속도가 빠르게 나타나기보다는 세트 별로 차이가 나는 경우가 많았다. 예를 들어 개인 참여자 별로 볼 때, 참여자 4명(6, 10, 12, 13)은 세트 3에서는 번역을 하는 경우보다 포스트에디팅의 속도가 더 느린 것을 볼 수 있다. 그러나 이 중 참여자 10의 경우 세트 1에서는 번역 대비 포스트에디팅의 생산성이 4.25배로 1위, 참여자 6의 경우 2.87배로 2위를 했다.

따라서 적어도 본 연구에 국한해서는 번역 속도와 포스트에디팅 속도는 연관이 있다고 보기 어려워 보인다. 그러나 이는 매우 제한된 연구 결과이기 때문에 개인별 생산성 차이에 대해 더 객관적인 연구를 위해 다양한 텍스트에 대해 개인의 번역 속도와 포스트에디팅 속도를 측정하는 후속 연구가 필요할 것이다.

4.4.2 기계번역 결과물의 영향

아란베리 외(2014)는 기계번역 결과물(raw machine translation)의 품질이 포스트에디팅의 생산성에 영향을 미칠 수 있음을 주장하였다. 즉 기계번역 결과물에 오류가 많고, 수정에 큰 노력이 필요하다면 포스트에디팅 생산성은 감소할 수밖에 없다는 것이다. 이에 본 연구에서는 기계번역 결과물을 문장 단위로 분석하여 오류가 있는지, 수정에 필요한 노력이 어느 정도인지에 근거하여 다음과 같이 각 문장에 정량적 태깅을 시행하였다.

〈표 8〉 기계번역 결과물 에러와 수정 노력 태깅

수정 필요한 부분	수정에 필요한 노력
수정할 부분이 전혀 없음	0
오류가 없으며 1단어 수준의 수정 필요	0.5
큰 오류가 없으며 2~3단어 수준의 수정 필요	1.0
큰 오류가 있어 대안 모색에 인지적 노력이 필요하며 2~3단어 이상의 수정 필요	2.0
중차대한 오류가 있어 대안 모색에 인지적 노력이 필요하며 4~5단어 이상의 수정 필요	3.0

(연구자가 작성한 기준)

예를 들어 악성 요소를 검사하는 행위를 나타내기 위해 출발언어 텍스트에서는 “scan”을 동사 혹은 명사로 자주 사용하고 있다. 하지만 기계번역 결과물에서는 “scan”이라는 동일한 단어를 “스캔”, “검색”, “검사” 등으로 번역하고 있다. 이 경우 작업자가 의사결정만 한다면 하나의 등가로 통일해서 작업할 수 있는 부분이다. 따라서 매우 단순한 수정이라고 볼 수 있고 본 분석에서는 0.5의 노력으로 태깅하였다.

위의 태깅에 근거하여 문장 단위로 분석한 결과는 <표 9>와 같으며, 추가적 노력이 있어야 하는 오류가 세트 3에서 가장 많고(15.5), 세트 1에서 가장 적다(9.5)는 것을 알 수 있다.

〈표 9〉 기계번역 에러 수정에 필요한 노력

	총 추가 노력	1.0 이상의 비중	분당 단어 처리 수
세트 1	9.5	1.0: 2회	23.38
세트 2	10.5	1.0: 3회 2.0: 2회	21.22
세트 3	15.5	1.0: 8회 3.0: 1회	20.07

이를 생산성 지표인 분당 단어 처리 수와 연계해서 본다면, 기계번역 결과물의 오류의 빈도와 심각도가 증가하면서 생산성이 떨어졌음을 관찰할 수 있었다. 특히 심각한 오류가 발생하면 번역사는 문제를 탐지, 식별, 해결책을 제안, 검토하는 과정에서 많은 시간을 쓸 수밖에 없으므로, 심각도가 높은 오류의 빈도가 높으면 생산성이 저해된 것으로 풀이된다. 이는 영어 한국어 언어쌍에서

도 기계번역 결과물의 오류의 빈도와 심각도가 생산성에 영향을 줄 수 있다는 관찰이기에 의미가 있으며, 향후 해당 주제에 대해 다양한 오류 유형에 대한 인지적 노력의 차이에 대해 심도 있는 논의가 필요해 보인다.

하지만 본 분석에서 더욱 유의해서 보아야 할 점은 연구자가 분석한 세트 1, 2, 3의 80문장 중, 전혀 수정이 필요 없는 문장이 무려 34문장이나 되었으며 추가 노력 0.5로 연구자가 설정한 매우 단순한 1단어 수준의 수정이 필요한 경우도 무려 29문장이나 되었다는 것이다.

이상의 분석이 의미하는 바는 본 연구에 사용된 기계번역 결과물은 오류의 수가 적었으며, 그 품질이 매우 높았다는 것이다. 그리고 이처럼 높은 품질 때문에 김순미 외(2019)에서 3%의 TER이 관찰된 것으로 풀이된다. 실제로 본 연구 참여자들도 기계번역 결과물의 품질에 대해 13명이 “좋음”, 1명이 “매우 좋음”이라 설문에서도 응답하였다. 따라서 김순미 외(2019)에서 관찰된 인간번역 대비 라이트 포스트에디팅의 생산성 우위 78% 및 본 연구에서 관찰된 인간번역 대비 풀 포스트에디팅의 생산성 우위 69.66%를 모든 상황에 일반적으로 적용해서는 안 될 것이다. 즉, 본 연구에서 사용된 기계번역 결과물은 이례적으로 높은 품질을 보였으며, 이보다 낮은 품질의 기계번역 결과물에 대하여 포스트에디팅을 진행한다면 생산성이 본 연구에서 보고한 수치보다 낮을 가능성이 있다.

5. 논의 및 결론

5.1 분석 결과에 대한 논의

본 연구를 통해 포스트에디팅이 인간번역보다 높은 생산성 그리고 인간번역 대비 열등하지 않은 수준의 품질을 재확인했다는 점에서 의미가 있다. 또한, 최소 수정을 통한 적정 수준의 번역을 추구하는 라이트 포스트에디팅의 경우 인간번역 수준을 추구하는 풀 포스트에디팅과 생산성 등에 차이가 있을 수밖에 없음을 관찰하였다.

우선 분당 처리 단어 수를 통해 살펴본 인간번역 대비 생산성 우위에 있어,

풀 포스트에디팅에서는 라이트 포스트에디팅 대비 생산성 우위가 78%에서 69.66%로 약 8% 포인트 감소한 것을 볼 수 있었다. 본 연구는 라이트 포스트에디팅 생산성을 연구한 김순미 외(2019)와 같은 'IT 매뉴얼' 텍스트를 활용했기에 영한 풀 포스트에디팅에서는 영한 라이트 포스트에디팅 대비 생산성이 감소할 수 있음을 실험적으로 증명했다는 의의가 있다.

따라서 라이트 포스트에디팅과 풀 포스트에디팅은 생산성 차이가 있으며, 그 차이를 명확하게 서비스 제공사나 고객 모두 이해할 필요가 있을 것이다. 특히 본 연구의 결과에 따르면 라이트 포스트에디팅과 풀 포스트에디팅의 속도 차이가 10% 미만이라는 관찰 사항은 포스트에디팅 서비스 요율 산정에 참조할 수 있는 초기 수치가 될 수 있을 것이다. 하지만 두 그룹 참여자의 포스트에디팅 교육 수준이 달랐기에, 추후 연구를 통해 동일 참여자의 라이트 포스트에디팅과 풀 포스트에디팅 생산성 비교 연구가 필요할 것이다.

둘째, 본 연구는 포스트에디팅의 생산성 우위가 번역 품질을 희생하고 확보한 것이 아님을 샘플 데이터를 통해 확인하였다. 특히 인간번역 대비 포스트에디팅 생산성 우위가 높은 그룹과 낮은 그룹의 품질 차이가 없음도 관찰하였다. 또한, 본 연구에서 사용된 것처럼 기계번역 결과물의 품질이 높다면 인간번역보다 포스트에디팅의 품질이 더 우수할 수 있음을 알 수 있었다. 이를 확대하여 해석하면, IT 매뉴얼이나 의학, 전자, 기계, 군사 등 분야의 정형화된 텍스트의 경우 포스트에디팅을 통해 생산성을 증가시키면서도 품질을 희생하지 않는 번역이 가능함을 알 수 있다는 것은 본 연구의 큰 의미라 할 수 있다.

셋째, 풀 포스트에디팅과 라이트 포스트에디팅 간 생산성 차이가 의외로 작았던 것에 비해 TER로 대표되는 수정량은 큰 차이가 있었다. 즉, 라이트 포스트에디팅 수정량 3% 대비 풀 포스트에디팅의 수정량은 23%로 7배 이상 상승한 것이다. 상기 관찰이 업계에 가지는 시사점은 수정량이 생산성을 완전하게 대변할 수는 없다는 것이다. 통계적 상관관계 분석을 통해서도 수정량이 생산성을 일부만 대변하는 변수임을 확인할 수 있었다. 따라서 수정량 자체로 요율을 계산하는 것에 주의가 필요할 것이다. 두 번째로 풀 포스트에디팅 참여자들이 적은 시간 내에 많은 수정을 했다는 것이다. 이는 김순미 외(2019) 연구에서 피실험자에 대한 포스트에디팅 교육이 제대로 이루어지지 않았던 것에 큰 부분 기인할 수도 있다. 따라서 포스트에디팅 훈련 정도, 번역 교육 경험, 번역 능력

등을 중심으로 라이트 포스트에디팅과 풀 포스트에디팅을 비교하는 연구가 실행된다면, 어떤 교육을 포스트에디팅 업무 시작 전에 제공해야 할지 파악이 가능할 것이다.

넷째, 본 연구는 기존의 연구에서 실행하지 않았던 기계번역 결과물 오류 빈도와 양상이 생산성에 미치는 영향에 주목하였다. 단순 반복적인 IT 매뉴얼 텍스트를 사용했기 때문에 기계번역 결과물의 오류 빈도도 낮았고 치명적인 오류도 적음을 텍스트 분석을 통해 알 수 있었고, 오류의 빈도와 치명도가 생산성에 부정적인 영향을 줄 수 있음을 관찰하였다. 이 관찰은 번역 업계에 중요한 시사점을 지닌다. 또한, 본 연구 및 김순미 외(2019)에서 제시된 인간번역 대비 생산성 우위는 예외적으로 높은 수치이며, 이 수치를 참조하여 작업 생산성을 예측하는 것은 무리가 있을 수 있다.

5.2 연구의 한계와 의의

본 연구는 최근 증가하는 풀 포스트에디팅이 번역 서비스로서 의미를 가지기 위한 생산성과 품질 모두에 대해서 주목하였다는 점에서 의미가 있다. 또한, 단순 생산성 분석에 그치지 않고 생산성에 영향을 줄 수 있는 요소를 다양한 각도에서 고민하고 미래 연구 주제를 제시하고 있다. 다만 본 연구는 제한적 데이터를 기반으로 결과를 제시한 것이기에, 본 연구의 결과를 일반론적으로 해석하는 데는 주의가 따른다. 따라서 본 연구에서 제시하고 있는 연구 주제와 관련 추가적인 연구가 필요하며, 다양한 환경에서 도출되는 결과를 추가로 분석할 필요가 있다.

참고문헌

김순미, 신호섭, 이준호 (2019)「번역학계와 언어서비스업체(LSP)간 산학협력연구: '포스트에디팅 생산성'과 '기계번역 엔진 성능 비교」,『번역학연구』20(1): 41-76.

김자경 (2022)「한영 포스트에디팅 과정에서의 노력 탐색 – 시간, 기술적 노력, 검색을 중심으로」,『통번역학연구』26(2): 1-24.

박혜경 (2018)「석사 과정의 기계번역 수업에 대한 소고: 한일번역 전공생의 포스트에디팅 사례를 통하여」,『번역학연구』19(3): 163-193.

신지선 (2020)「기계번역 포스트에디팅에 관한 해외 연구 동향」,『번역학연구』21(4): 87-114.

이준호 (2021a)「영한 포스트에디팅 생산성에 대한 고찰 – 시간적 노력을 중심으로」,『통번역학연구』25(2): 55-83.

이준호 (2021b)「한영 포스트에디팅 노력 예비연구: 트랜스로그 II를 활용한 한영인간번역과 포스트에디팅의 차이 분석」,『번역학연구』22(5): 271-298.

최문선 (2018)「국내 번역학 기계번역 연구 동향: 내용 분석과 키워드 분석을 중심으로」,『언어학연구』24(1): 275-297.

Akoglu, Haldun (2018) 'User's Guide to Correlation Coefficients', *Turkish Journal of Emergency Medicine* 18(3): 91-93.

Aranberri, Nora, Gorka Labaka, Arantza Díaz de Ilarraza and Kepa Sarasola (2014) 'Comparison of Post-editing Productivity between Professional Translators and Lay Users', *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, 20-33.

Daems, Joke and Lieve Macken (2019) 'Interactive Adaptive SMT versus Interactive Adaptive NMT: A User Experience Evaluation', *Machine Translation* 33(1): 117-134.

Daems, Joke, Sonia Vandepitte, Robert Hartsuiker and Lieve Macken (2017) 'Translation Methods and Experience: A Comparative Analysis of Human Translation and Post-editing with Students and Professional Translators', *Meta: Translators' Journal* 62(2): 245-270.

Gaspari, Federico, Antonio Toral, Sudip KumarNaskar, Delcan Groves and Andy Way (2014) ‘Perception vs. Reality: Measuring Machine Translation Post-editing Productivity’, *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, 60-72.

Koehn, Philipp (2009) *Statistical Machine Translation*, New York: Cambridge UP.

O’Brien, Shanon (2011) ‘Towards Predicting Post-editing Productivity’, *Machine Translation* 25(3): 197-215.

Plitt, Mirko and Masselot François (2010) ‘A Productivity Test of Statistical Machine Translation Post-editing in a Typical Localisation Context’, *The Prague Bulletin of Mathematical Linguistics* 93(January 2010). Available at http://ufal.mff.cuni.cz/pbml/93/art-plitt-masselot.pdf.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul (2006) ‘A Study of Translation Edit Rate with Targeted Human Annotation’, *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223-231.

[Abstract]

An Investigation into English to Korean Full Post-editing: Factors Affecting Productivity of Full Post-editing

Jun-ho Lee & Soon-mi Kim
(Hankuk University of Foreign Studies & Sookmyung Women's University)

This research investigates the nature of English to Korean full post-editing and what factors can affect the productivity of full post-editing. The first goal is to examine if machine translation post-editing is faster than human translation and if the adequacy and fluency are not inferior to human translation. The data collected from 14 students showed that full post-editing is 69.66% faster than human translation. A sampling evaluation found that post-editing quality is not inferior to human translation. Furthermore, a statistical analysis found that editing volume does not correlate closely with productivity. This analysis implies that editing volume is one of the core factors that can explain productivity, but there can be other factors. In addition, a text analysis found that the frequency and severity of errors in raw machine translation are directly related to productivity losses. This paper discusses the reasons behind these findings and suggests how the industry and academia should use these findings.

Keywords: machine translation, post-editing, full post-editing, light post-editing, TER
주제어: 기계번역, 포스트에디팅, 풀 포스트에디팅, 라이트 포스트에디팅, TER

이준호(1저자)
한국외대 EICC학과 강사
cuefit@gmail.com
관심 분야: 기계번역, 포스트에디팅, 번역교육

김순미(공동저자)
숙명여대 영어영문학부 부교수
smikim@sookmyung.ac.kr
관심 분야: 기계번역, 포스트에디팅, 과학기술 발전과 통번역, 문화콘텐츠 번역

논문 투고: 2022년 11월 6일
1차 심사 완료: 2022년 12월 9일
2차 심사 완료: 2022년 12월 17일
게재 확정: 2022년 12월 24일