번역학연구, 25(2), 11-37 (2024년 6월 30일)
https://doi.org/10.15749/jts.2024.25.2.001

챗GPT, 파파고, 인간 번역가 간의 한영 문학번역 차이점 연구*

이창수(한국외국어대학교)

1. 서론

2016년에 구글이 내놓은 신경망번역(NMT)은 기계번역의 품질을 획기적으로 개선하며 기계번역 상용화의 시대를 열었다(Wu et al., 2016). 이를 계기로 번역학계에서는 딥러닝에 기초한 기계번역이 최고 고난도 번역으로 간주하는 문학번역을 할 수 있을까 하는 문제에 관심을 두기 시작했다(Rothwell et al,, 2023). 연구 결과를 보면 문학번역에서도 NMT가 기존 기계번역기와 비교하여 번역 품질이 크게 향상되었지만 직역식 기계번역의 특징에 따른 번역 오류와 창의성 미비로 출판 가능한 품질에는 크게 못 미친다는 의견이 우세하다(cf. Toral & Way, 2018). 이런 배경에서 NMT가 번역한 결과물을 인간 번역가나 에디터가 수정하는 MTPE 방식에 관심을 갖게 되었다. 그러나 이 방식조차도 기계번역의 결과물에 기초하기 때문에 번역가의 창의성을 제약한다. 따라서 전문 문학번역가들은 MTPE 방식보다는 처음부터 독자적으로 번역하는 것을 선호한다는 부정적 평가가 많았다(Guerberof-Arenas & Toral, 2022; Kolb, 2023; Moorkens et al., 2018).

이 같은 흐름에 큰 변화를 일으킨 것이 챗GPT의 등장이다. 챗GPT가 기

* 본 연구는 2024년도 한국외국어대학교 교내연구비 지원을 받아 작성되었음.

계번역을 주목적으로 한 기술은 아니지만 프롬프트에 따라 번역 결과물이 달라지며 맥락을 반영한 결과물을 내놓는 등 NMT와는 다른 번역 방식을 제공한다. 현재까지 챗GPT의 번역 능력에 대한 연구를 보면 일반 텍스트의 경우 번역 품질이 MNT에 유사한 수준이거나 더 우수하며, 특히 번역 수요가 많은 언어 쌍의 경우 번역 품질이 크게 향상된 결과를 보여준다(Hendy et al., 2023; Jiang & Zhang, 2024). 반면에 문학번역의 경우 챗GPT의 번역 결과물도 여전히 여러 문제를 갖고 있어 인간 번역가의 추가 수정이 필요하다는 의견이 우세하다(Karabayeva & Kalizhanova, 2024; Sanz-Valdivieso & López-Arroyo, 2023). 그러나 현시점에서 챗GPT의 문학번역 능력에 관한 연구는 매우 제한적이기 때문에 다양한 언어 쌍과 장르에서 연구 성과를 쌓는 것이 필요하다.

이런 배경에서 본 연구는 한영 문학번역에서 인간번역, NMT, 챗GPT 간에 최빈도 어휘 사용 면에서의 문체 및 창의성과 번역 오류 면에서의 차이점을 정량 분석하는 것을 목표로 설정하였다. 분석데이터로는 황석영의『삼포 가는 길』단편 소설을 2명의 인간 번역가가 번역한 결과물, NMT의 경우는 한글에 특화된 파파고의 번역 결과물, 그리고 문학이라는 장르 프롬프트를 주고 얻어낸 챗GPT의 결과물을 사용하였다. 분석은 2단계로 진행하였다. 첫 단계에서는 코퍼스 내의 최빈도 어휘 80개를 사용하여 주성분분석을 통해 문체 차이를 분석했다. 본 연구에서 '문체'는 최빈도 어휘 분포를 의미하며, 최빈도 어휘는 전산문체학에서 텍스트 저자나 종류를 구분하는 데 가장 널리 쓰이는 언어 표지다. 둘째 분석 단계에서는 소설 초반부에 대하여 수작업 내용 분석을 통해 창의성과 오류 분석을 실시하였다. 이를 통해 최빈도 어휘 사용 면에서 챗GPT가 인간번역과 NMT중 어느 쪽에 더 가까운지, 창의성과 오류 면에서 챗GPT가 NMT와 인간번역에 어떻게 비교되는지 등을 밝히고자 한다.

2. 선행연구 고찰

2016년 구글 신경망 기계번역(GNMT)의 등장은 기계번역의 획기적 전환

점이었다. 해당 시스템을 개발한 구글 팀이 내놓은 논문에서(Wu et al., 2016) 저자들은 GNMT가 구글의 기존 구문 기반 기계번역기와 비교해 번역 오류가 60퍼센트 감소했고 일부 문장에서는 인간 번역가와 동등한 품질을 달성했다고 보고했다. 이 같은 GNMT의 등장으로 기계번역에 대한 관심이 폭증하였고, 그런 관심은 품질이 크게 향상된 NMT가 문학번역까지 수행할 수 있을까 하는 질문으로 이어졌다. 문학번역은 주제나 개인적 취향에 따라 변하는 작가의 문체, 문화 같은 텍스트 외적 지식을 요하는 텍스트 해석 등 여러 요소 때문에 번역 형식 중 가장 고난도의 번역 작업으로 간주한다 (Sanchez, 2009, p. 134; Pârlo, 2019). 따라서 문학번역은 기계번역이 인간 번역가에 버금할 수 있느냐는 질문에 대한 최종의 시금석 같은 의미가 있다.

NMT의 문학번역 능력에 대하여 개발자들은 과거 기계번역기와 비교하여 크게 향상된 품질을 제시하면서 잠재력을 강조하지만, 번역학자나 실제 번역가들 사이에서는 부정적 평가가 주류를 이룬다. 이런 부정적 관점은 NMT 문학번역에 오류가 많아서 인간 번역가나 에디터의 개입과 수정이 필요하다는 차원을 넘어선다. 보다 근본적으로 에디터의 손을 거치건 말건 기계번역에 기초한 번역은 문학번역의 핵심인 창의성과 융통성을 제약한다는 지적이 있다. Guerberof-Arenas와 Toral(2022)은 영어 단편 소설을 카탈리아와 네덜란드어로 번역하는 실험에서 인간번역, 문학 데이터로 훈련된 NMT, MTPE 등 세 가지 방식의 번역 결과물을 비교하고 평가자들에게 창의성 점수를 매기도록 하였다. 그 결과 인간번역의 창의성 점수가 가장 높았고, MTPE, NMT 순이었다. 그러나 MTPE에서도 번역가의 창의성이 크게 제약되어 출판에 적합한 수준의 품질에 도달하지 못했다고 평가했다. 이는 문맥에 적합한 다양한 번역 솔루션을 생각해 낼 수 있는 창의력이 MT와 인간번역을 구분하는 핵심 요소임을 보여준다. Kolb(2023)도 인간 문학번역과 MTPE 문학번역 과정을 비교하였는데 MTPE 결과물은 인간번역 결과물에 비하여 번역물 간 유사성이 더 높게 나타났다. 즉, 인간번역에 비하여 기계번역에 기초한 문학번역은 선택의 다양성을 제약하는 결과를 가져온다. 국내에서도 비슷한 연구가 있었는데 마승혜(2018)는 한강 저『채식주의자』의 Deborah Smith 영역본과 GNMT의 영역본을 비교한 결과 기계번역은 문맥에 바탕을 둔 해석, 창의적 재현, 독자 소통 능력에서 인간번역에 크게 못 미친

것으로 평가했다. 문학번역에서 기계번역과 인간번역의 차이는 컴퓨터 어휘 분석에 기초한 문체에서도 드러났다. 이창수(2019, 2021)는 한국어 장·단편 소설 28편을 대상으로 3가지 온라인 신경망 기계번역기의 번역 결과물과 인간번역 결과물을 선형판별분석(LDA), 주성분분석(PCA), 랜덤포리스트(RF) 분석을 통해 최빈도 어휘 분포를 조사했다. 2019년과 2020년 두 차례 같은 분석을 실시하였는데 양 연도 분석에서 인간번역과 기계번역은 서로 뚜렷한 군집을 형성하며 거리를 두어 문체에서 확실한 차이가 나타났다. 또한 1년 사이에 기계번역기 간의 거리가 좁혀져 시간이 지나면서 기계번역기들의 문체가 점점 더 균질화되고 있음을 알 수 있었다. 결과적으로 NMT가 아무리 발전한다고 해도 문장 단위 직역에 기초한 알고리즘으로는 문맥을 반영한 창의적 번역을 핵심으로 하는 문학번역의 요건을 충족하기 어려워 보인다.

그런데 2022년 11월에 발표된 챗GPT는 문학번역에서 인간번역과 기계번역이 명확히 구분되던 구도에 변화를 예고한다. 기존 NMT와 달리 챗GPT는 프롬프트에 맥락 정보를 줄 경우 이를 반영해서 번역한다(류친, 2024). 따라서 인간 평가자들에게서 NMT보다 번역 품질이 눈에 띄게 좋다는 평가를 받고 있다(Jiang & Zhang, 2024). 그러나 번역 언어 쌍에 따라 번역 품질이 크게 차이가 나서 훈련 데이터가 많은 언어 쌍에서는 경쟁력 있는 번역 품질을 보이지만 그렇지 않은 언어 쌍의 경우 품질이 크게 떨어진다(Hendy et al., 2023). 또한 문장 단위 NMT와 비교하여 문단이나 담화 차원에서 맥락을 반영할 수 있는 능력을 갖춘 챗GPT의 경우 번역이나 문법 오류가 훨씬 적고 문체의 일관성도 개선된 결과를 보여준다. 그와 동시에 번역을 생략하는 등 치명적 약점도 있다(Karpinska & Iyyer, 2023). 또한 거대언어모델에 기초한 번역시스템의 경우 프롬프트에 따라 번역 결과가 달라지기 때문에 적절한 프롬프트 전략의 중요성이 강조되기도 한다(Zhao et al., 2023). 챗GPT 같은 생성형 AI의 문학번역 능력에 대한 연구는 초기 단계로 현재까지의 연구 결과만으로 어떤 결론을 내리기 어렵다. 따라서 다양한 언어 쌍과 문학 장르를 대상으로 지속적인 연구가 이뤄질 필요가 있다.

3. 연구 방법

분석데이터로는 황석영의 단편 소설『삼포 가는 길』의 인간번역과 기계번역 결과물을 사용하였다. 인간번역은 2008년에 김다희(Kim Dahee)와 2012년에 김우창(Kim U-Chang)이 'The Road to Sampo'란 동일 제목을 영역한 2개의 번역본을 사용하였고 기계번역은 2024년 1월 말에 파파고와 챗GPT에서 얻어낸 결과물을 사용하였다. 챗GPT의 경우는 "The following is a Korean literary text. Translate it into English with a literary flair(다음은 한국 문학 텍스트로 문학적 풍미가 있도록 영역하라)"란 프롬프트를 주어 번역을 시행하였다.

『삼포 가는 길』은 상황 묘사가 세부적이고 구체적이어서 번역물을 비교평가하는데 용이하며, 대화에서는 한국어 특유의 표현과 어법이 등장하기 때문에 기계번역이 이런 구문을 제대로 다룰 수 있는지를 시험하는 데 적합한 텍스트라고 판단하였다. 번역문의 단어 수를 비교해 보면 김다희 번역본(이하 HT2로 지칭)이 8,097, 김우창 번역본(이하 HT1으로 지칭)이 7,381, 챗GPT가 6,935, 파파고가 6,873 순이었다. 그런데 어휘 다양성의 지표 중 하나인 표준화된 타입/토큰 비율(STTR)을 보면 챗GPT가 49.32, HT1이 45.36, 파파고가 42.7, HT2가 41.56 순이었다. 챗GPT 출현 전에 동일 원문에 대한 인간번역과 기계번역의 문체 비교 연구에서도 STTR 값은 HT1>기계번역>HT2의 순으로 나타났다(이창수, 2019). 챗GPT의 STTR이 인간번역보다 높게 나타난 것은 흥미로운 점이나 현재의 데이터만 놓고 볼 때는 STTR 값은 인간번역, NMT, 챗GPT번역을 구분하는 변별력을 갖고 있지 않다.

데이터 분석은 두 가지 방법을 사용하였다. 처음에는 4종의 번역문 전체를 사용하여 컴퓨터로 주성분분석(PCA)를 통해 최빈도 어휘 분포의 차이를 분석하였다. 앞서 본 연구에서 '문체'는 최빈도 어휘의 분포 패턴을 의미한다고 정의하였는데, 최빈도 어휘는 전산문체학에서 텍스트의 저자나 장르를 구분할 때 가장 흔하게 사용되는 언어 자질이다(Burrows, 2002). 먼저 R 프로그램의 stylo패키지를 사용하여 번역문에서 최빈도 어휘를 추출하였다. 여기에 <그림 1>과 같이 HT1(김우창 2012 번역본), HT2(김다희 2008 번역본), Papago(파파고), ChatGPT(챗GPT)를 구분하는 TR이란 범주 변수와 HT(인간

번역), MT(기계번역)를 구분하는 Group이란 범주 변수를 추가하였다. 이렇게 준비된 데이터를 R의 FactorMinerR 패키지로 불러들여 PCA 분석을 시행하였다. PCA 분석은 샘플 수가 다수여야 하므로 각 번역문을 같은 위치에서 사 등분 하여 총 16개의 분석 샘플을 만들었다. 각 번역문의 샘플은 <그림 1>의 id 항목에서 보듯이 알파벳 a, b, c, d로 구분하였다.

그림 1

PCA 분석데이터

id	TR	Group	the	a	to	and	you	of	i	it	in.	was
Sampo_4_ChatGPT_a	ChatGPT	MT	7.456404	3.006615	2.405292	2.10463	0.721587	1.503307	1.082381	1.383043	1.142514	1.142514
Sampo_4_ChatGPT_b	ChatGPT	MT	7.040762	3.493912	2.170461	2.170461	1.482266	1.64108	0.794071	1.799894	1.323452	0.847009
Sampo_4_ChatGPT_c	ChatGPT	MT	5.450237	2.42891	1.836493	1.954976	1.244076	1.718009	1.658768	0.829384	1.184834	0.473934
Sampo_4_ChatGPT_d	ChatGPT	MT	6.325778	2.583026	2.372167	1.4233	1.212441	1.634159	1.212441	1.265156	1.370585	1.001581
Sampo_4_HT1_a	HT1	HT	6.531882	2.488336	2.851218	2.851218	1.244168	1.244168	1.399689	1.658891	0.933126	1.918092
Sampo_4_HT1_b	HT1	HT	5.243644	2.966102	2.383475	2.701271	2.012712	1.324153	1.430085	1.588983	1.377119	1.271186
Sampo_4_HT1_c	HT1	HT	4.533745	2.472952	2.679031	2.833591	2.163833	1.648635	2.112313	0.618238	1.648635	1.081917
Sampo_4_HT1_d	HT1	HT	5.926321	2.242392	2.722904	2.562734	1.761879	2.295782	1.708489	1.121196	1.494928	1.227977
Sampo_4_HT2_a	HT2	HT	7.439614	2.125604	2.560386	1.497585	1.207729	1.690821	1.304348	0.917874	1.594203	1.594203
Sampo_4_HT2_b	HT2	HT	6.500489	2.492669	2.394917	2.052786	2.150538	1.368524	0.879765	1.564027	1.173021	0.83089
Sampo_4_HT2_c	HT2	HT	5.365355	1.83955	1.992846	2.095043	2.452734	1.83955	1.48186	0.817578	1.277466	1.175268
Sampo_4_HT2_d	HT2	HT	5.702851	2.351176	2.601301	2.101051	1.550775	1.650825	1.350675	0.90045	1.250625	1.2006
Sampo_4_Papago_a	Papago	MT	6.716418	2.985075	3.04248	2.525832	1.607348	1.377727	1.894374	2.06659	1.033295	2.583238
Sampo_4_Papago_b	Papago	MT	7.26979	3.338718	2.154012	2.584814	2.207862	1.938611	2.369413	2.477114	1.615509	1.561659
Sampo_4_Papago_c	Papago	MT	6.453455	3.198172	2.113078	2.912621	1.484866	1.541976	2.45574	1.827527	1.427756	1.656196
Sampo_4_Papago_d	Papago	MT	6.169377	2.411666	2.748177	2.35558	1.682557	1.570387	1.962984	2.019069	2.019069	1.906898

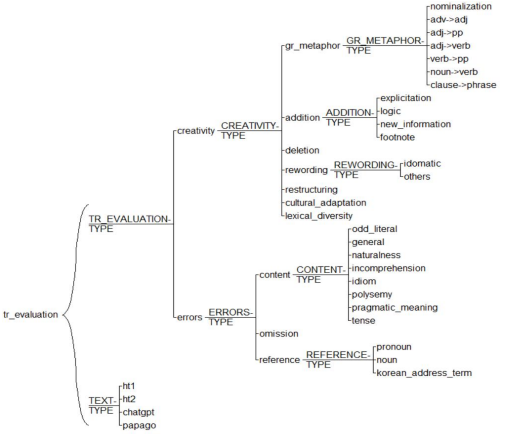
PCA 분석에서는 최빈도 어휘 몇 개를 사용할 것인가를 결정해야 하는데 특별히 정해진 기준은 없고 분석의 목적과 데이터의 종류에 따라 연구자가 결정한다. 최빈도 어휘를 사용한 PCA 저자판별 분석으로 잘 알려진 Burrows(2002)는 50에서 100 사이의 최빈도 어휘 사용을 권장하지만, 분석가들에 따라 더 많은 어휘를 사용하기도 한다. 본 연구에서는 다음과 같은 기준에 따라 최빈도 어휘 80개를 사용하였다. <그림 1>에서 보듯이 최상위 빈도 어휘는 대부분 the, a 같은 관사, to, of 같은 전치사, and 같은 접속사, you, I, it 같은 대명사들이다. 이런 어휘는 기능어라고 불리며 텍스트 내용의 영향을 받지 않는다. 그런데 빈도수가 내려갈수록 school, season 같은 명사나 send, slept 같은 동사 등 내용어가 나타난다. 따라서 텍스트의 내용의 간섭을 최소화한 상태에서 충분한 변수를 사용하여 문체를 분석하기 위해서 상위 80개 단어를 선택했다.

두 번째 분석에서는 번역 품질을 구체적으로 분석하기 위하여 원문의 4분의 1(27%)에 해당하는 번역문을 수작업으로 분류하여 주석을 달았다. 분석에서는 수작업 분류의 효율성과 통계의 정확성을 확보하기 위하여 UAM

코퍼스 툴(UAM Corpus Tool) 2.8 버전을 사용하였다. 분석은 <그림 2>의 주석 달기 분류도에서 보듯이 창의성(creativity)과 번역 오류(errors) 두 범주로 나누고 각 범주에 하위 범주를 두었다. '창의성'이란 용어는 사용하는 사람에 따라 다양한 것을 의미할 수 있다. 번역은 그 자체가 창의적인 작업이기 때문에(Ho, 2024, p. 8) 구체적으로 어떤 요소를 창의성의 발로로 볼 것인가를 결정하는 것은 쉬운 일이 아니다. 그러나 번역에서 창의성은 '고쳐 쓰기(rewriting)'(Merkle, 2010, p. 18)로 정의되기도 하며, 글을 고쳐 쓴다는 것은 기본적으로 원문에서 탈피, 또는 벗어나는 것을 의미한다. Guerberof-Arenas와 Toral(2002)은 번역에서의 '창의성'이란 개념을 설명하며 기본적으로 원문과 다른 새로운 것(novelty)을 창의성으로 정의한다. 이런 의미에서 창의성은 원문과 달라진 표현, 즉 '쉬프트(shift)'로 나타난다. 그런 의미에서 본 연구에서는 오역이 아닌 상태에서 원문의 직역에서 벗어난 경우를 창의성으로 분류하였다.

그림 2

2차 분석 주석달기 분류도



<그림 2>의 분류도는 창의성과 오류라는 두 가지 대범주만 정해 놓고 실제 번역물을 원문과 비교해 가며 별도 하위 범주가 필요하다고 판단될 경우 해당 범주를 실시간으로 분류도에 추가하였다. 따라서 기존 번역 품질평가 모델과 비교하여 빠지거나 추가된 하위 범주가 있을 수 있으며 <그림 2>

의 분류는 총괄적 평가모델이 아니라 본 논문에서 인간-챗GPT-파파고의 번역 결과를 비교하기 위한 목적에 맞춰 만들어진 것임을 밝혀둔다. 세부적인 하위 항목에 대한 논의는 분석 결과를 제시할 때 다루도록 한다.

4. 분석 결과

분석 결과는 먼저 최빈도 어휘 80개를 활용한 PCA 문체 분석 결과를 살펴본 후에 수작업으로 창의성과 번역 오류를 분류한 결과를 논하도록 한다.

4.1 최빈도 80개에 기초한 PCA 문체 분석

<그림 1>의 데이터를 사용하여 PCA 분석을 실시하면 2차원 평면에 분석 텍스트와 어휘를 배치한다. 이때 어휘 분포 상관관계가 높을수록 텍스트들은 가까운 거리에 배치되고 낮을수록 거리가 멀어진다. 따라서 해당 그래프를 보면 시각적으로 어떤 텍스트들이 최빈도 어휘 분포에서 유사하거나 차이가 나는지를 파악할 수 있다.

분석 결과는 <그림 3>, <그림 4>, <그림 5>에 도형으로 제시하였다. 먼저 <그림 3>에서 각 번역문 샘플의 위치를 비교해 보자. 사각형으로 표시된 것이 파파고 번역문인데 4개의 샘플 모두 도형 상단 위쪽에 다른 번역문과 떨어진 상태로 배치되어 있다. 세모와 마름모꼴은 각각 HT2와 HT1을 나타내며 HT2는 중간선 부근에 HT1은 그보다 좀 더 아래쪽에 자리 잡고 있다. HT1의 HT1_a 샘플이 HT2 샘플들 사이에 끼어있는 것을 제외하면 두 번역문도 어느 정도 집단으로 구분이 된다. 마지막으로 육각형으로 표시된 챗GPT는 도형 아래쪽에 배치되어 있는데 ChatGPT_a와 ChatGPT_d가 각각 HT2와 HT1의 군집에 들어가 있다. 즉, 파파고의 번역 샘플은 홀로 떨어져 하나의 군집을 이루지만 챗GPT는 인간번역문과 혼재되어 있다. 즉, 최빈도 어휘 사용에 기초한 문체에서 파파고는 인간번역과 뚜렷이 구분되지만, 챗GPT는 구분되지 않는다.

그림 3

PCA 문체 분석 1

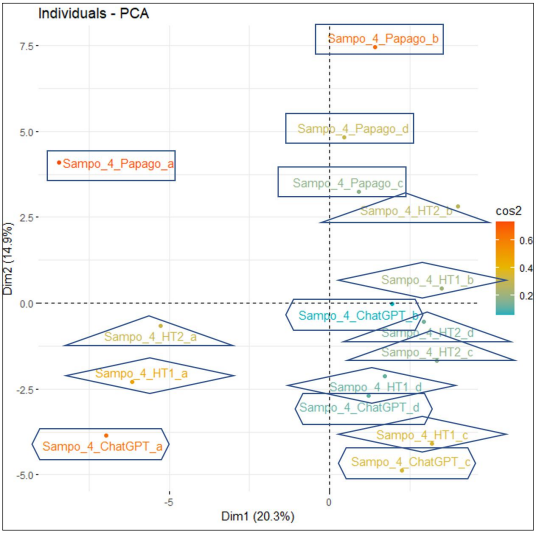
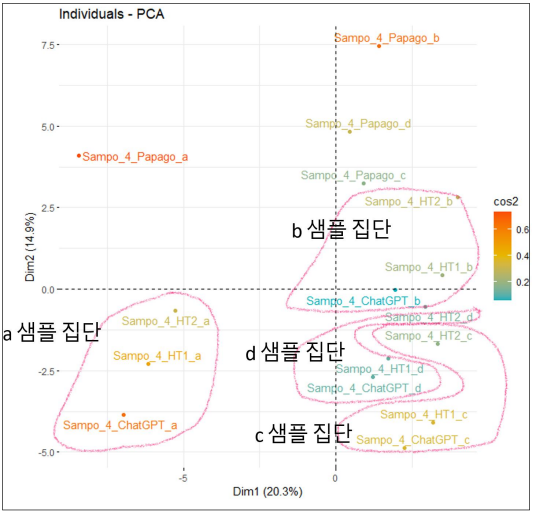


그림 4

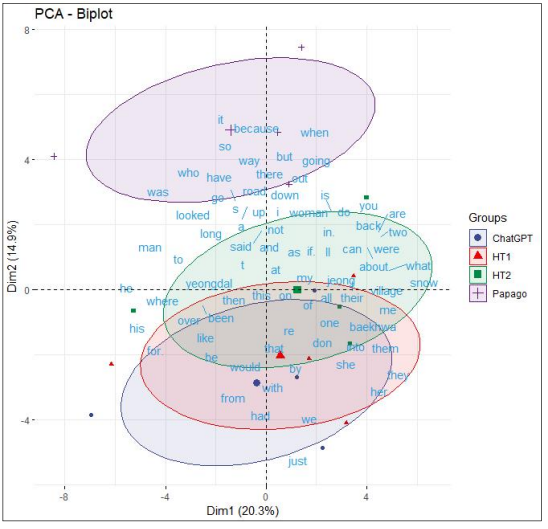
PCA 문체 분석 2



한 가지 흥미로운 점은 <그림 4>에서 보듯이 파파고를 제외하면 나머지 3개의 번역문의 샘플들은 a, b, c, d 분류를 중심으로 모여 있다는 점이다. a, b, c, d는 원문을 사 등분 한 것이고, 샘플들이 이 알바벳을 중심으로 뭉쳐 있다는 것은 원문의 내용에 따라 모여있다는 것을 의미한다. 파파고를 제외한 나머지 3 번역문 간에는 최빈도 어휘 사용에서 통계적으로 유효한 차이가 없다 보니까 이야기 내용을 반영하는 내용어가 부각되어 이야기 흐름에 따라 동일한 번역 내용을 가진 샘플들끼리 모인 결과이다.

그림 5

PCA 문체 분석 3



이를 좀 더 상세하게 분석하기 위하여 <그림 5>의 어휘 배치도를 살펴보자. 도형 내의 원은 번역문별로 어휘 분포의 표준점을 중심으로 그려진 95% 신뢰 타원으로 원이 서로 떨어져 있으면 통계적으로 차이가 있고 겹친 경우에는 유효하지 않다. HT2, HT1, 챗GPT는 원이 상당 부분 서로 겹쳐 있어 통계적으로 유효한 차이가 존재하지 않는다. 이에 반해 파파고 타원은 홀로 떨어져 있어 최빈도 어휘 사용에서 다른 번역문과 분명히 구분된다.

<그림 5>의 어휘를 보면 대부분이 기능어들로, 각 번역문 원안에 들어

있는 어휘는 해당 번역문이 특징적으로 많이 사용한 것들이다. 파파고 원안에 있는 어휘는 사용 빈도에서 다른 번역문과 통계적 차이가 날 만큼 파파고가 특징적으로 사용한 어휘다. 이를 보면 파파고는 because, when, but, so 같은 연결어를 많이 사용한다. 또 '어디에 무엇이 있다'는 구문에 사용되는 there도 파파고에서 두드러진다. 이와 대조적으로 나머지 번역문은 상당수의 어휘를 공유하고 있기 때문에 원이 겹쳐 있다. 특히 HT1과 챗GPT는 대부분의 어휘를 공유하고 있어 둘 사이의 문체가 상당히 유사하다는 것을 알 수 있다. 이에 반하여 HT2는 어느 정도 독자적인 특징어를 갖고 있다. 이는 세 번역문 사이에 통계적 차이는 없지만 HT2보다는 HT2가 챗GPT와 더 유사하다는 것을 보여준다.

특이한 점은 yeongdal(영달), jeong(정씨), baekhwa(백화) 등 주인공의 이름이 아래 세 번역문이 겹치는 부분에 자리 잡고 있다는 점이다. yeongdal은 화면 오른쪽 중심선에 있는데 이곳은 <그림 5>에서 볼 때 샘플 a와 b 집단의 경계선이다. 또 jeong은 샘플 b 집단 내에 있고 baekhwa는 샘플 d와 c 사이에 위치한다. 이것은 HT2, HT1, 챗GPT 사이에 기능어 빈도 차에 따른 통계적 구분이 명확하지 않기 때문에 내용어인 사람 이름이 상대적으로 부각된 결과다. 이는 이야기 전반부부터 등장하는 '영달', 중간에 등장하는 '정씨' 그리고 그 후에 등장하는 '백화'의 흐름을 쫓아가고 있다. 같은 맥락에서 village(마을), snow(눈) 같은 내용어도 서로 겹친 3개 번역문 군집에서 발견된다.

이상의 분석 결과를 종합하면 최빈도 어휘 분포로 살펴본 본 문체에서 NMT인 파파고는 인간번역과 뚜렷한 차이를 보이지만 챗GPT는 구분되지 않는다.

4.2 수작업 분류에 의한 창의성, 번역 오류 분석

이번에는 창의성과 번역 오류라는 품질 면에서 4가지 번역문을 비교해 본다. 연구 방법에서 설명하였듯이 전체 원문의 4분 1(샘플 a 부분에 해당) 분량에 대하여 <그림 2>의 분류도를 기준으로 수작업 코딩 분류를 수행하였다. 여기서는 지면 제약상 주요 분류 항목을 중심으로 논의를 전개하도록 한다. 먼저 <표1>의 창의성 분석 결과를 살펴보자. 표의 맨 위에

CREATIVE-TYPE 행에 총 발생 건수가 나와 있다. HT2가 69개로 가장 많고, HT1이 44개, 챗GPT가 27개, 파파고가 1개 순이다. 그 밑에 세부 분류를 보면 HT1의 경우 gr_metaphor(문법적 은유)가 가장 많고, HT2는 addition(추가)이 압도적으로 많다. 챗GPT는 문법적 은유와 추가가 비슷한 숫자로 나타나고 나머지 항목에서도 인간번역과 비교할 때 비교적 고른 발생 분포를 보여준다.

좀 더 세부적으로 살펴보면 문법적 은유(Halliday & Matthiessen, 1999)의 경우 원문의 동사, 형용사 등을 명사로 바꿔 표현한 nominalization(명사화)이 압도적으로 많다. 특히 HT1이 이런 기법을 특징적으로 많이 쓰고 있다. <예시 1>을 보면 “뒤를 돌아 보다”에서 “보다”를 다른 번역문은 전부 원문처럼 동사를 써서 look으로 했는데 HT1은 glance란 명사를 써서 표현했다. 이 같은 명사화는 문법 은유의 가장 핵심적 요소로 영어 소설에서 자주 쓰이는 수법이다(이창수, 2019, p. 178). 챗GPT도 명사화를 HT2에 버금가는 빈도로 사용했다.

표 1

창의성 분석 결과

	ht1		ht2		chatgpt		papago	
Feature	N	Percent	N	Percent	N	Percent	N	Percent
Total Units	44		69		27		1	
CREATIVITY-TYPE	N=44		N=69		N=27		N=1	
- gr_metaphor	17	38.64%	9	13.04%	8	29.63%	1	100.00%
- addition	9	20.45%	43	62.32%	8	29.63%	0	0.00%
- deletion	3	6.82%	5	7.25%	1	3.70%	0	0.00%
- rewording	8	18.18%	4	5.80%	5	18.52%	0	0.00%
- restructuring	1	2.27%	4	5.80%	3	11.11%	0	0.00%
- cultural_adaptation	0	0.00%	0	0.00%	1	3.70%	0	0.00%
- lexical_diversity	6	13.64%	4	5.80%	1	3.70%	0	0.00%
GR_METAPHOR-TYPE	N=17		N=9		N=8		N=1	
- nominalization	12	70.59%	7	77.78%	8	100.00%	1	100.00%
- adv->adj	1	5.88%	0	0.00%	0	0.00%	0	0.00%
- adj->pp	0	0.00%	1	11.11%	0	0.00%	0	0.00%
- adj->verb	1	5.88%	0	0.00%	0	0.00%	0	0.00%
- verb->pp	1	5.88%	0	0.00%	0	0.00%	0	0.00%
- noun->verb	1	5.88%	0	0.00%	0	0.00%	0	0.00%
- clause->phrase	1	5.88%	1	11.11%	0	0.00%	0	0.00%
ADDITION-TYPE	N=9		N=43		N=8		N=0	
- explicitation	7	77.78%	42	97.67%	4	50.00%	0	0.00%
- logic	1	11.11%	0	0.00%	0	0.00%	0	0.00%
- new_information	1	11.11%	1	2.33%	3	37.50%	0	0.00%
- footnote	0	0.00%	0	0.00%	1	12.50%	0	0.00%
REWORDING-TYPE	N=7		N=4		N=5		N=0	
- idomatic	2	28.57%	1	25.00%	0	0.00%	0	0.00%
- others	5	71.43%	3	75.00%	5	100.00%	0	0.00%

addition 항목은 원문에 명시적으로 없는 정보를 추가하는 경우로 함축된 정보를 명시화하는 explicitaiton과 새로운 정보를 추가한 new_information, 절간 논리적 관계를 추가한 logic, 언어나 문화적 차이를 고려하여 주석 형식으로 설명을 추가한 footnote로 구분하였다.

<예시 1>
원문: 그는 뒤도 돌아보지 않고 질척이는 둑길을 향해 올라갔다.
HT1: Without a backward glance, the guy headed for the muddy path along the dike.

HT2가 explicitation에서 압도적으로 많은 수치를 기록했는데 이는 대부분이 원문의 직접 인용문에 '누가 어떻게 말했다'는 맥락 정보를 추가한 경우이다. <예시 2>에서 다른 번역문은 원문의 직접 인용문을 그대로 전달했는데 HT2는 Yeoungdal said(영달이 말했다)란 인용절과 admiringly(감탄하듯이)라는 태도를 나타내는 부사구를 추가하였다. 이 같은 화법 명시화는 HT1과 HT2의 비율(7 대 42)에서 보듯이 번역자에 따라 선호도가 차이 난다. HT2의 경우는 특별하다고 할 정도로 많이 사용하였다. 주목할 점은 인간 번역가들이 필요시 수행하는 화법 명시화 전략을 챗GPT도 적게나마 사용하고 있다는 점이다.

<예시 2>
원문: "... 정말 든든하시겠구만."
HT2: "... you must feel very secure," Yeongdal said admiringly.

챗GPT는 원문에 없는 정보도 글의 흐름에 따라 필요하다고 판단하면 자의적으로 첨가하는 경향을 보였다. <예시 3>을 보면 원문 어디에도 pretending to be a merchant(상인인체하며)란 내용이 없다. 열차에서 개피떡과 인절미를 팔았다는 맥락에서 그 같은 내용을 끄집어내어 창의적으로 삽입한 것으로 보인다. 그뿐 아니라 '개피떡'을 dog tteok이라고 번역하면서 괄호 안에 Korea rice cake(한국식 쌀떡)이란 문화 정보까지 주석 형식으로 추가하였다. 이 같은 창의적 원문 개입은 HT, HT2에서는 각각 1건이 발견되

었는데 챗GPT는 3건을 기록했다.

<예시 3>
원문: 개피떡이나 인절미를 전방으로 호송되는 군인들께 팔았다는 것인데 딴은 열차를 타며 사내들 틈을 누비던 계집이 살림을 한답시고 들어앉아 절름발이 천가 여편네 노릇을 하려니 따분했을 것이었다.
챗GPT: The woman had apparently sold dog tteok (Korean rice cake) and injeolmi to soldiers being escorted ahead. While others covertly infiltrated between the soldiers on the train, pretending to be a merchant, she set up her home to play the role of Cheon's wife.

마지막으로 rewording은 의미의 유사성은 유지하면서 원문과 뚜렷하게 다른 표현 방식을 사용한 경우인데 여기서도 챗GPT는 인간 번역가에 견줄 수 있는 창의성을 보였다. <예시 4>에서는 HT2와 챗GPT가 rewording을 한 것으로 코딩했는데 HT2는 '잠바 깃 속에 턱이 파묻혀있다'는 말을 '깃이 뺨까지 올라와 있었다'로 번역했다. 챗GPT는 여기서 한발 더 나아가 '재킷이 얼굴 대부분을 숨겼다'로 번역했다. 그리고 '알아볼 수 없었다'를 다른 번역문은 대부분 원문대로 번역했지만, 챗GPT는 '그 결과 턱이 보일락말락 하게 되었다'로 번역하였다. 여기서 주목할 점은 챗GPT의 창의적 개입이 맥락을 이해하고 번역에 반영하는 능력에 기초하고 있다는 점이다. 이는 문장 단위 번역에 제한되어 맥락에 맞는 어휘나 지시어 선택에서 큰 어려움을 겪는 NMT와는 분명히 구별되는 특징이다.

<예시 4>
원문: 검게 물들인 야전 잠바의 깃 속에 턱이 반 남아 파묻혀서 누군지 쌍통을 알아볼 도리가 없었다.
HT2: It was hard to tell who he was, as the turned-up collar of his army field jacket came up to his cheeks.
챗GPT: The collar of his dark-stained field jacket concealed most of his face, leaving only a hint of his chin visible.

그럼에도 챗GPT는 번역에서 많은 오류를 기록했다. <표2>를 보면 챗

GPT는 인간번역보다 몇십 배 많은 번역 오류를 범했고, 파파고는 챗GPT보다 2배에 많은 오류를 기록했다. 가장 많은 오류는 내용을 잘못 번역한 content인데 주목할 점은 파파고는 '황당한' 수준의 직역식 오류(odd_literal)가 15건이 있었던 반면 챗GPT는 1건 밖에 없다는 점이다. 그리고 일반적 내용 오역도 파파고는 52건인데 반하여 챗GPT는 절반 수준인 26건이다.

표 2

번역 오류 분석 결과

	ht1		ht2		chatgpt		papago	
Feature	N	Percent	N	Percent	N	Percent	N	Percent
Total Units	5		10		92		178	
ERRORS-TYPE	N=5		N=10		N=92		N=178	
- content	2	40.00%	3	30.00%	63	68.48%	113	63.48%
- omission	3	60.00%	7	70.00%	7	7.61%	7	3.93%
- reference	0	0.00%	0	0.00%	22	23.91%	58	32.58%
CONTENT-TYPE	N=2		N=3		N=63		N=113	
- odd_literal	0	0.00%	0	0.00%	1	1.59%	15	13.27%
- general	2	100.00%	1	33.33%	26	41.27%	52	46.02%
- naturalness	0	0.00%	0	0.00%	3	4.76%	9	7.96%
- incomprehension	0	0.00%	0	0.00%	12	19.05%	13	11.50%
- idiom	0	0.00%	2	66.67%	12	19.05%	11	9.73%
- polysemy	0	0.00%	0	0.00%	0	0.00%	3	2.65%
- pragmatic_meaning	0	0.00%	0	0.00%	8	12.70%	8	7.08%
- tense	0	0.00%	0	0.00%	1	1.59%	2	1.77%
REFERENCE-TYPE	N=0		N=0		N=22		N=58	
- pronoun	0	0.00%	0	0.00%	13	59.09%	50	86.21%
- noun	0	0.00%	0	0.00%	1	4.55%	1	1.72%
- korean_address_term	0	0.00%	0	0.00%	8	36.36%	7	12.07%

odd_literal의 예를 보면 <예시 5>에서 챗GPT와 파파고는 우리말의 '외진 시골'이란 뜻의 '벽지'를 직역해서 wall(벽)과 wallpaper(벽지)로 번역했다. 사실 이 경우는 기계번역이 한국어 어휘 의미를 이해하지 못해 발생한 오류(incomprehension)나 다의어 번역 오류(polysemy)로 분류할 수도 있다. 이렇듯 내용 오류를 하위 범주로 분류할 때는 범주가 겹치는 경우가 종종 있다. 이 경우 어느 범주에 포함할까 하는 것은 분석가의 주관적 기준과 판단에 의존할 수밖에 없다. <예시 6>에서 보면 "분풀이로 청주댁을 후려 패다"는 부분이 있는데 파파고는 이를 '불(with a fire)로 청주집을 때리다'로 엉뚱하게 번역했다. 이것은 '분풀이로... 패다'에서 '분풀이'를 때리는 도구로 이해하고 밑도 끝도 없는 fire(불)란 단어를 첨가한 결과다. 이에 반하여 챗GPT는 그런 해석의 오류가 없이 무난하게 번역했다. 파파고 번역에서는 이 같은 '황

당한’ 수준의 오류가 상대적으로 빈번하게 발생했다.

<예시 5>
원문: 거긴 벽지나 다름없잖소.
챗GPT: That place is practically a wall,
파파고: That’s like a wallpaper.

<예시 6>
원문: 천가가 분풀이로 청주댁을 후려 패는 동안 방아실에 숨어 있었다.
챗GPT: he hid in the storage room while Cheon vented his frustration by beating his wife.
파파고: ... hid in the chamber while the heavenly family beat the Cheongju house with a fire.

다음에 incomprehension의 경우는 챗GPT와 파파고가 거의 비슷한 오류 숫자를 보였는데 대부분 동일 원문을 이해하지 못한 경우였다. <예시 7>에서 영달이 담배꽁초를 건네며 ‘버리슈’라고 한 말을 인간 번역가들은 ‘나는 필요 없으니 쓰고 버리라’는 의미로 정확하게 번역했지만, 챗GPT와 파파고는 ‘-하슈’란 사투리 어투를 이해하지 못해 원문의 의도와 달리 번역했다. 그런데 여기서도 파파고는 이해 못 한 상태에서 포기하고 영어로 음차해서 전달했지만, 챗GPT는 무엇인가를 건네는 맥락을 고려하여 다른 사람에게 물건을 건넬 때 흔히 쓰는 말로 대체하여 글의 흐름을 살렸다. 챗GPT가 문장의 의미를 해석할 때 문맥을 고려한다는 사실을 뒷받침하는 예이다.

<예시 7>
원문: “버리슈.” 담배 꽁초를 건네주며 영달이가 퉁명스럽게 말했다.
챗GPT: “Here you go”
파파고: “Burish,”

idiom(관용어)의 경우는 ‘발랑 까졌다’, ‘사내 재미를 보다’, ‘말뚝 박고 살다’와 같은 우리말 관용표현을 오역한 경우로 챗GPT와 파파고에서 유사

한 오류 건수가 나타났다. 다만 이 경우도 <예시 8>에서 보듯이 “발랑 까졌다”에서 파파고는 ‘발’ ‘까지다’를 각각 직역해서 feet(발이) peeled(껍질이 벗겨지다)로 ‘엉뚱하게’ 오역했지만, 챗GPT는 글의 흐름으로는 이상하게 읽히지 않는 gone crazy(미쳤다)로 번역했다. 이는 챗GPT가 이해하지 못한 부분은 글의 맥락을 고려하여 자연스럽게 녹아들 수 있는 말로 메워 넣는 능력이 있음을 보여준다.

<예시 8>
원문: 모두들 발랑 까졌다고 하지만서두.
챗GPT: Everyone says she’s gone crazy, but still...
파파고: Everyone’s got their feet peeled, but hurry.

<예시 9>
원문: 사내가 목장갑 낀 손으로 코 밑을 쓱 훔쳐냈다.
챗GPT: running his gloved hand under his nose.
파파고: The man stole under his nose with his glove-gloved hand.

polysemy(다의어)는 파파고에서만 오류가 발생했다. <예시 9>를 보면 원문의 ‘훔치다’는 우리말은 ‘물건을 훔친다’는 뜻과 ‘닦다’는 의미가 있는데 파파고는 stole(물건을 훔쳤다)로 오역했지만 챗GPT는 ‘코밑에서 물건을 훔치는 것’은 닦는 행위일 것이라는 논리적 판단에 따라 동사 run으로 정확하게 번역했다. 이는 문장 내 어휘 선택에서 파파고는 단어 대 단어의 기계적인 일대일 치환에 의존하는 반면 챗GPT는 문장 내 다른 단어들과의 통합관계를 파악하여 이치에 맞는 단어를 선택하는 능력이 있음을 보여준다.

pragmatic_meaning은 우리말 표현이 직역과 다른 화용적 또는 함축된 의미가 있을 때 그 의미를 해석하지 못하고 직역한 경우로 챗GPT와 파파고 둘 다 이 부분에서 취약함을 드러냈다. <예시 10>에서 ‘인사가 늦었다’는 우리말 표현은 ‘인사하는 행동’을 ‘늦게 했다’는 직역 의미가 아니라 ‘미처 자신의 소개를 하지 못했다’는 화용적 의미로 쓰인다. 인간번역에선 이런 화용적 의미가 정확히 표현됐지만 챗GPT와 파파고는 우리말을 직역해서 오류가 발생했다. 특히 <예시 11>은 원문에 암시된 의미를 파악하는데 컴퓨터

언어 모델이 어려움을 겪는 것을 잘 보여준다. 원문에서 파파고와 챗GPT는 ‘현장에서 잡히다’를 직역했다. 그런데 글의 맥락상 그 현장은 천가 아내와 몰래 잠자리를 하다 들킨 상황으로 HT1은 caught the two in the act(두 명이 그 짓을 하는 것을 잡다)란 관용표현을 써서, HT2는 좀 더 명시적으로 catch him in bed with his wife(아내와 잠자리에 있는 그를 잡다)라고 정확히 전달했지만, 파파고뿐만 아니라 챗GPT도 ‘건설 현장에서 잡혔다’고 오역했다.

<예시 10>
원문: 인사 늦었네요.
챗GPT: You greeted me late. I’m No
파파고: You’re late to say hello.

<예시 11>
원문: 역에 나갔던 천가 놈이 예상 외로 이른 시각인 다섯 시쯤 돌아왔고 현장에서 덜미를 잡혔던 것이었다.
챗GPT: [...] He got caught up at the construction site.
파파고: the man [...] was caught at the scene.

idiom, polysemy, pragmatic_meaning 등의 범주에서 파파고와 챗GPT가 비슷한 수준의 오류를 보인 것은 컴퓨터 언어 모델은 문학 언어의 특징인 관용어, 다의어, 모호성, 유머, 은유 표현, 긴 문장 등을 처리하는 데 어려움을 겪는다는 기존 관점과 일치한다(Youdale, 2020, p. 23).

이외에 기타 내용 오류는 general로 분류하였는데 파파고 보다는 적지만 챗GPT도 여전히 많은 번역 오류를 기록했다. 그중에는 원문과 정반대로 번역한 경우도 여럿이다. <예시 12>를 보면 원문에서는 ‘몇 걸음 남겨 놓고 섰다’고 했는데 챗GPT는 반대로 ‘몇 걸음 갔다’고 번역했다. 파파고도 ‘가다가 선’ 맥락을 반영하지 않고 단순히 ‘서 있다’고 번역해서 오류로 코딩되었지만, 정도를 따지면 챗GPT의 오류가 더 치명적이다.

<예시 12>
원문: 그는 몇 걸음 남겨 놓고 서더니 털모자의 챙을 이마빡에 붙도록 척 올리면서

챗GPT: After taking a few steps and adjusting the brim of his fur hat to cover his forehead,
파파고: He stood a few steps away and spoke, chucking up the brim of his woolly hat to his forehead.

마지막으로 기계번역의 큰 취약점 중 하나인 지시어 선택의 오류를 분석해 보자. 기계번역은 앞에 나온 명사를 지칭하는 대명사 선택에서 오류를 범하는 경우가 많다(Wehrli & Nerima, 2013). 이 같은 조응어 관계 오류는 주어가 흔히 생략되는 한국에서는 더 큰 문제가 된다. 주어가 없는 한국어 문장의 경우 영어로 번역할 때는 필수적으로 주어를 넣어야 하므로 번역가는 문맥을 고려하여 누가 행동이나 말의 주체이고 대상인가를 파악해야 한다. 문장 단위 번역에 제한된 기계번역에서는 글로 명시화되지 않은 지시어 연결 고리(reference chain)를 파악하는 것이 큰 난제다. 이는 본 연구의 오류 분석에서 파파고가 50건의 대명사(pronoun) 오류를 범한 것만 봐도 알 수 있다. 챗GPT도 이런 오류에서 완전히 벗어나지 않지만, 발생 건수가 13개로 훨씬 적다. 이것은 챗GPT가 맥락을 고려하여 대명사를 선택하는 능력이 제한적으로나마 작동하기 때문이다.

<예시 13>을 보면 우리말 문장의 주어 두 개가 생략되어 있다. 문맥은 술집 아낙 청주댁에 관한 이야기로 인간번역에선 정확하게 주어 자리에 여성 일인칭 대명사인 she를 넣어 번역했다. 그런데 챗GPT와 파파고는 첫 문장 번역에서 이런 맥락을 놓치고 주어로 it과 that을 선택했다. 그런데 다음 문장에서는 챗GPT는 앞 문장에 woman(여성)이란 단어가 나온 것을 고려하여 문장의 주어가 이 여성일 것이라는 공지시 관계(coreference)를 정확히 파악해서 she를 주어로 사용했다. 그러나 파파고는 밑도 끝도 없이 일인칭 대명사 I를 넣어 번역했다. 파파고는 이런 식으로 he나 she를 넣어야 할 곳에 무작위로 you나 I를 넣어 번역한 예가 많다. 그에 반해 챗GPT에서는 그 같은 오류가 상대적으로 훨씬 적다.

<예시 13>
원문: “여자야 그만이었죠. 처녀 적에 군용차두 탔답니다.”
챗GPT: “It was enough with the woman. She even rode military vehicles

during her maiden days. [...]"
파파고: "That was enough, girl. I also rode a military car when I was a girl."

<예시 14>
원문: 그런데 노형은 어디로 가쇼?
챗GPT: But where is Nohyung headed?
파파고: But where are you going, bro?

그러나 우리말 특유의 호칭(korean_address_term)을 번역하는 데서는 챗GPT와 파파고가 비슷한 어려움을 겪었다. <예시 14>에서 상대방을 가르치는 호칭인 '노형'을 챗GPT는 이름으로 해석해서 Nohyung으로 음차했지만, 파파고는 you로 정확하게 번역했다. 그러나 반대로 앞서 <예시 6>에서는 '천가' '청주댁'이란 호칭을 파파고는 직역해서 heavenly family(하늘의 가족), the Cheongju house(청주 집)로 오역했지만, 챗GPT는 정확하게 번역했다. 특히 청주댁이 '천가'의 아내라는 관계가 이전 문맥에 나온 것을 반영하여 인간번역 HT1과 HT2처럼 his wife(그의 아내)로 번역한 것이 인상적이다.

5. 결론

이상의 분석 결과를 종합해 보면 다음과 같다. 인간번역과 NMT 계열인 파파고 번역 간에는 최빈도 어휘 사용 패턴에 기초한 문체나 번역의 창의성과 오류 면에서 현격한 차이가 존재한다. 파파고는 문체 면에서 인간번역과 뚜렷한 차이를 보였으며 창의력은 미미하지만, 수많은 번역 오류를 기록했다. 특히 문맥이나 이치로 '황당하다'고 할 수 있는 오류가 빈번했다. 이에 반하여 생성형 AI인 챗GPT는 문체 면에서 인간번역과 통계적으로 구분되지 않았으며 제한적으로나마 인간번역에서 나타나는 창의적 개입을 수행하는 능력을 보여주었다. 오류 건수도 파파고에 비하여 적었을 뿐만 아니라 '황당한' 수준의 오류가 극히 적었다. 파파고와 챗GPT 간의 이 같은 차이는 기본적으로 맥락 이해 및 활용 능력의 차이에서 비롯된다. 분석 결과를 보면

파파고의 번역은 문장 단위 자구적 번역에 절대적으로 의존하고 있다. 이에 반하여 챗GPT는 문맥 정보를 활용하여 오류 건수를 크게 줄이고 번역문에 창의적으로 개입하기도 했다. 그럼에도 챗GPT 번역은 의미를 정반대로 전달하는 치명적 오류를 포함하여 많은 오류를 기록했다. 또한 한국어 특유의 호칭, 관용구, 사투리, 화용적 의미 등을 처리하는 데는 파파고와 비슷한 어려움을 보였다. 이런 상황을 종합적으로 고려한다면 현시점에서 챗GPT의 위치는 인간번역과 기계번역 중간에 자리 잡고 있다고 볼 수 있다.

앞으로 챗GPT가 문학 데이터를 지속적 학습한다면 오류를 상당 수준 줄일 가능성이 있다. 그렇다면 챗GPT를 포함한 컴퓨터 번역과 인간번역을 구분하는 핵심은 번역 오류가 아니라 창의성이 될 가능성이 높다. 본 연구에서 챗GPT가 어느 정도 창의성을 보여주었지만, 이해 못 할 경우 자의적으로 말을 만들어 넣는 등 문제도 드러냈다. 번역에서 '창의성' 개념은 번역이 원문의 파생적 텍스트가 아니라 번역가의 창의적 인지 작업의 산물이란 관점에서 제기되었다(O'Sullivan, 2013). Malmkjær(2020, p. 29)는 번역에서의 창의성을 사모아 섬 댄서들이 전통춤의 기본 스텝과 순서는 유지하면서 자신의 스타일을 창조해 내는 것에 비유한다. 즉, 번역은 사모아 댄서들과 전통춤의 관계처럼 원문과의 관계 속에서 번역가가 자신의 창의성을 실현하는 과정이다. Scott(2018, p. 48)은 원문과 번역문의 관계를 논하면서 번역은 "ST를 새로운 기표(signifier)로 재구성해서 ST를 새로운 존재로 투영하는 작업"으로 정의하였다.

이런 관점에서 원문과 번역문의 관계는 번역의 목적, 번역가의 취향, 스타일에 따라 단단할 수도 있고 느슨할 수도 있다. 때로는 번역문이 원문에 매인 듯 보일 수도 있고 반대로 원문에서 매우 자유로울 수도 있다(Malmkjær, 2020, pp. 29-30). 따라서 창의성 영역은 열린 공간이다. 이 공간에서 인간 번역가는 문화, 저자, 독자, 글의 내용, 맥락, 등장인물의 성격, 상황 등 수많은 복합적 배경 정보를 활용하여 번역문에 어떤 식으로 어느 정도의 창의적 개입을 할지를 결정한다. 이 과정은 주체적이고 의식적이다. 챗GPT가 아무리 인간번역의 창의성을 흉내 낸다고 해도 인간만 가지고 있는 주체성과 의식이 없는 한 과정이 다를 수밖에 없고 따라서 결과물도 다를 수밖에 없다. Kaindl(2021, p. 20)의 말대로 인간 번역가는 '중도적 중재자가

아니라 사회 정치적 맥락에서 의식적 결정을 하는 주체'이기 때문이다.

끝으로 본 연구는 다음과 같은 한계점을 갖고 있다. 먼저, 본 연구는 한 종의 문학 작품에 대한 인간, 챗GPT, 파파고의 번역문을 비교한 것으로 연구 결과를 일반화하기 어렵다. 이를 검증하기 위해서는 유사한 추가 연구가 뒤따라야 할 것이다. 또한, 챗GPT는 프롬프트의 내용에 따라 다른 결과물을 내기 때문에 본 연구에서 사용한 것과 다른 프롬프트를 주었을 경우 다른 번역 결과물을 산출했을 가능성이 있다. 따라서 챗GPT의 번역 수행 능력을 종합적으로 평가하려면 다양한 프롬프트에 따른 결과물을 분석에 포함할 필요가 있다. 다만, 그 같은 분석은 인간-챗GPT-파파고 3종 비교 분석이라는 연구의 초점을 흐리고 연구 범위를 지나치게 확장할 가능성이 있어 본 연구에서는 시도하지 않았다. 마지막으로 본 연구에서 시도한 분석을 비문학 텍스트에 적용했을 때도 유사한 결과가 나올지를 알아보기 위한 후속 연구도 필요해 보인다.

참고문헌

<1차 자료>

Hwang, S-Y. (2008). *The road to sampo*. (D. Kim Trans.), Avil.

Hwang, S-Y. (2012). *The road to sampo* (U-C. Kim Trans.), Asia Publishers.

<2차 자료>

류친. (2024). GPT를 활용한 중-한 문학번역에 대한 고찰: 장편 소설 "인생(活著)"의 관용구 번역을 중심으로. 문화와융합, 46(1), 273-287.

마승혜. (2018). 문학작품 기계번역의 한계에 대한 상세 고찰. 통번역학연구, 22(3), 65-88.

이창수. (2019). 문학번역에서의 기계번역과 인간번역 문체에 대한 전산문체학적 비교 연구. 번역학연구, 20(2), 111-130.

이창수. (2021). 한영문학번역에서의 번역문체 연구. 통역과 번역, 14(2), 173-195.

Burrows, J. F. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, *17*(3), 267–87.

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

Halliday, M.A.K. & Matthiessen, C. (1999). *Constructing experience through meaning: A language-based approach to cognitio*n. Continuum.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify M., & Awadalla, H. H. (2023). *How good are GPT models at machine translation? A comprehensive evaluation.* arXiv preprint arXiv:2302.09210.

Ho, N-K. M. (2024). *Appraisal and the transcreation of marketing texts: Persuasion in Chinese and English*. Routledge.

Jiang, Z., & Zhang, Z. (2024). *Can ChatGPT rival neural machine translation? A comparative study*. arXiv preprint arXiv:2401.05176.

Kaindl, K. (2021). (Literary) Translator studies: Shaping the field. In K. Kaindl, W. Kolb & D. Schlager (Eds.), *Literary translator studies* (pp. 1-38). John Benjamins..

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

Karpinska, M., & Iyyer, M. (2023). *Language models effectively leverage document-level context for literary translation, but critical errors persit.* arXiv preprint arXiv:2304.03245.

Kolb, W. (2023). 'I Am a bit surprised': Literary translation and post-editing processes compared. In A. Rothwell, A. Way & R. Youdale (Eds.), *Computer-assisted literary translation* (pp. 53-68). Routledge.

Malmkjær, K. (2020). *Translation and creativity*. Routledge.

Merkle, D. (2010). Censorship. In Y. Gambier & L. V. Doorslaer (Eds.), *Handbook of translation studies* (Vol. 1, pp. 18-21). John Benjamins.

Moorkens, J., Toral, A., Castilho, S., & Way, A. (2018). Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, *7*(2), 240-262.

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

Pârlo, A-C. (2019). *Intersemiotic translation: Literary and linguistic multimodality.* Palgrave.

Rothwell, A., Way, A., & Youdale, R. (Eds.). (2023). *Computer-assisted literary translation*. Routledge.

Sanchez, M. T. (2009). *The Problems of literary translation: A study of the theory and practice of translation from English into Spanish*. Peter Lang.

Sanz-Valdivieso, L., & López-Arroyo, B. (2023). Google translate vs. ChatGPT: Can non-language professionals trust them for specialized translation?. In *International conference human-informed translation and interpreting technology* HiT-IT, 2023, 97-107.

Scott, C. (2018). *The work of literary translation.* Cambridge University Press.

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.

Wehrli, E., & Nerima, L. (2013). Anaphora resolution, collocations and translation. In J. Monti, R. Mitkov, G. C. Pastor & Seretan, V. (Eds), *Proceedings of the workshop on multi-word units in machine translation and translation technology* (pp. 12-17).

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., ŁKaiser, u., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., ... Dean, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation.* arXiv preprint arXiv:1609.08144.

Youdale, R. (2020). *Using computers in the translation of literary style: challenges*

and opportunities. Routledge.

Zhao, A., Huang, K., Yu, H., & Huang, D. (2023). DUTNLP system for the WMT2023 discourse-level literary translation. *Proceedings of the eighth conference on machine translation,* 296-301.

Differences in Korean-English literar translation among ChatGPT, Papago, and human translators

Chang-soo Lee (soolee@hanmail.net)
Hankuk University of Foreign Studies

Abstract

The present study explores stylistic and quality differences in English translations of a Korean short story by ChatGPT, Papago, and two human translators. Results of two types of quantitative analysis are reported. A PCA analysis showed that, in the distribution of 80 topmost frequent words, Papago was clearly distinguished from human translators while no such distinction was found vis-a-vis ChatGPT. In a follow-up manually-annotated creativity and quality analysis, Papago was again far from measuring up to human translators with near-zero creativity and numerous errors. ChatGPT also made many errors (about half of Papago's) but exhibited a measure of creativity of the kinds seen in human translation. ChatGPT's display of creativity, along with a significant reduction in errors compared to Papago, is attributed to ChatGPT's ability to consider the context in comprehending the source text and transferring the meaning into the target language.

Ke words: ChatGPT; NMT; human translation; literary translation; quantitative analysis

키워드: 챗GPT, NMT, 인간번역, 문학번역, 정량분석

이창수
한국외국어대학교 교수
soolee@hanmail.net

논문 투고: 2024년 4월 4일
1차 심사 완료: 2024년 5월 30일
2차 심사 완료: 2024년 6월 9일
게재 확정: 2024년 6월 15일