https://doi.org/10.15749/jts.2023.24.3.017

번역학연구
24권 3호
(2023.9.30.)

챗GPT 출현 이후 기계 번역과 인간 번역 간의 번역 문체 차이 변화 연구*

이 창 수
(한국외대)

1. 서론

2016년에 구글에서 구글 신경망 기계 번역(GNMT)을 내놓으면서 기계 번역의 품질이 획기적으로 높아졌다(Wu et al. 2016). GNMT가 발표된 이후 한국의 파파고, 마이크로소프트사의 빙 번역기도 신경망 기계 번역 기술을 도입했고, 2017년에는 유럽에서 역시 신경망 기술에 기초한 DeepL이 출범했다. 이러한 흐름에 발맞추어 번역학에서도 기계 번역과 인간 번역의 결과물을 비교하는 연구가 관심을 끌기 시작하였다. 이런 연구는 주로 기계 번역과 인간 번역 사이에는 언어학적으로 분명한 차이가 존재한다는 주장을 뒷받침하는 결과를 내놓았다(cf. Lars 2017; Webster et al. 2020; 이창수 2021; 이현주 2022; 전혜진 2019; 한승희 2020). 그런데 2022년에 생성형 AI 모델인 거대 언어 모델(LLM)에 기초한 챗GPT가 기계 번역 시장에 합류하면서 또 다른 큰 변화를 몰고 왔다. 챗GPT는 기존 신경망 기계 번역에 콘텐츠를 생성할 수 있는 기술을 결합

* 본 연구는 2023년도 한국외국어대학교 교내연구비 지원을 받아 작성되었음.

하여 1차 기계 번역에 이어 2차로 자가 수정까지 할 수 있으며 필요시 창의적 내용까지 첨가한 결과물을 내놓고 있다. 이는 하나의 입력 텍스트에 대하여 하나의 고정된 번역 텍스트만 내놓는 기존 기계 번역기와는 분명히 다른 번역 솔류션이다. 헨디 외(Hendy et al. 2023)는 번역이 많이 이뤄지는 언어 쌍에서는 챗GPT의 번역 품질이 기존 예측형 AI 기계 번역기의 번역 품질에 뒤지지 않는다는 연구 결과를 내놓았다.

챗GPT 번역과 관련하여 주목할 만한 점은 챗GPT가 대화체보다는 문어체를 선호하는 편향성을 갖고 있다는 지적이다. 미트로빅 외(Mitrović et al 2023: 2)는 챗GPT가 답변에서 사용하는 언어가 일반적으로 “격식체적(formal)이고, 공손하고(polite) 비개인적이다(impersonal)”라고 평가하였다. 보르지(Borji 2023: 31)도 “인간은 답변에서 캐주얼하고 친숙한 표현을 쓰는 반면 챗GPT는 비격식체를 피하도록 프로그래밍 되어 답변이 격식체적인 경향이 있다”고 하였다. 세진 외(Cegin et al 2023)는 자연언어 모델을 학습시키는 데 사용되는 데이터의 다양성을 확보하는 방법으로 챗GPT를 사용하여 기존 문장을 바꿔쓰는 패러프레이징(paraphrasing) 실험을 했는데 챗GPT가 생성한 문장이 이례적으로 문어체적인 경우를 보고하였다. 이 같은 관찰은 챗GPT가 생성하는 번역문에서도 문어체적 특징이 두드러질 수 있다는 점을 시사하며, 이런 추정이 사실이라면 기존 기계 번역기의 결과물과 문체적으로 구별되는 요인이 될 수 있다.

본 연구는 이 같은 기계 번역 분야의 최근 변화를 반영하여 다음과 같은 연구 질문에 답하려 한다. 첫째, 인간 번역과 기계 번역은 기존 연구 결과대로 여전히 문체 면에서 뚜렷이 구분되는가? 둘째, 챗GPT 번역물은 문어체적 특징이 강한가? 그리고 이런 특징 때문에 기존 기계 번역기의 결과물과 문체적으로 구분되는가? 셋째, 챗GPT가 1차 번역물을 자가 수정했을 때 결과물은 1차 번역물과 문체적으로 구분되고 인간 번역에 더 가까워지나?

이 같은 질문에 대한 답을 구하기 위하여 중앙일보, 한겨레, 경향신문 사설을 인간 번역사와 4가지 인터넷 기계 번역기(Papago, Google, DeepL, 챗GPT)가 번역한 결과물을 분석 코퍼스로 사용하고 바이버(Biber 1988, 1995)가 레지스터 변이 연구에 사용한 67개의 어휘 및 구문 표지를 분석 자질로 채택하여 다차원 통계분석 기법 중 하나인 주성분 분석(PCA)을 실시하였다.

2. 선행 연구

최근에 기계 번역의 품질이 괄목하게 향상되면서 기계 번역과 인간 번역의 결과물을 비교하는 연구가 많아지고 있다. 기계 번역 개발자 관점에서는 기계 번역이 인간 번역의 품질에 도달했느냐는 논점을 중심으로 연구가 진행되어 온 반면(Läubli et al. 2018; Toral 2020; Toral et al. 2018) 번역학 관점에서는 기계 번역과 인간 번역의 차이점을 규명하는데 더 큰 관심을 보여 왔다. 후자의 경우, 기계 번역의 에러, 어휘 다양도 같은 통계 수치, 컴퓨터를 사용한 문체 차이 분석 등 다양한 관점에서 연구가 이뤄져 왔다. 가령, 라스(2017)는 영국 논설 영어 기사를 인간 번역사와 기계 번역기가 스웨덴어로 번역한 결과물에서 어순의 변화를 비교 분석하였는데 인간 번역물에서 압도적으로 많은 어순 변화가 발생하였다. 이러한 어순 변화는 대부분 문법적 차이를 반영했지만 인간 번역물에서는 원문의 스타일을 향상하려는 의도의 변화도 목격되었다. 반면 기계 번역 결과물에서는 인간 번역에서 볼 수 있는 문장 나누기, 원문의 텍스트 기능 변화, 명시화, 관점 변화, 의역 등의 시도가 관찰되지 않았다. 웹스터 외(2020)는 구글과 DeepL 기계 번역기를 사용하여 4편의 영어 고전 소설을 네덜란드어로 번역한 후 인간 번역과 차이점을 비교 분석하였다. 수작업에 의한 오류 분류, 어휘 다양도, 어휘 결속성, 원문과의 구문적 차이에 대한 통계분석, 버로우의 델타값(Burrows Delta)을 사용한 문체 차이 등을 분석하였다. 분석 결과 기계 번역에서는 인간 번역에서는 찾아보기 힘든 오류가 많이 발견되었고 인간 번역보다 어휘 다양도 및 결속성 수준이 낮았다. 또한 기계 번역은 인간 번역보다 원문의 구문 구조를 쫓아가는 경향을 보였으며 문체에서도 인간 번역과 차이를 보였다.

국내 연구를 보면 한승희(2020)는 체계 기능 언어학 이론에 따라 3가지 의미 층에서 여러 언어 자질의 발생 빈도를 비교 분석한 결과 인간 번역, 기계 번역, 컴퓨터 보조 번역 등의 번역 방식 간에 뚜렷한 차이가 발견되었다. 이현주(2022)도 중국어 소설을 인간 번역사와 기계 번역기가 한국어로 번역한 결과물에서 지시 및 대용, 접속, 반복 등의 결속 구조를 대조 분석한 결과 기계 번역이 인간 번역에 비하여 결속 구조를 적절하게 활용하지 못하는 것으로 드러났다. 전혜진(2019)은 톨스토이의 『유년시절』의 인간 번역과 기계 번역 한국어

결과물을 비교 분석하였다. 저자는 어휘, 문법, 화용, 문체, 문화 등 다양한 층위에서 예문 분석을 통해 텍스트 분석, 맥락 이해, 창의성 등을 요하는 문학번역에서 기계 번역은 번역 적절성을 달성하기 어렵다고 평가하였다. 이창수(2021)는 한국어 장-단편 소설 28편의 인간 번역 및 기계 번역 결과물의 문체 차이를 비교 분석하였다. 특히 기계 번역 결과물은 2019년과 2020년 두 차례에 걸쳐 수집하여 기계 번역의 변화 추이도 분석하였다. 동 연구에서는 분석 언어 자질로 최빈도 어휘를 활용하였으며 차원축소분석법 중 하나인 선형판별분석(LDA)와 랜덤포리스트(RF) 기계학습 알고리즘을 사용하여 문서 식별의 정확도를 분석하였는데 2019년과 2020년에 모두에서 기계학습 알고리즘은 인간 번역과 기계 번역 결과물을 정확히 예측해냄으로써 두 번역 방식 간에 분명한 문체 차이가 있음을 입증하였다. 또한 1년 동안 기계 번역기들과 인간 번역과의 거리는 좁혀지지 않았지만, 기계 번역기 간에 거리가 좁아져 시간이 지나면서 문체에서 서로 비슷해지는 경향이 관찰되었다.

이상에서 보듯이 인간 번역과 기계 번역 간의 언어적 차이를 규명하려는 연구에서는 다양한 언어 표지와 분석 방법이 사용되고 있다. 그중 컴퓨터를 사용하여 문체 차이를 분석한 연구에서는 웹스터 외(2020)와 이창수(2021) 연구에서 보듯이 최빈도 어휘를 주로 활용한다. 최빈도 어휘는 문헌의 저자 판별을 핵심으로 하는 전산문체학(stylometry) 연구에서 가장 많이 활용되는 언어 표지이다(Oakes and Pichler 2013: 224). 이 경우 앞서 언급한 LDA, RF나 주성분분석(PCA). 요인 분석(FA), 대응분석(CA)같은 차원축소분석법이 흔히 사용된다. 최빈도 어휘를 번역학 연구에 적용한 초기 연구로는 버로우즈(Burrows 2002)를 들 수 있다. 동 저자는 최빈도 어휘에 기초하여 버로우즈 델타라는 문서 거리 값을 산출한 후 영국의 왕정복고 시대에 라틴어 시인 주벨라가 쓴 시를 영역한 영국 시인이 누구인지를 밝혀내는 연구를 했다. 분석 결과 번역 시인에 따라 번역문에서 자신의 문체가 드러나는 경우와 그렇지 않은 경우가 공존했다. 유사한 목적과 방법을 사용한 번역 연구로는 리빅키(Rybicki 2008, 2012)와 리(Lee 2018)를 들 수 있다.

저자 판별 분석에서 사용되는 언어 표지는 최빈도 어휘 외에도 매우 다양하다. 스타마타토스 외(Stamatatos et al. 2000)는 문체 분석에서 사용되는 언어 표지를 크게 (1) 어휘 수, 문장 수 등이 포함된 토큰(단어) 차원, (2) 수동태 구

문 수, 명사구 수 등 구 차원, (2) 총 어휘 수, 어휘 다양도(TTR) 차원, (4) 최빈도 어휘 차원 등으로 구분하였다. 최빈도 어휘는 저자를 분별하는 데는 매우 효과적이지만 어휘 빈도에 기초한 통계이기 때문에 형태학적이나 통사적 구조에 따른 문법적 문체 차이를 반영하는 데는 한계가 있다. 이에 대한 대안으로 바이버(1988)가 장르 간 문체 변이를 연구하는 데 사용했던 총 67개의 언어 표지에 주목할 필요가 있다. 바이버는 특별히 고안한 컴퓨터 프로그램을 사용하여 21개 장르, 481개의 텍스트로 구성된 코퍼스에서 67개 언어 자질의 발생 빈도를 추출하였다. 이 언어 자질들은 텍스트별로 빈도 차이가 있는데 이를 통계 용어로 분산이라고 한다. 그런데 자질 중에는 서로 유사한 분산 패턴을 보이는 것들이 있다. 바이버는 요인 분석법(factor analysis)이란 다차원 분석법을 사용하여 분산 패턴이 유사한 자질을 하나의 세트로 묶어 총 6개의 새로운 변수를 도출하였는데 이를 요인이라고 한다. 바이버는 이 요인들이 텍스트의 기본적 의사소통 기능을 반영한다고 가정하고 각 요인과 상관관계가 강한 언어 자질을 분석하여 각 요인의 의사소통 기능을 규정하였다. 이렇게 해석된 요인은 차원(dimension)이라고 불렀다. 가령, 차원 1과 긍정적 상관관계를 가진 언어 자질에는 see, perceive, think, believe 등 지각 및 인지 의미를 가진 사적 동사(private verbs)와 that을 생략한 that 절 등이 있는데 이는 구어 텍스트의 특징이다. 반면에 차원 1과 부정적 상관관계를 가진 언어 자질에는 명사, 스펠링이 긴 단어 등이 있는데 이는 정보 전달이 목적인 문어 텍스트의 특징이다. 이를 바탕으로 차원 1은 개입형(involved) 텍스트와 정보형(informational) 텍스트를 구분하는 의사소통 기능을 가진 것으로 해석하였다. 이러한 과정을 거쳐 총 6개 차원의 의사소통 기능을 규정하고 각 차원과 장르 간의 관계를 분석하였다.

번역학에서도 바이버의 영향을 받아 다양한 문법적 언어 자질을 사용하여 번역과 비번역 텍스트의 문체 차이를 규명하려는 연구가 일부 시도되었다. 가령, 쿠니로프스카야와 라프시보나-콜튼스키(Kunilovskaya and Lapshinova-Koltunski 2019)는 총 49개의 어휘 및 구문 언어 표지를 주성분 분석의 분석 자질로 사용하여 영-러 번역에서 번역과 비번역 텍스트 간에 문체 차이를 분석하였다. 그 결과 학생과 전문번역사 번역물 모두에서 번역 텍스트는 비번역 텍스트와 문체적으로 명확하게 구분되었다. 후 외(Hu et al 2019)는 바이버가 사용한 67개의 언어 자질을 이용하여 7개 장르에서 번역된 영어 텍스트가 비번역

텍스트와 어떻게 차이가 나는지를 연구하였다. 그 결과 총 26개의 언어 자질이 번역과 비번역 텍스트를 구분하는 역할을 하는 것으로 밝혀졌다. 이 같은 선행 연구에 기초하여 본 연구에서도 인간 번역과 기계 번역 텍스트의 문체 차이를 규명하는데 바이버의 67개 언어 자질을 활용하였다.

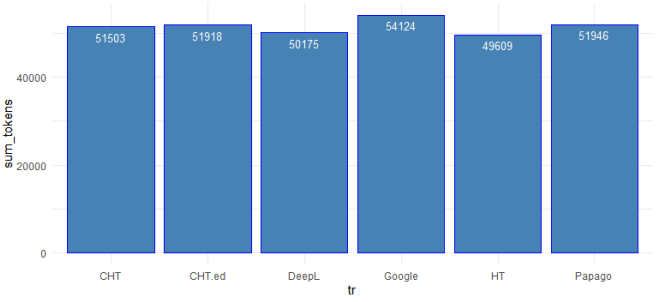
3. 연구 방법

3.1 연구 데이터

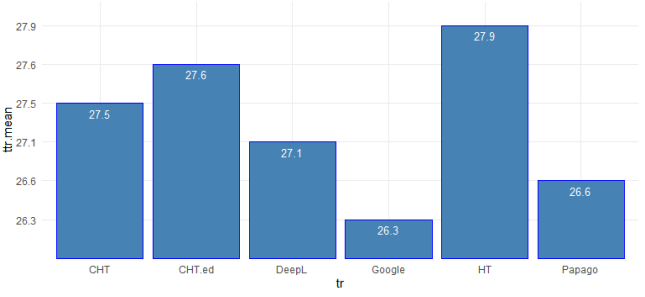
본 연구에 사용된 분석 데이터는 2017년 5월에 중앙일보, 경향신문, 한겨레 신문 웹사이트에서 수집한 총 48개의 한글 사설의 영어 번역문과 2023년 6월에 파파고, 구글, DeepL, 챗GPT 등 4개의 인터넷 기계 번역기에서 추출한 각 48개 영어 번역문, 그리고 챗GPT에게 “Proofread the translation(번역문을 감수하라)”이란 프롬프트를 주어 자기 번역문을 수정하도록 하여 얻어진 48개의 수정 번역문으로 구성되었다. 이 같은 6가지 번역 방식은 분석 데이터에 각각 HT, Papago, Google, DeepL, CHT, CHT_ed 등으로 표기하였다. 총 분석 텍스트 수는 인간 번역문 48개와 기계 번역 5개 모드에서 추출한 48개 번역문 240개를 더한 총 288개이다. 신문사별 사설 수는 경향신문 11개, 중앙일보 20, 한겨레 17개이다. 신문사별로 영역을 제공하는 사설 수가 다르기 때문에 수집된 사설 수에서도 차이가 있다.

코퍼스에 대한 기본 통계 정보를 보면 코퍼스의 총 단어 수는 309,257로 각 번역 방식 별 총 단어 수와 어휘 다양도를 나타내는 TTR(총 단어 수 대 단어 유형 수)은 각각 <그림 1> 및 <그림 2>와 같다. 번역된 총 단어 수는 HT가 가장 적은 데 반하여 TTR은 HT가 가장 높다. 기계 번역기 중에서는 Papago가 가장 높고 챗GPT의 경우는 CHT을 자가 수정한 CHAT_ed에서 TTR이 조금 상승한 것으로 나타났다. 문장 길이를 나타내는 문장의 평균 단어 수는 <그림 3>에 나와 있는데 HT는 기계 번역기들에 비하여 중간 정도의 수준이다.

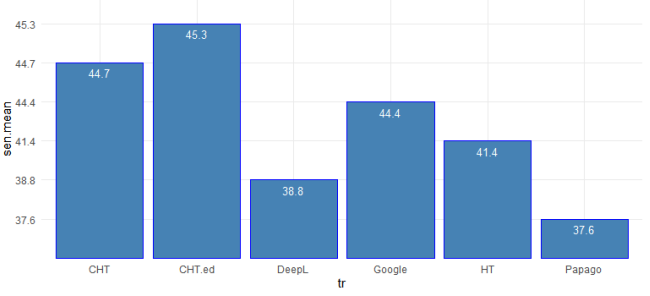
〈그림 1〉 번역 모드 별 총 단어 수



〈그림 2〉 번역 모드 별 TTR



〈그림 3〉 평균 문장 길이 (단어 수)



3.2 분석 방법

앞서 2절에서 언급하였듯이 본 연구에서는 바이버(1988, 1995)가 사용한 67개의 형태소, 구문, 어휘 표지를 분석 언어 자질로 사용하였다. 언어 자질은 크

게 (1) 시제 및 양상 표지 (2) 장소 및 시간 부사 (3) 대명사 및 프로 do 동사 (4) 의문문 (5) 명사구 (6) 수동태 (7) 종속 구조 (8) 전치사구 (9) 형용사 (10) 부사 (11) 어휘 특이성 (12) 어휘 집단 (13) 조동사 (14) 특수 동사 집단 (15) 축약형 및 비연속 구조 (16) 등위 관계 (17) 부정어 등으로 분류된다.

상기 언어 자질을 사용하여 분석 텍스트에 주석을 달고 통계치를 구하는 작업은 '다차원 분석 태거(Multidimensional Analysis Tagger: MAT)'(Nini 2019)란 프로그램을 사용하였다. MAT에 분석 텍스트를 탑재하면 단어별로 바이버의 67개 자질을 주석 형태로 첨부하고 텍스트별 발생 빈도를 엑셀용 csv파일로 생성한다. 이렇게 생성된 csv에 <그림 4>와 같이 텍스트별 id, 신문사(newspaper), 번역 모드(tr), 인간 번역과 기계 번역을 구분하는 범주 항목(group) 변수를 추가하였다.

이 같은 csv 분석 자료를 R 컴퓨터 프로그램에 탑재한 후에 FactoMineR이란 패키지를 사용하여 바이버가 사용했던 요인 분석과 유사한 다차원 분석법 중 하나인 주성분 분석(PCA)를 실시하였다. 두 분석법은 차원 축소법이란 점은 공통적이지만 요인 분석은 사전에 데이터에 어떤 잠재 변수가 존재한다고 가정하고 이와 관련된 변수들을 요인(factor)으로 묶어 가설을 검증하는 데 사용하는 반면에 PCA는 다차원 데이터에서 유사한 변이 패턴을 보이는 변수들을 선형적으로 묶어 주성분이란 새로운 변수를 도출한다. 본 연구처럼 특정한 가설이 없이 다양한 언어 자질과 텍스트가 어떻게 연관되어 있는지를 탐구 관찰하는 목적에는 PCA가 적합하다.

〈그림 4〉 본 연구에 사용된 분석 데이터

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	newspape	tr	group	AMP	ANDC	AWL	CAUS	CONC	COND	CONJ	DEMO	DEMP
2	HK_01_Ch	Hankyore	CHT_ed	MT	1.19	0.27	1.75	-0.65	1.75	1.5	2.87	0.88	0.67
3	HK_01_Ch	Hankyore	CHT	MT	1.15	0.25	1.6	-0.65	1.75	1.45	2.81	0.81	0.63
4	HK_01_De	Hankyore	DeepL	MT	-1.04	-0.48	1.37	-0.65	2.12	1.91	2	-0.24	0.9
5	HK_01_Go	Hankyore	Google	MT	0.65	-0.02	1.75	0.65	2.12	1.86	3.38	1.31	1.33
6	HK_01_HT	Hankyore	HT	HT	-0.27	-0.08	0.97	-0.65	-0.63	2.55	1.81	-1.88	-0.96
7	HK_01_Pa	Hankyore	Papago	MT	-0.19	-0.48	1.15	0.65	2.12	1.86	4.75	-0.79	0.88
8	HK_02_Ch	Hankyore	CHT_ed	MT	0.58	-0.06	1.27	-0.65	2	1.73	3.19	-1.36	-0.52
9	HK_02_Ch	Hankyore	CHT	MT	0.54	-0.08	1.25	-0.65	2	1.68	3.13	-1.38	-0.52
10	HK_02_De	Hankyore	DeepL	MT	-0.19	1.31	0.6	-0.65	-0.63	1.82	1.94	-1.83	-0.06
11	HK_02_Go	Hankyore	Google	MT	0.77	-0.46	0.73	0.71	2.25	2.05	3.62	-2.36	0.02
12	HK_02_HT	Hankyore	HT	HT	0.38	-0.94	0.32	-0.65	-0.63	1.36	2.69	-2.36	0.96
13	HK_02_Pa	Hankyore	Papago	MT	-0.23	-0.06	0.52	-0.65	2	1.73	1.88	-1.86	-0.08
14	HK_03_Ch	Hankyore	CHT_ed	MT	0.73	-0.94	1.53	-0.65	2.25	-1.14	2.13	2.05	0
15	HK_03_Ch	Hankyore	CHT	MT	1.58	-0.94	1.7	-0.65	2.25	-0.09	2.06	2.5	-0.02
16	HK_03_De	Hankyore	DeepL	MT	0.69	-0.48	0.85	-0.65	-0.63	-0.14	-0.75	1.36	0.9
17	HK_03_Go	Hankyore	Google	MT	0.5	-0.94	1.03	-0.65	-0.63	0.68	0.5	1	-0.13

4. 분석 결과

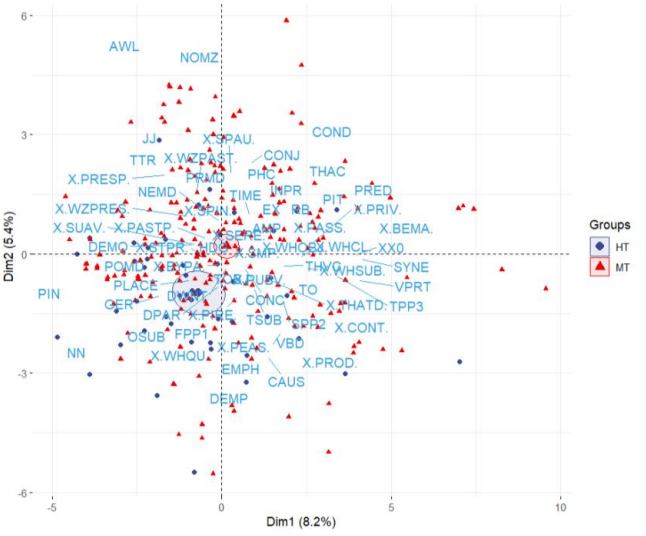
본 장에서는 인간 번역과 기계 번역 및 기계 번역기 간의 문체 차이와 각 번역 방식을 1 대 1로 비교 분석한 문체 차이를 논하도록 한다.

4.1 인간 번역 대 기계 번역 간의 일반적 문체 차이

먼저 PCA 분석 결과 중 분석 텍스트가 인간 번역(HT) 대 기계 번역(MT)으로 명확히 구분되는지를 살펴보도록 한다. <그림 5>는 <그림 4>의 분석 텍스트에서 group을 범주 변수로 잡고 2차원 화면에 288개의 텍스트와 언어 자질을 동시에 배치한 바이 플롯(biplot)이란 그래프이다. 본 연구의 분석 데이터는 67개의 언어 자질을 분석 변수로 한 67개 차원을 가진 데이터이다. 앞서 간단히 언급했듯이 PCA는 이 같은 다차원 데이터의 차원을 선형적으로 합쳐서 소수의 새로운 차원을 도출하는 차원 축소 분석법이다. <그림 5>의 그래프는 이런 방법으로 새롭게 도출한 차원 중 Dim 1을 x축, Dim 2를 y축으로 한 그래프이다. 기타 Dim 3, Dim 4 ... 등도 분석할 수 있지만 이들 기타 차원을 살펴본 결과 번역 방식 간 문체 차이를 가장 확실히 반영한 것은 Dim 1과 Dim 2였기 때문에 여기서는 이 두 개 차원을 중심으로 논의를 전개하도록 한다.

<그림 5>의 그래프를 보면 HT와 MT의 텍스트가 중간에 일부 섞여 있지만 일반적으로 HT는 Dim 2를 따라 아래쪽으로 MT는 위쪽으로 퍼져나가면서 분포해 있다. 두 집단의 중간에 있는 작은 타원은 각 집단의 평균점을 연결한 신뢰 타원(confidence ellipse)이다. 두 타원이 겹치지 않고 명확한 거리를 두고 떨어져 있다. 이는 두 집단의 문체가 평균적으로 분명한 차이가 있다는 것을 의미한다.

〈그림 5〉 HT 대 MT



이 두 집단을 구분하는데 가장 큰 기여를 하는 언어 자질은 두 신뢰 타원의 중심을 직선으로 연결한 선(y축이 x플러스 방향으로 약간 기운 선)에 가장 가까운 것들을 살펴보면 된다. 이 선을 따라 밖으로 나갈수록 상관관계 강도가 높다. 지면 제약상 모든 언어 자질을 다 분석할 수 없지만 몇 가지 눈에 띄는 것을 살펴보기로 한다. 먼저 MT 영역(우상단)에서 가장 눈에 띄는 것은 COND(if 조건절), CONJ(접속사), THAC(형용사 보어 that절; 'I'm afraid that...')등 절을 연결하는 언어 자질이다. 이는 MT가 HT에 비하여 종속절이나 내포절 같은 복합절을 상대적으로 더 많이 사용한다는 것을 의미한다. 또 EX(there 존재절; 'There is ...'), BEMA(be동사 구문), PASS(by가 없는 수동태 구문: 'X is needed')등도 MT를 특징짓는 언어 자질이다. 또한 MT는 HT에 비하여 PIT(it 대명사)와 INPR(부정대명사; 'all', 'any', 'none', 'one', 'each' 등)의 발생 빈도가 높다. 이에 비하여 HT의 경우는 상대적으로 NN(일반 명사), OSUB(기타 종속절), GER(동명사), FPP1(1인칭 대명사), WHQU(wh-의문문), DEMP(지시대명사), EMPH(강조어), PLACE(장소 부사) 등이 특징적 언어 자질로 나타난다.

MT가 구문 구조를 크게 바꾸지 않고 원문을 직역하는 특성(Rothwell et al.

2023: 110-111)을 감안하면 MT와 연관된 언어 자질 중 상당수는 한국어 원문의 구조가 반영된 결과일 가능성이 크다. EX(there 존재절)를 예로 들어 보자. 우리말에선 '-이 있다/없다'는 구문이 자주 쓰인다. 이를 영어로 직역하면 'There+be동사'가 된다. 예문 (1)을 보면 원문이 '...이 없지 않다'로 끝난다. 기계 번역의 결과물을 보면 DeepL을 제외하고 모두 there로 시작하는 직역 문장이다. 이에 반하여 HT는 we를 주어로 사용하여 직역을 피하였다.

예문 (1) JA01
ST: 아쉬운 대목도 없지 않다.
Papago: There are some disappointing points.
Google: There are no missing parts.
DeepL: It is not without its regrets.
CHT: There are also regrettable aspects.
HT: But we have some regrets, too.

예문 (1)에서 한 가지 더 주목할 점은 HT에서 주어를 we로 사용하였다는 점이다. 이는 HT의 특징적 언어 자질에 FPP1(1인칭 대명사)가 포함된 것과 연관 있다. 여기서 we는 사설 집필자(또는 신문사)와 독자를 총칭하는데 영어 신문 사설에서 이런 식의 we가 자주 사용되어 '사설-we(editorial-we)'라고 불린다 (Westin 2002: 44). HT 언어 자질에 FPP1이 포함된 것은 기계 번역기에 비하여 인간 번역사가 사설-we를 사용하는 경향이 강하다는 의미이다. 우리말 사설의 경우 '우리'란 대명사는 '우리 경제' '우리나라의 저출산 정책'같이 대한민국 공동체를 의미하는 뜻으로 주로 쓰이며, 영어식 사설-we의 의미로 쓰이는 경우는 거의 없다. 흥미로운 점은 우리말 사설에서 대한민국 공동체를 지칭하는 '우리'의 경우 인간 번역이나 기계 번역 모두에서 예문 (2)처럼 Korea로 바꿔 번역한다는 점이다. 이런 상황을 놓고 보면 HT 번역문에서 FPP1이 특징적 언어 자질로 나타난 것은 인간 번역사가 영어 사설 규범에 맞춰 '사설-we'를 창의적으로 삽입한 결과로 여겨진다. 데이터에서 추적이 비교적 용이한 there 구문과 1인칭 대명사 we를 예로 들어 설명하였지만 예문 (1)의 분석은 인간 번역과 기계 번역의 문체 차이를 발생시키는 근본적 요인은 기계 번역의 직역식 특징과 인간 번역의 창의성이란 점을 보여준다.

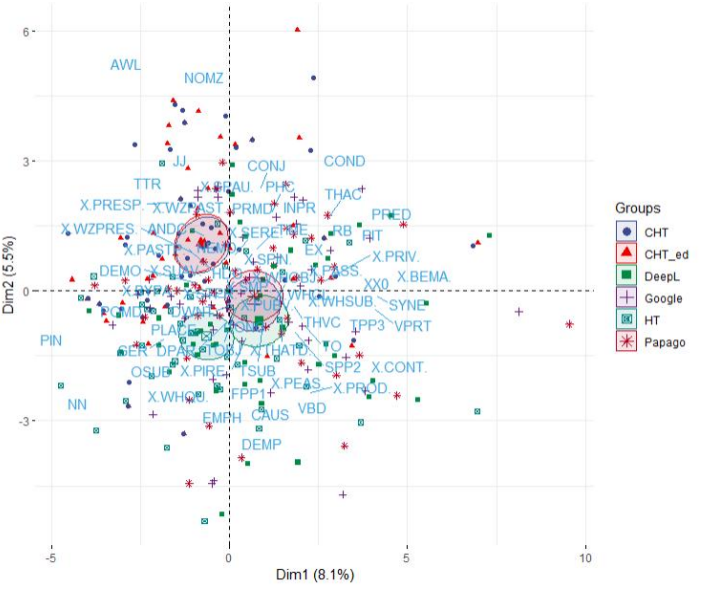
예문 (2) JA03
ST: 지금 우리나라의 저출산 정책이 계속 헛발질을 하는 데에는 바로 이 이유가 클 것이다.
HT: This is the very reason why Korea's policies to boost the birthrate constantly fail.
Papago: This must be the reason why Korea's low birth rate policy continues to be in vain.

그렇다면 기계 번역기들 사이에는 문체 차이가 없을까? 특히 2022년에 새로 등장한 챗GPT와 기존 기계 번역기 간에 문체 차이는 없을까? 이런 질문에 대한 답을 구하기 위하여 이번에는 <그림 4>의 분석 데이터에서 tr을 범주 변수로 하여 PCA 분석을 한 결과를 살펴보자. 그 결과는 <그림 6>의 그래프에 나와 있다. 이 그래프에서는 몇 가지 눈여겨볼 점이 있다. 첫째, 맨 위에 CHT_ed와 CHT의 신뢰 타원이 거의 완전히 겹쳐있다. <그림 1>, <그림 2>, <그림 3>에서 보면 GHT_ed는 CHT보다 단어 수에서는 큰 차이가 없지만, 어휘 다양도(TTR)는 증가하고 문장 평균 길이도 늘어났다. 그러나 <그림 6>의 그래프는 이 같은 변화가 문체를 크게 바꿀 정도는 아니란 것을 보여준다. 즉, 어휘 수준에서의 수정은 있지만 구문 구조를 바꾸는 차원의 수정은 거의 없다고 볼 수 있다. 둘째, Dim 2축을 따라 CHT와 CHT_ed 밑에는 Papago, Google, DeepL이 하나의 군집을 형성하고 있다. Papago와 Google이 거의 겹쳐있고 DeepL이 이들과 반쯤 겹쳐 있다. 이 군집은 챗GPT 집단과 거리를 두고 떨어져 있다. 이는 기존 기계 번역기 간에는 통계적으로 유의미한 문체 차이가 존재하지 않지만 챗GPT와는 뚜렷한 차이가 있다는 것을 의미한다. 셋째, 인간 번역인 HT는 여전히 챗GPT 및 기타 기계 번역기와 떨어져 있지만 거리상 기존 기계 번역기 군집이 HT에 더 가까이 포진해있다. 챗GPT가 기존 기계 번역기보다 인간 번역과 문체 차이가 더 크다는 점은 흥미로운 결과이다.

결론적으로 <그림 6>의 그래프에 따르면 HT, 챗GPT, 기타 기계 번역기 등 세 집단 간에 뚜렷한 문체 차이가 존재한다. 이 같은 상황은 앞서 <그림 5>의 그래프에서 챗GPT와 기타 기계 번역기를 하나의 집단으로 묶어 분석한 것이 인간 번역과 기계 번역 간의 문체 차이를 제대로 반영하지 않았을 가능성을 시사한다. 즉, 세 집단 간에 문체 차이가 존재하기 때문에 이들 집단을 각각 1대

1로 비교해봐야 좀 더 정확한 문체 차이를 규명할 수 있을 것이다. 이 같은 배경에서 (1) HT 대 챗GPT, (2) HT 대 기타 기계 번역기 (3) 챗GPT 대 기타 번역기 등 3가지 차원에서의 비교분석을 시도해 보았다.

〈그림 6〉 전체 번역 모드



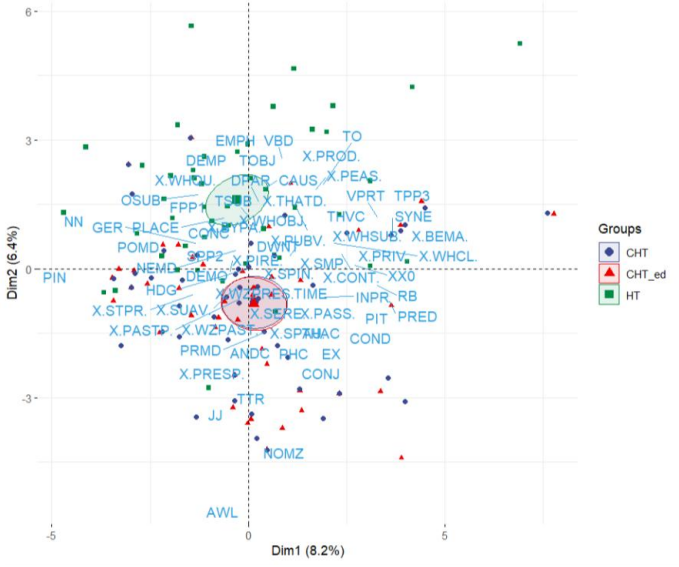
4.2 HT, 챗GPT, 기타 기계 번역기 간의 1 대 1 대조 분석

먼저 HT 대 챗GPT 간의 문체를 비교해 보면 <그림 7>과 같다. 이 그래프에서는 앞선 두 그래프와 달리 상단에 HT가 있고 하단에 CHT, CHT_ed가 배치되어 있다. 이 그래프를 보면 <그림 5> 그래프에서 HT의 특징이라고 분석했던 COND(if 조건절), CONJ(접속사), THAC(형용사 보어 that절), EX(there 존재절), PASS(by가 없는 수동태 구문) 등이 여전히 챗GPT의 영역에 배치되어 있다. 이에 반하여 HT와 연관된 언어 자질을 보면 <그림 5>에서 HT의 특징으로 분석한 자질 중에는 WHOU(목적어 자리의 관계절), EMPH(강조어), DEMP(지시대명사), FPP1(1인칭 대명사) 등이 좀 더 중심적 자리로 옮겨와 있고 VBD(과거시제), TOBJ(목적어 자리의 that 관계절) 등이 추가되었다. 약간의 변

동이 있지만 <그림 5>에서 분석했던 인간 번역과 기계 번역을 구분하는 언어 자질이 거의 대부분 HT와 챗GPT를 구분하는데도 유효함을 알 수 있다.

그러나 <그림 7>의 그래프에서 챗GPT를 가장 특징짓는 언어 자질은 AWL(평균 단어 길이), NOMZ(명사화; -ity, -ment, -ization, -ation 등의 접미사가 붙은 명사), TTR(어휘 다양도), JJ(일반 형용사) 등이다. 이중 AWL, NOMZ, TTR은 바이버(1988: 160-164)의 연구에서 정보형 문어체 텍스트의 특징으로 분류되었다. 이는 인간 번역과 대조할 때 챗GPT는 문어체적 특성이 더 강하다는 것을 의미한다. 이는 챗GPT의 언어가 문어체 특징이 강하며 그런 특징이 번역 결과물에도 반영될 것이라는 추측을 뒷받침하는 결과이다. 즉, 문체에서 챗GPT가 인간 번역과 뚜렷한 차이가 나는 이유는 한국어 구문 구조를 직역하는 기계 번역의 내재적 특징과 더불어 챗GPT가 갖고 있는 문어체 편향성이다.

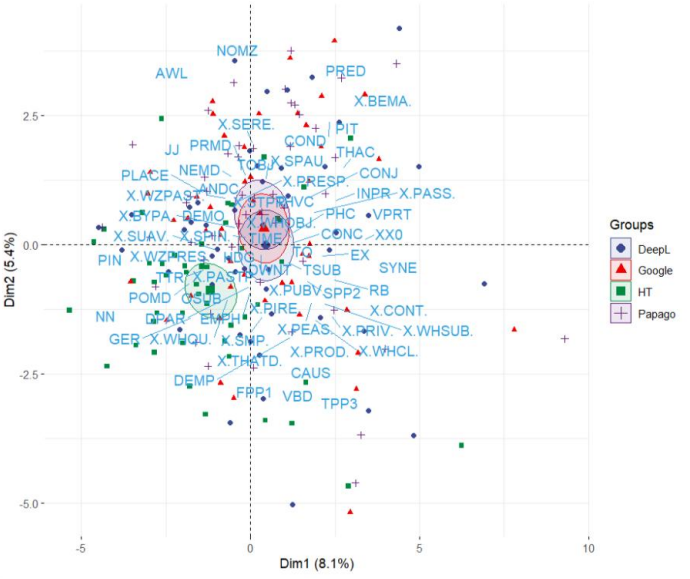
〈그림 7〉 HT 대 챗GPT



다음에는 <그림 8> 그래프에 나와 있는 HT 대 기타 기계 번역기 간의 문체 차이를 살펴보자. 이 그래프에서는 HT는 Dim 2 축 하단에, 기타 기계 번역기들은 상단에 자리 잡고 있다. 이 그래프에서도 앞서 MT의 특징적 언어 자질

로 분석했던 COND(if 조건절), CONJ(접속사), THAC(형용사 보어 that 절), EX(there 존재절), PASS(by가 없는 수동태 구문), PIT(it 대명사), INPR(부정대명사) 등이 모두 기계 번역기 군집 오른편에 배치되어 있다. 또한 HT와 연관 있는 것으로 분석했던 NN(일반 명사), OSUB(기타 종속절), GER(동명사), FPP1(1인칭 대명사), WHQU(wh-의문문) 등도 여전히 HT의 영역에 배치되어 있다. 특이한 것은 <그림 5>에서는 HT와 MT의 신뢰 타원 중심을 연결한 선에서 오른쪽으로 비켜나 있던 PRED(서술 형용사)와 BEMA(주동사 be)가 <그림 8>에서는 MT의 가장 특징적 언어 자질로 부상한 점이다. 이는 챗GPT를 제외했을 때 기계 번역은 인간 번역에 비하여 명사 앞에 오는 서술 형용사와 be 동사 구문을 더 많이 사용한다는 것을 의미한다.

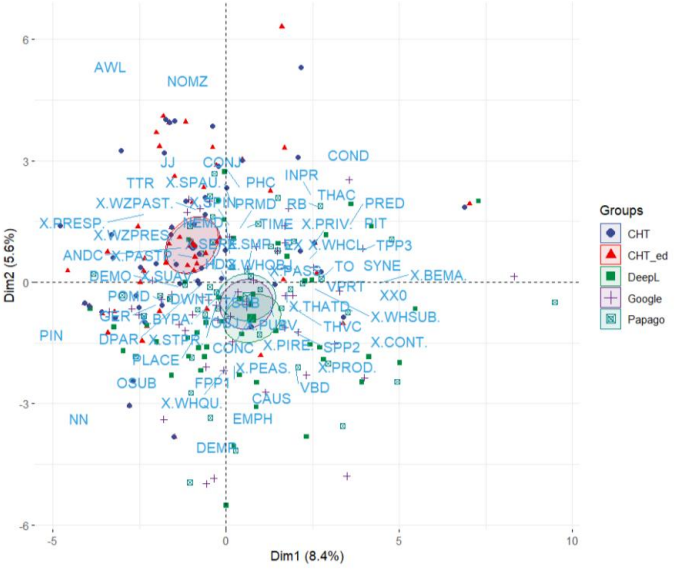
〈그림 8〉 HT 대 기타 기계 번역기



마지막으로 챗GPT 대 기타 기계 번역기 간의 문체 차이를 분석해 보자. <그림 9>의 그래프를 보면 CHT와 CHT_ed는 좌상단에, 기타 기계 번역기는 우하단에 배치되어 있다. 앞서와 마찬가지로 두 집단의 신뢰 타원의 중심을 연

결하는 직선을 따라 특징적 언어 자질을 살펴보면 <그림 7>과 <그림 8>의 그래프에서 HT와 대조하여 챗GPT 및 기타 기계 번역기의 문체를 구분했던 언어 자질들이 더 이상 그 같은 역할을 하지 않고 있음을 알 수 있다. 대신에 챗GPT에 대하여 기타 기계 번역기의 문체를 특징짓는 언어 자질에는 VBD(과거형), CAUS(계속적 관계절), EMPH(강조어), PROD(대동사 do), SPP2(2인칭 대명사), PEAS(과거완료형), FPP1(1인칭 대명사) 등이 있다. 이에 비하여 챗GPT와 상관관계가 높은 언어 자질로는 AWL(평균 단어 길이), NOMZ(명사화), JJ(일반 형용사), TTR(어휘 다양도), WZPAST(과거완료 수식절 wh삭제), SPAU(분리 조동사: “has already seen”, “will never see”), PRESP(현재완료 수식절) 등이 있다. 즉 기타 기계 번역기와 비교할 때도 챗GPT의 문어체 편향성은 문체 차이를 초래하는 핵심 요소이다.

〈그림 9〉 챗GPT 대 기타 기계 번역기



5. 논의 및 결론

이상의 논의를 종합하여 서론에서 제기했던 연구 문제에 대한 답을 정리하면 다음과 같다. 첫째, 인간 번역과 기계 번역 간에는 여전히 뚜렷한 문체 차이가 존재한다. 둘째, 기계 번역기 내에서는 챗GPT와 기존 기계 번역기 간에 확실한 문체 차이가 존재한다. 셋째, 챗GPT의 번역물과 이를 자가 수정한 결과물 간에는 큰 문체 차이가 없다. 넷째, 인간 번역 및 기타 기계 번역기와 대조할 때 챗GPT의 문체 차이를 야기하는 핵심 요소는 챗GPT의 문어체 편향성이다.

기존 연구와 비교하여 본 연구에서 새롭게 밝혀진 점은 2022년에 번역 시장에 새롭게 진입한 챗GPT가 인간 번역뿐만 아니라 기존 기계 번역기들과도 문체에서 뚜렷한 차이를 보인다는 점이다. 이는 기계 번역기에 사용된 AI와 훈련 방식의 차이에서 비롯된 것으로 보인다. 예측형 AI에 기초한 기존 신경망 기계 번역기는 통계적 확률이 가장 높은 답안을 제시한다. 이에 반하여 챗GPT는 학습 과정에서 추가적으로 인간 트레이너의 피드백을 반영한다. 이 같은 과정 때문에 챗GPT의 답변은 트레이너의 개인적 선호를 반영하는 편향성을 갖고 있다(Kocoń et al. 2023). 앞서 언급했던 챗GPT의 문어체 선호 성향도 이 같은 학습 과정의 결과이다. 본 연구 결과는 챗GPT의 문어체 편향성이 번역에서도 작동하며 그 결과 문체에서 인간 번역 및 기존 기계 번역기와 뚜렷한 차이가 있다는 점을 보여준다.

본 연구에서는 신문 사설이란 특정 장르에서의 문체 차이를 연구하였는데 본 연구에서 관찰된 결과가 다른 장르의 번역에서도 유효한지를 확인하는 추가 연구가 필요해 보인다. 또한 챗GPT와 인간 번역이 문체에서 차이가 나는 것으로 나타났기 때문에 두 번역 방식 간의 문체 차이를 보다 심층적으로 분석하는 추가 연구도 필요해 보인다. 특히 작업 프롬프트에 따라 다른 결과물을 내놓는 챗GPT의 특성을 고려하여 1차 번역물에 대하여 다양한 수정 프롬프트를 제시하여 궁극적으로 1차 번역물과 문체적으로 다른 결과물이 나오는지도 실험해 보면 흥미로울 것이다.

참고문헌

이창수 (2021)「기계학습 알고리즘을 활용한 문학번역에서의 기계 번역과 인간 번역 결과물 분류 연구」,『번역학연구』22(1): 199-217.

이현주 (2022)「문학작품의 중-한 기계번역 결과의 결속구조 분석 – 인간번역과의 비교를 중심으로」,『중어중문학』87: 259-282.

전혜진 (2019)「AI 시대, 문학번역에서 기계번역과 인간번역 비교분석 연구: 똘스또이의 '유년시절' 번역 분석을 중심으로」,『노어노문학』31(1): 111-154.

한승희 (2020)「인간번역, 기계번역, 컴퓨터보조번역 간 문체 비교 연구: SFL 이론을 중심으로」, 박사학위논문, 한국외국어대학교.

Biber, Douglas (1988) *Variation across Speech and Writing*, Cambridge: Cambridge UP.

BIber, Douglas (1995) *Dimensions of Register Variation: A Cross-linguistic Comparison*, Cambridge: Cambridge UP.

Ali Borji (2023) *A Categorical Archive of ChatGPT Failures*, arXiv:2302.03494 [cs.CL].

Burrows, John (2002) 'The Englishing of Juvenal: Computational Stylistics and Translated Texts', *Style* 36(4): 677-699.

Cegin, Jan, Jakub Simko and Peter Brusilovsky (2023) *ChatGPT to Replace Crowdsourcing of Paraphrases for Intent Classification: Higher Diversity and Comparable Model Robustness*, arXiv:2305.12947 [cs.CL]

Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify and Hany Hassan Awadalla (2023) *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation,* arXiv:2302.09210 [cs.CL]

Hu, Xianyao, Richard Xiao and Andrew Hardie (2019) 'How Do English Translations Differ from Non-Translated English Writings? A Multi-Feature Statistical Model for Linguistic Variation Analysis', *Corpus Linguistics and Linguistic Theory* 15(2): 347-382.

Kocoń, Jan, Igor Cichecki1, Oliwier Kaszyca1, Mateusz Kochanek1, Dominika Szydło1, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak and Przemysław Kazienko (2023) 'ChatGPT: Jack of All Trades, Master of None', *Information Fusion* 99: 1-37.

Kunilovskaya, Maria and Ekaterina Lapshinova-Koltunski (2019) 'Translationese Features as Indicators of Quality in English-Russian Human Translation', in *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, Shoumen, Bulgaria, 47-56.

Lars, Ahrenberg (2017) 'Comparing Machine Translation and Human Translation: A Case Study', in Irina Temnikova, Constantin Orasan, Gloria Corpas and Stephan Vogel (eds) *RANLP 2017 The First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT) Proceedings of the Workshop*, Shoumen, Bulgaria: Association for Computational Linguistics, 21-28.

Lee, Changsoo (2018) 'Do Language Combinations Affect Translators' Stylistic Visibility in Translated Texts?', *Digital Scholarship in the Humanities* 33(3): 592-603.

Läubli, Samuel, Rico Sennrich and Martin Volk (2018) *Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation,* arXiv:1808.07048 [cs.CL].

Mitrović, Sandra, Davide Andreoletti and Omran Ayoub (2023) *ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text*, arXiv:2301.13852 [cs.CL].

Nini, Andrea (2019) 'The Multi-Dimensional Analysis Tagger', in Sardinha, T. Berber and Pinto M. Veirano (eds) *Multi-Dimensional Analysis: Research Methods and Current Issues*, London: Bloomsbury Academic,

67-94.

Oakes, Michael and Alois Pichler (2013) 'Computational Stylometry of Wittgensteins "Diktat für Schlick"', *Bergen Language and Linguistics Studies (BcLLs)* 3(1): 221-240.

Rothwell, Andrew, Joss Moorkens, María Fernández-Parra, Joanna Drugan and Frank Austermuehl (2023) *Translation Tools and Technologies,* London and New York: Routledge.

Rybicki, Jan (2008) 'Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and Its Two English Translations', *Literary and Linguistic Computing* 21(1): 91-103.

Rybicki, Jan (2012) 'The Great Mystery of the (Almost) Invisible Translator: Stylometry in Translation', in Michael P Oakes and Meng Ji (eds) *Quantitative Methods in Corpus-Based Translation Studies*, Amsterdam: John Benjamins, 231-248

Stamatatos, Efstathios, Nikos Fakotakis and George Kokkinakis (2000) 'Automatic Text Categorization in Terms of Genre and Author', *Computational Linguistics* 26(4): 471–495.

Toral, Antonio, Sheila Castilho, Ke Hu and Andy Way (2018) *Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation*, arXiv:1808.10432 [cs.CL].

Toral, Antonio (2020) *Reassessing Claims of Human Parity and Super-Human Performance in Machine Translation at WMT 2019*, arXiv:2005.05738 [cs.CL].

Webster, Rebecca, Margot Fonteyne, Arda Tezcan, Lieve Macken and Joke Daems (2020) 'Gutenberg Goes Neural: Comparing Features of Dutch Human Translations with Raw Neural Machine Translation Outputs in a Corpus of English Literary Classics', *Informatics* 7(3): 32. Available at https://doi.org/10.3390/informatics7030032.

Westin, Ingrid (2002) *Language Change in English Newspaper Editorials*, Amsterdam: Rodopi.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes and Jeffrey Dean (2016) *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*, arXiv:1609.08144 [cs.CL]

[Abstract]

A Follow-up Study of Stylistic Differences between Human and Machine Translation with ChatGPT Added in the Mix

Chang-soo Lee
(Hankuk University of Foreign Studies)

The present study explores whether new shifts have developed in the stylistic landscape of human vs. machine translation in the wake of ChatGPT's arrival. For this purpose, it conducts a series of principal component analyses (PCAs) on a normalized frequency dataset comprising 67 morphological and syntactic linguistic features borrowed from Biber's (1988) research on register variation. The dataset is derived from a corpus of Korean editorials from three Korean newspapers, their human English translations, and English translations generated by four machine translation systems (Papago, Google, DeepL, ChatGPT), including ChatGPT's self-proofread versions. The analyses indicate that human and machine translation remain distinctly differentiated in terms of style, as demonstrated in previous studies. However, among the machine translation systems, ChatGPT, both in its translations and self-proofread versions, deviates significantly from the others. A closer examination of the linguistic features strongly associated with ChatGPT reveals that this difference can be attributed to the model's intrinsic preference for a formal, written style. Notably, there are no substantial stylistic divergences between ChatGPT's translations and its self-proofread versions.

Keywords: human translation, machine translation, stylistic analysis, ChatGPT, newspaper editorials
주제어: 인간 번역, 기계 번역, 문체 분석, 챗GPT, 신문 사설

이창수
한국외국어대학교 통번역대학원 교수
soolee@hanmail.net
관심 분야: 문학번역, 전산문체학, 코퍼스언어학, 체계 기능 언어학

논문 투고: 2023년 7월 13일
1차 심사 완료: 2023년 8월 31일
2차 심사 완료: 2023년 9월 10일
게재 확정: 2023년 9월 19일