

비교말뭉치 기반 한중 번역문 중국어의 언어적 특성 분석: AI 학습용 번역문과 일간지 번역문을 중심으로*

황은하(배재대학교)
비립(숙명여자대학교)

1. 서론

본 연구는 생산 조건이 상이한 두 한중 번역문 집단이 비번역 중국어와 공통으로 구별되는 분포적 특징을 보이는지, 그리고 두 집단 사이에 어떤 분포적 차이가 관찰되는지를 말뭉치언어학적으로 분석하는 것을 목적으로 한다. 이를 위해 비번역 중국어 기사와 국내 주요 일간지의 한중 기사 번역문, AI 학습용 한중 신문 기사 번역문으로 3원 비교말뭉치를 구성하고, 표기, 어휘, 통사 층위에서 그 차이를 검토한다.

번역의 목적과 기능은 텍스트 유형, 상정된 수용자, 의뢰와 사용의 조건에 따라 달라진다(Hermans, 1999; Toury, 1995). 출발 텍스트를 목표어 독자가 읽을 수 있게 하는 것은 번역이 성립하기 위한 기본 조건이지 번역의 목적 전부가 아니다. 최근에는 인공지능 학습용 데이터 구축을 위해 대규모 번역문이 생산 공개되고 있는데, 이 번역문은 인간 독자의 이해가 아니라

* 본 논문은 2026학년도 배재대학교 교내학술연구비 지원에 의하여 수행됨.

언어모델의 학습을 최종 용도로 한다는 점에서 일반 독자를 위한 번역과 생산 목적 및 사용 환경을 달리한다. 본 연구가 주목하는 것은 바로 이처럼 생산 조건이 구별되는 번역문이다.

한국지능정보사회진흥원이 운영하는 AI 허브의 한중 한영 한일 병렬말뭉치는 인공지능 학습용 데이터 구축을 목적으로 하며, 다수의 문장에 대해 기계번역 또는 클라우드 위커의 초벌 번역과 후편집을 거쳐 산출된다(한국지능정보사회진흥원, 2020). 반면 일간지 중국어판의 번역 생산 방식은 공개된 자료로 확인되지 않는다. 따라서 본 연구는 이 두 번역문 집단을 생산 조건, 즉 생산 목적, 상정된 수용자, 편집 단위, 품질 검수 방식의 차이로 구분한다. AI 학습용 번역문은 전통적인 출판 언론 번역과 생산 조건이 다르지만, 학습 데이터의 언어적 특성이 모델 출력에 전이될 수 있다는 점에서 목표어 텍스트로서의 특성을 점검할 필요가 있다(관련 논의는 2.1 참조).

번역문 텍스트가 목표어 원어민의 비번역문 텍스트와 계량적으로 구별된다는 사실은 중국어를 목표어로 하는 연구에서도 여러 차례 보고되었다(Xiao et al., 2010; 상세한 검토는 2.2 참조). 다만 이 연구들은 주로 영중 언어쌍에 집중되어 있으며, 생산 조건이 다른 번역문을 함께 대조한 사례는 드물다.

한편 원문과 번역문으로 구성된 병렬말뭉치는 대조언어학, 번역학, 외국어로서의 한국어 교육 등 응용언어학 연구에 요긴한 자료라는 측면에서도 그 언어적 특징의 분석이 필수적이다. 한중 언어쌍의 병렬말뭉치 기반 연구가 어느 정도 축적되어 왔으나, 개인 구축의 소규모 자료, 특정 장르 편중, 제한된 번역자 수 등의 문제로 대표성을 확보하기 어려운 경우가 있다(황은하, 2016, 2021). 그런 점에서 AI 허브의 대규모 병렬말뭉치는 이 문제의 대안이 될 수 있으나, 언어 연구를 위해 설계된 자료가 아니므로 활용에 앞서 목표어 텍스트, 번역문으로서의 실제성과 분포적 특성의 검토가 요구된다.

본 연구는 중국어 원어민이 생산한 비번역 중국어 신문 기사 말뭉치를 참조물로 삼고, AI 허브의 「한국어-중국어 번역 말뭉치 2(사회과학)」에 수록된 중국어 번역문(이하 ‘AI용 번역문’)과 국내 주요 일간지 중국어판의 기사 텍스트(이하 ‘일간지 번역문’)를 함께 대조하는 3원 비교말뭉치를 구성한다. 세 하위말뭉치는 신문 기사라는 매체 유형과 사회 경제 문화라는 대분류를

맞추고, 분야별 8만 문장씩 총 72만 문장, 약 1,655만 단어로 구성하였다. 이를 통해 다음의 연구 문제를 검토한다.

연구 문제 1. 두 번역문 집단은 비번역 중국어 기사와 구별되는 분포적 특징을 공통으로 보이는가? 차이가 있다면 어떤 지표에서 어떤 방향으로 나타나는가?

연구 문제 2. 두 번역문 집단 사이에 체계적 차이가 있는가? 차이가 있다면 어떤 지표에서 어떤 방향으로 나타나는가?

연구 문제 3. 연구 문제 1과 2에서 관찰된 차이는 번역 생산 조건의 어떤 측면과 정합적으로 해석되며, 어떤 대안 설명이 배제되지 않는가?

본 논문은 다음과 같이 구성한다. 2장에서는 번역투, 번역문 중국어의 계량적 접근에 관한 선행연구를 검토한다. 3장에서는 말뭉치 구성, 분석 도구와 지표, 통계 처리 방법을 제시한다. 4장에서는 표기 어휘 통사 층위의 분석 결과를 논의하고, 5장에서는 결과와 한계를 정리한다.

2. 이론적 배경과 선행 연구

2.1 번역 보편소와 번역투, 후편집투

번역 텍스트가 비번역 텍스트와 체계적으로 다른 언어적 특징을 보인다는 관찰은 번역학의 오랜 주제이다. 엘레르스탐(Gellerstam, 1986)은 스웨덴어 번역소설의 특징에 ‘번역어(translationese)’라는 이름을 붙였고, 베이커(Baker, 1993, 1996)는 이러한 차이가 개별 언어쌍이나 번역자에 한정되지 않을 가능성을 번역 보편소(translation universals)로 개념화하였다. 주요 후보로는 단순화(simplification), 명시화(explicitation), 정규화(normalisation), 간섭(interference)이 제시되어 왔다.

단순화는 번역문이 어휘적 통사적으로 단순해지는 경향이며, 라비오사(Laviosa, 1998)는 영어 비교말뭉치에서 번역 텍스트의 어휘 밀도와 어휘 다양성이 낮음을 보였다. 명시화는 원문에 함축된 정보가 번역문에서 표면화되는 현상이고, 정규화는 목표어의 전형적 표현으로 수렴하는 경향이다. 간

섭 또는 출발어 투과(source language shining through)는 출발어 구조가 번역문에 남는 현상으로, 언어쌍에 특징적이다(Teich, 2003). 한국어와 중국어의 언어쌍에서는 장현주(2006)가 중한 번역의 정보 배열에 나타난 번역투를 분석한 바 있다.

이 경향들을 보편소로 부를 수 있는지에 대해서는 논쟁이 계속되어 왔다. 체스터먼(Chesterman, 2004)은 원문과 번역문의 관계에서 관찰되는 S-보편소와 번역문과 비번역문의 관계에서 관찰되는 T-보편소를 구분하였고, 하우스(House, 2008)는 번역이 언어쌍 장르 시대의 조건 아래 이루어진다는 점에서 보편소 개념을 비판하였다. 본고는 번역 보편소와 번역투를 다음과 같이 구분한다. 번역 보편소는 언어쌍 장르 시대의 조건을 넘어 성립한다고 가정되는 일반 경향에 관한 가설이며, 그 성립 여부 자체가 논쟁의 대상이다. 번역투는 특정 언어쌍과 생산 조건 아래에서 목표어 텍스트에 관찰되는 표지를 가리킨다. 본 연구는 후자를 관찰 대상으로 삼고, 전자는 관찰된 분포를 해석하는 틀로만 참조하며 검증 대상으로 삼지 않는다. 따라서 번역투를 단일하고 불변하는 속성으로 전제하지 않으며, 생산 조건이 다른 번역문을 함께 대조하여 번역문과 비번역 텍스트 사이의 차이, 그리고 두 번역문 집단 사이의 차이를 기술하는 데 논의를 한정한다.

이러한 구분에 따를 때, 본 연구는 출발 텍스트를 분석 대상으로 삼지 않는 비교말뭉치 설계이므로 T-보편소 후보만을 다룬다. 명시화와 간섭은 출발 텍스트와 번역문의 대응을 전제하는 S-보편소에 해당하므로 본 설계로는 판정할 수 없으며, 이하에서는 선행연구를 소개하는 범위로만 언급한다. 본 연구가 검토하는 것은 단순화와 정규화 두 후보이다. 품사별 어휘 다양성은 단순화 및 정규화와, 문장부호는 목표어 표기 관행과 관련된다. 다만 하나의 지표는 복수의 설명과 양립할 수 있으므로, 관찰된 분포 차이를 특정 개념의 실현으로 단정하지 않는다.

한편, 기계번역과 후편집의 확산은 번역투 연구에 새로운 비교 대상을 제공하였다. 토랄(Toral, 2019)은 기계번역 후편집 텍스트가 인간 번역보다 어휘 다양성이 낮고 원문 구조에 더 밀착하는 경향을 ‘후편집투(post-edited)’로 논의하였다. 담스 등(Daems et al., 2017)과 반마센호버 등(Vanmassenhove et al., 2021)도 후편집 또는 신경망 기계번역 산출물이 인간

번역과 구별되는 언어적 특성을 보일 수 있음을 보고하였다. 본 연구는 AI 학습용 번역문의 언어적 특징을 후편집투와 동일시하지는 않지만, 기계번역 후편집을 포함하는 대규모 번역 생산 환경에서 목표어 텍스트의 특성을 점검할 필요가 있다는 점에서 이 논의와 연결된다.

2.2 번역 중국어에 관한 계량 연구

국내 번역학에서 자동 계량 평가 연구는 주로 원문과 번역문 사이의 대응 정확성, 즉 ‘무엇을 옮겼는가’를 평가하는 데 초점을 두어 왔다(김보영 등, 2020; 정혜연, 2018; 정혜연, 박현일 등, 2021; 정혜연, 최지수 등, 2022; 최지수, 2023; 최효은 등, 2023). 이에 비해 번역문 자체를 비번역 목표어 텍스트와 대조하여 ‘어떻게 옮겼는가’를 살피는 접근은 상대적으로 적다. 본 연구는 후자의 관점에서 목표어 텍스트의 분포적 특성을 분석한다.

번역 중국어의 계량적 특성에 관한 연구는 번역중국어 말뭉치(ZJU Corpus of Translational Chinese, ZCTC)와 비번역 중국어 말뭉치(Lancaster Corpus of Mandarin Chinese: LCMC)의 비교를 중심으로 축적되어 왔다(Xiao et al., 2010). 샤오(Xiao, 2010)와 샤오와 다이(Xiao & Dai, 2014)는 번역 중국어와 원어민 중국어 사이의 어휘 밀도, 기능어 분포, 구조조사(结构助词)¹⁾ ‘的’의 사용, 문절 길이 등에서 체계적 차이를 보고하였다. 이들은 원어민 중국어의 어휘 밀도가 더 높았으나, 표준화 타입-토큰 비율(STTR)²⁾에서는 유의한 차이가 없었다고 보고하였다. 또한 번역 중국어에서 ‘的’의 사용이 두드러지며, 이는 관형 수식 구조 및 문절 길이와 관련될 수 있음을 지적하였다.

다이(Dai, 2016)는 경동사³⁾ ‘进行’과 상 표지의 사용을 통시적으로 검토

- 1) 중국어에서 구조조사는 ‘的 地 得’와 같이 수식 관계를 표시하는 조사이고, ‘了 吗 呢’와 같이 문장의 어기를 표시하는 어기조사(语气助词)와 구분된다.
- 2) STTR은 텍스트를 1,000단어 구간으로 나누어 각 구간의 타입-토큰 비율을 구한 뒤 평균한 값이다. 본 연구가 3.3.1에서 쓰는 ‘표준화 TTR’은 공통 표본 크기에서 기대되는 타입 수를 산출한 것이므로 표준화의 방식이 다르다.
- 3) 경동사(light verb)는 ‘进行研究(연구를 진행하다)’의 ‘进行(진행하다)’처럼 그 자체의 어휘 의미가 거의 없고 뒤에 오는 명사가 실질 의미를 담당하는 동사를 말한다. 한국어의 ‘하다’, ‘실시하다’와 기능이 비슷하다.

하여, 일부 현상이 번역의 영향뿐 아니라 중국어 내부의 변화와도 관련됨을 보였다. 이는 번역 중국어의 분포적 특징을 해석할 때 출발어 간섭, 목표어 내부 변화, 장르와 시기 효과를 함께 고려해야 함을 시사한다. 병렬말뭉치 연구에서는 황리보(黃立波, 2008)가 영중 번역에서 인칭대명사 주어가 영어 원문보다 감소함을 보였는데, 이는 비교말뭉치에서의 번역 중국어 대 원어민 중국어의 차이와 병렬말뭉치에서의 번역문 대 출발어 원문의 차이가 서로 다른 질문에 답함을 보여 준다.

한국어를 출발어로 하는 번역 중국어에 대해서는 명시화 현상을 다른 소규모 연구가 있으나(신지언, 2017; 웨이잉, 2019), 어휘 다양성, 기능어 분포, 의존관계 구조를 비교말뭉치로 함께 분석한 연구는 드물다. 본 연구는 한중 언어쌍에서 번역 중국어의 계량적 특성을 기술하고, 생산 배경이 다른 두 번역문 집단을 함께 비교한다는 점에서 기존 연구를 보완한다.

이러한 공백을 메우기 위해서는 적절한 계량 지표가 필요하다. 목표어 텍스트의 복잡도를 계량화하는 방법은 제2언어 작문 연구에서도 발전해 왔다. 문장 절 구의 길이와 수는 통사 복잡도의 대표 지표이며(Hunt, 1965; Lu, 2010), 중국어에서는 연어와 의존관계에 기반한 지표가 학습자 작문의 숙달도 및 품질과 관련됨이 보고되었다(胡韧奋, 2021; 胡韧奋과 肖航, 2019; Hu et al., 2022). 본 연구는 이 가운데 품사별 어휘 다양성과 구 의존관계 기반 지표를 번역문 비교에 활용한다.

2.3 AI 학습용 한중 번역 데이터의 품질 관리

AI 허브의 한중 번역 말뭉치는 구축 가이드라인에 따라 정제와 품질 검수를 거친다. 자료 설명서에 따르면 한국어 원문은 일정 길이 범위에서 정제되며, 개인정보를 포함하거나 일부 필수 성분이 결여된 문장은 배제된다. 번역문은 기계번역 또는 클라우드 워커의 초벌 번역에 MTPE 1단계 또는 필요에 따라 2단계를 적용해 생성되며, 전체의 5% 표본에 대한 블랙박스 평가로 검수된다(한국지능정보사회진흥원, 2021a, 2021b).

이 절차는 한국어 원문-중국어 번역문 쌍의 대응 정확성을 중심으로 설계되어 있다. 규범적 문체적 적합성이 평가 항목에 포함되더라도, 번역문을 목표어 원어민 텍스트와 대조하여 그 언어적 분포를 점검하는 절차는 명시

되어 있지 않다. 또한 문장 단위로 정제 저장된 자료의 구성 방식은 번역문의 언어적 특성에 영향을 줄 수 있다. 본 연구는 이와 같은 관점에서 AI 학습용 번역문의 표기 어휘 통사적 분포를 검토한다.

3. 연구 대상과 방법

3.1 비교말뭉치의 구성

본 연구는 AI 학습용 번역문, 일간지 번역문, 비번역 중국어로 비교말뭉치를 구성하였다. 세 하위말뭉치의 구성 절차와 특징은 다음과 같다.

3.1.1 AI 학습용 번역문

AI 학습용 번역문은 AI 허브의 「한국어-중국어 번역 말뭉치 24(사회과학)」(한국지능정보사회진흥원, 2020)의 중국어 측 자료이다. 이 말뭉치는 세부 분야별로 나뉜 7종의 파일로 배포되며, 파일별로 교육, 금융/증시, 문화재/향토/K-FOOD, K-POP/대중문화공연콘텐츠, 조례, 정치/행정, 사회/노동/복지 등의 내용이다. 본 연구는 출처 확인을 거쳐 선별한 세 파일에서 분야별 8만 문장씩, 총 24만 문장을 무작위 추출하였다.

선별은 원천 CSV에 포함된 ‘출처’ 필드를 기준으로 하였다. 이 필드에는 원 기사의 URL 또는 한국언론진흥재단의 UCI 식별자가 기록되어 있다. 7종 파일의 출처 분포를 전수 집계한 결과, 사회/노동/복지, 금융/증시, K-POP/대중문화공연콘텐츠 세 파일은 출처가 기재된 문장이 모두 한국 언론사의 기사에서 온 것이었다. 반면 문화재/향토/K-FOOD 파일은 한국향토문화전자대전과 국제문화교류진흥원 등 비언론 출처가 대부분이었고 출처 미기재 비율도 56.4%에 이르렀다. 이에 본 연구는 앞의 세 파일만을 표집 대상으로 삼고, 이를 각각 사회 경제 문화 분야에 대응시켰다. 표집 대상 세 파일의

- 4) 이 번역 말뭉치는 문어와 구어, 메신저 대화 등 여러 사용역과 인문 사회 기초과학 방송 등 여러 도메인을 포괄한다. 본 연구가 사용한 자료는 2024년 5월 17일에 내려받았으며, 배포처가 정한 이용 조건에 따라 연구 목적으로 활용하였다.

출처 분포는 <표 1>과 같다.

표 1

AI 학습용 번역문 표집 대상 파일의 출처 분포

분야 (원천 파일)	문장 수/ 출처 기재 규모	확인된 출처
사회 (ko2zh_society, 사회/노동/복지)	160,000문장/ 113,185문장 (70.7%)	- 노컷뉴스 18,128, 한국경제 17,303, 언론진흥재단 UCI 16,651, 서울경제 15,913, 국민일보 13,728, 서울신문 12,727, 아시아경제 10,998, 파이낸셜뉴스 5,068, 전자신문 2,578 — 비언론 출처 0건 - 출처 URL 기준 개별 기사 13,319건, 기사당 평균 8.5문장
경제 (ko2zh_finance, 금융/증시)	160,000문장/ 159,534문장 (99.7%)	- 서울경제 43,751, 한국경제 23,502, 아시아경제 22,316, 국민일보 15,306, 언론진흥재단 UCI 14,468, 서울신문 12,647, 노컷뉴스 11,386, 전자신문 8,197, 파이낸셜뉴스 7,892 — 비언론 출처 0건 - 출처 URL 기준 개별 기사 38,908건, 기사당 평균 4.1문장
문화 (ko2zh_kpop, K-POP/대중문 화공연콘텐츠)	160,000문장/ 159,981문장 (100.0%)	- 서울경제 71,638, 국민일보 33,075, 서울신문 23,699, 노컷뉴스 19,241, 언론진흥재단 UCI 12,315 — 비언론 출처 0건 - 출처 URL 기준 개별 기사 44,779건, 기사당 평균 3.6문장

세 파일을 합하면 48만 문장 가운데 출처가 기재된 43만 2,700문장(90.1%)이 모두 한국 언론사 기사이며, 비언론 출처는 한 건도 확인되지 않았다. 출처가 기재되지 않은 4만 7,300문장(9.9%)은 대부분 사회 분야에 몰려 있으며, 문장 내용을 표본 검토한 결과 모두 기사문으로 판단되었으며, 따라서 본 연구는 이 문장들을 배제하지 않고 표집에 포함하였다.

아울러 AI 학습용 번역문 역시 출처 URL을 통해 기사 단위 연결정보를 보존한다. 표집 대상 문장은 사회 13,319건, 경제 38,908건, 문화 44,779건의 개별 기사에서 왔으며 기사당 평균 문장 수는 3.6-8.5문장이다. 곧 세 하위 말뭉치는 모두 문장이 기사 안에 군집된 위계적 자료이며, 이 점은 5장에서

다시 논의한다.

3.1.2 일간지 번역문

일간지 번역문은 중앙일보 조선일보 동아일보 아주일보⁵⁾ 중국어판의 온라인 홈페이지에서 수집하였다(황은하와 비립, 2023).

2023년 9월 15일을 기준으로 각 매체의 중국어판 홈페이지에서 검색 가능한 기사 목록 1~100면을 순차적으로 열람하여 기사 제목, 본문, 바이라인, 게시일자를 수집하고 1차 자료를 구성하였다. 이 가운데 각 매체가 부여한 섹션 또는 태그 정보를 기준으로 사회, 경제, 문화(예능 한류 포함), 정치(한반도 관련 포함)에 해당하는 기사를 추출한 뒤, 본 연구의 비교 대상인 사회 경제 문화 세 분야만을 남기고 정치 분야는 제외하였다. 이어서 제목 본문 바이라인 게시일자가 모두 갖추어져 있는지를 확인하고, 중복 게재된 기사와 기사 내 출처 또는 바이라인이 불분명하거나 통신사 전제로 판단되는 기사를 제외하였다. 이 절차를 거쳐 확정된 자료에서 분야별로 8만 문장씩 무작위 추출하였다. 이 텍스트는 중국어권 독자를 수용자로 상정해 편집 공개되었으며, AI 학습용 번역문과 생산 조건을 달리한다. 바이라인 정보가 있는 조선일보 중앙일보 기사 10,808건을 조사한 결과, 한국 매체를 출처로 하는 기사가 전체의 98.2%를 차지하였다.⁶⁾

3.1.3 비번역 중국어

비번역 중국어는 칭화대학(清华大学)이 공개한 중문 뉴스 말뭉치 THUCNews의 신량망(新浪新闻) 기사 가운데 사회 경제 문화에 해당하는 자

5) 아주일보는 다른 세 매체에 비해 검색 가능한 기사 수가 적었다. 수집 시점의 중국어판 홈페이지에서 주로 2022년까지의 기사만 검색되었기 때문이다. 이에 따라 일간지 번역문의 생산 시기 분포는 매체에 따라 다르다.

6) 국제 뉴스의 언어 간 이동에는 축자 번역 외에 발췌 요약 재배열이 포함될 수 있다(Bielsa & Bassnett, 2009). 바이라인 분석에서 한국 기자의 실명이 표기된 비율은 매체와 분야에 따라 18.6~99.7%로 달랐으나, 기관명 표기까지 포함하면 한국 매체를 출처로 하는 기사는 여섯 하위말뭉치에서 94.0~100.0%(전체 98.2%)였다. 비한국 출처는 中国新闻网 전체 28건(0.3%)이었다. 따라서 이 말뭉치는 대체로 한국어 원기사의 존재를 전제할 수 있는 텍스트로 구성되어 있다.

료에서 분야별 8만 문장씩 추출하였다(清华大学自然语言处理与社会人文计算实验室, 2016). THUCNews는 중문 텍스트 분류 도구 THUCTC(孙茂松 등, 2016)의 배포 페이지에서 함께 공개된 뉴스 말뭉치이다. 이 말뭉치는 문서 단위 자료이므로 종지 부호(。 ! ?)를 기준으로 문장을 분할한 뒤 표집하였으며, 같은 기준을 세 하위말뭉치에 적용하였다.⁸⁾

3.1.4 비교 가능성: 통제 변수와 한계

비교말뭉치 설계에서 변수는 세 부류로 나뉜다. 첫째, 통제 대상이 아닌 변수이다. 어휘 다양성, 문체, 통사 구조는 본 연구가 측정하려는 대상, 곧 종속변수이므로 사전에 일치시킬 수 없다. 이를 통제하면 관찰할 대상 자체가 사라진다. 둘째, 통제한 변수이다. 본 연구는 장르(신문 기사, 출처 확인으로 입증), 주제 영역(사회 경제 문화), 분야별 문장 수를 일치시켰다. 이는 선행 비교말뭉치 연구의 통제 수준에 부합한다. ZCTC와 LCMC(Xiao et al., 2010)는 장르 구성과 표본 비율을 일치시켰으나 표집 시기에는 약 10년의 차이가 있다. 셋째, 통제하지 못한 변수이다. 수집 시기(AI용 2020년, 일간지 번역문 2014~2023년, 비번역 기사 2005~2011년), 텍스트의 기능, 상정된 수용자, 매체와 편집 관행이 여기에 속한다. 문장 수를 표집 단위로 삼은 결과 단어 수에는 최대 10.8%의 편차가 있으며, 이는 3.3.1의 표준화로 보정한다. 이에 따라 본 연구는 시기와 하위 주제에 민감한 내용어 및 연어 기반 관찰을 보조 자료로만 제시하고, 결론은 기능어와 구조 지표를 근거로 구성한다. 문화 분야를 제외한 재산출(4.2.2)은 이 원칙의 타당성을 점검하기 위한 것이다.

통제한 조건과 통제 불가 조건, 비교 조건 그리고 표집 결과의 규모를 하위말뭉치별로 정리하면 <표 2>와 같다.

-
- 7) 중국어에서 문장부호를 표점(标点)이라 부르며, 이 가운데 문장의 끝을 표시하는 마침표 느낌표 물음표(。 ! ?)를 종지 부호라 한다. 반점 쌍반점 쌍점 등 문장 내부의 표점 부호까지 경계로 삼아 구획한 단위를 본고에서는 표점 단위로 부른다. 표점 단위는 절(clause)이 아니라 부호로 나뉜 형식적 단위이다.
- 8) 인용부호 안의 종지 부호는 경계로 인정하지 않았고, 기사 제목과 목록형 항목은 분할 대상에서 제외하였다. 영문 약어와 소수점의 마침표는 반각이므로 경계로 계수되지 않는다.

표 2
비교말뭉치의 구성

구분	세부 구분	AI 학습용 번역문	일간지 번역문	비번역 중국어
통제	매체 유형	신문 기사	신문 기사	신문 기사
	주제 영역	사회 경제 문화	사회 경제 문화	사회 경제 문화
	분야별 문장 수	80,000	80,000	80,000
	문장 수 합계	240,000	240,000	240,000
비교 조건	생산 목적	기계번역 학습데이터	중국어권 독자 대상 보도	중국 독자 대상 보도
	상정된 수용자	대규모 언어모델(인간 독자 부재)	일반 독자	일반 독자
	편집 단위	문장	기사	기사
	품질 검수 방식	전체 5% 표본 블랙박스 평가	각 매체 편집국의 편집, 게재 절차	각 매체 편집국의 편집, 게재 절차
통제 불가	생산 시기	2020년	2014~2023년	2005~2011년
	출처	AI 허브 「한국어-중국어 번역 말뭉치 2(사회과학)」	중앙일보, 동아일보, 조선일보, 아주일보 중국어판	THUCNews (신량망 新浪新闻)
표집 결과 규모	단어(사회/경제/문화)	1,992,128 / 1,911,467 / 1,848,735	1,828,180 / 1,831,300 / 1,533,606	1,691,980 / 1,962,478 / 1,948,472
	단어 수(합계)	5,752,330	5,193,086	5,602,930
	토큰 수(합계)	6,708,185	6,191,910	6,650,263

3.2 분석 도구와 절차

전처리와 언어 분석에는 N-LTP(Neural Language Technology Platform)를 사용하였다(Che et al., 2021)⁹⁾. 세 하위말뭉치에 동일한 모델과 설정을 적용

9) N-LTP는 하얼빈공업대학(哈尔滨工业大学) 사회계산 정보검색연구센터가 개발 공

하여 단어 분리, 품사 부착, 의존구문 분석을 수행하여 품사별 타입과 토큰 수, 의존관계 유형별 빈도와 어휘 다양성, 문장 및 표점 단위의 길이 지표를 산출하였다. N-LTP의 품사 체계는 소분류까지 27종으로 이루어지며, 이 가운데 접두사와 접미사는 토큰 수가 적어(공통 표본 크기 각각 115, 3,607) 해석에서 제외하였다.

3.3 측정 지표와 통계 처리

3.3.1 어휘 지표: 어휘 다양성

어휘 층위는 품사 범주별 타입 수를 해당 범주의 토큰 수로 나눈 타입-토큰 비율(Type-Token Ratio, 이하 TTR)로 측정한다. 품사별 TTR을 사용한 것은 다양성의 차이가 나타나는 범주를 식별하기 위해서이다. TTR은 말뭉치 규모가 커질수록 낮아지므로, 각 품사에서 세 말뭉치의 최소 토큰 수를 공통 표본 크기 m 으로 정하고, m 개를 무작위로 뽑았을 때의 기대 타입 수 $E(Sm)$ 을 산출한다(Heck et al., 1975; Hurlbert, 1971).

$$E(Sm) = \sum_i [1 - C(N - fi, m) / C(N, m)] \quad (1)$$

여기서 N 은 해당 범주의 전체 토큰 수, fi 는 타입 i 의 빈도이다. $E(Sm)/m$ 을 이하 ‘표준화 TTR’로 부른다. 본 연구의 표준화 TTR은 공통 표본 크기에서 기대되는 타입 수를 추정해 값이므로 STTR¹⁰⁾과는 표준화의 방식이 다르다. 따라서 STTR을 사용한 선행연구의 수치와 직접 비교하지 않는다. 표준오차는 1,000회 몬테카를로 부분표집으로 추정하였다.¹¹⁾

개한 도구로, 본 연구가 이 도구를 택한 것은 분사 품사 부착 의존구문 분석을 하나의 파이프라인에서 동일한 토큰화 기준으로 처리할 수 있고, 세 하위말뭉치에 같은 모델과 설정을 적용해 비교 가능성을 확보할 수 있기 때문이다.

- 10) 각주 2에서 기술한 것처럼 기존의 STTR은 텍스트를 일정 길이의 구간으로 나누어 각 구간에서 타입-토큰 비율을 구한 뒤 그 값들을 평균한 지표이다.
- 11) 반복 횟수는 표준오차 추정의 상대 오차가 약 2.2%가 되도록 정하였다. 이 구간은 부분표집의 변이만 반영하며, 말뭉치 자체가 모집단에서 표집된 데서 오는 변이는 포함하지 않는다.

결과는 비번역 중국어 대비 비율과 95% 신뢰구간으로 제시한다. 비번역 기사 대비 비의 95% 신뢰구간이 사전에 설정한 실질적 차이 한계 [0.900, 1.111] 안에 온전히 포함될 때 ‘동등’으로 표시하였다. 이 기준은 탐색적 비교를 위한 기술적 판정 규칙이며, 개별 범주의 판정보다 세 말뭉치의 상대적 순서에 해석의 무게를 둔다.

3.3.2 통사 지표

이하에서 사용하는 분석 용어를 먼저 밝힌다. 의존관계(dependency relation)는 문장 안에서 한 어휘가 다른 어휘에 의존하는 관계이며, 의존구문 분석은 이 관계의 유형과 방향을 부여하는 절차이다. 핵심관계(HED)는 문장의 중심 술어를 문장 전체에 연결하는 관계로 문장당 한 번만 출현한다. 좌측 부가(左附加)와 우측 부가(右附加)는 접속사 조사처럼 독립된 성분을 이루지 않는 요소가 핵어의 왼쪽 또는 오른쪽에 붙어 핵어를 수식 보충하는 관계를 가리킨다. 전치 목적어(前置宾语)는 ‘把’자문 등에서 술어 앞으로 옮겨진 목적어이고, 겸어(兼语)는 앞 술어의 목적어가 뒤 술어의 주어를 겸하는 구조로, ‘让他走(그를 가게 하다)’의 ‘他’가 이에 해당한다. 술어-보어(动补)는 술어에 결과나 정도를 나타내는 보어가 결합한 구조이다. 그 밖의 주술(主谓) 동사-목적어(动宾) 관형어-중심어(定中) 부사어-중심어(状中) 병렬(并列) 관계는 한국어 문법의 대응 개념으로 이해할 수 있다. 이밖에, 계사(系辞)는 ‘甲是乙(갑은 을이다)’의 ‘是’처럼 주어와 술어를 이어 주는 요소이며, N-LTP는 이를 일반 동사와 구별해 별도 품사로 분류한다.

통사 층위는 문장, 표점 단위, 구 의존관계 수준에서 측정한다. 문장 길이는 단어 수로 산출하고, 길이 분포는 선행 연구와의 비교를 위해 글자 수로 제시한다. 표점 단위는 전각 반각 반점(, /), 쌍반점(: /;), 쌍점(: /:)을 경계로 분할한 조작적 단위이며, 절과 동일시하지 않는다.¹²⁾ 문장부호에 따른 결과를 보완하기 위해 문장당 주술관계 병렬관계 술어-보어 구조의 수도 함께 산출하였다.

12) 중국어에서 반점 쌍반점 쌍점은 절 경계 외에도 병렬 명사구, 긴 주어, 부사어 뒤에 사용될 수 있다(国家质量监督检验检疫总局과 国家标准化管理委员会, 2011). 표점 단위는 샤오와 다이(2014)가 sentence segment로 부른 단위와 같은 개념이다.

구 수준에서는 N-LTP가 부착한 의존관계 유형별 분포 비율과 관계별 어휘 다양성을 산출하였다. 분포 비율은 문장별로 계산하고, 두 말뭉치의 평균 차이를 합동 표준편차로 나눈 표준화 효과크기인 d (Cohen, 1988)를 함께 제시하였다. 다만 같은 기사에 속한 문장의 군집성과 문장 길이 분포의 차이를 고려하여, d 는 세 말뭉치의 차이 방향과 상대적 크기를 읽는 보조 지표로만 해석하며 모집단 수준의 추론으로 확장하지 않는다. 또한 어휘와 통사 지표의 척도가 다르므로, 4.5의 종합에서는 층위 간 효과크기를 직접 비교하지 않고 각 층위 안에서 세 말뭉치의 상대적 순서만을 논거로 삼는다.

의존관계의 분포 비율은 문장별로 산출한 뒤 평균하므로 문장 길이 분포에 민감하다. 핵심관계는 문장당 반드시 한 번만 출현하고 문장부호 관계(标点)는 표점 부호의 수에 직결되므로, 두 관계의 비율은 문장 길이의 역수에 직접 좌우되어 분포 비율 해석에서 제외한다. 중앙값이 0인 저빈도 관계도 함께 제외하였다. 나머지 관계에 대해서도 문장 길이의 영향이 완전히 제거되지는 않는다. 명사구의 확장에 관여하는 관계는 긴 문장에서 실현되므로 짧은 문장의 비중이 높은 말뭉치에서 그 값이 낮게 산출될 여지가 있다. 본 연구는 이 영향의 크기를 분리하지 못하였으므로, 해당 지표의 격차를 해석할 때 이 제약을 함께 밝힌다.

3.4 판정 논리

본 연구의 3원 비교는 두 단계로 판정한다. 첫째, 두 번역문 집단이 공통적으로 비번역 중국어와 갈리는 지점은 번역문 일반에서 관찰되는 경향의 후보로 본다. 둘째, 두 번역문 집단 사이가 갈리는 지점은 생산 조건과 관련된 변이의 후보이다. 각 지표에 대해 비번역 중국어 대비 비의 95% 신뢰구간과 등가 한계 [0.900, 1.111]을 적용하여 두 단계를 차례로 판정한다. 다만 둘째 단계에서 관찰된 차이를 특정 생산 조건의 효과로 단정하지 않는다. 이는 3.1.4에서 밝힌, 통제되지 않은 변수(생산 목적 수용자 편집 단위 검수 방식)가 동시에 존재하기 때문이다.

4. 분석 결과

분석 결과는 자료 구축 절차와 어휘 통사 세 층위로 나누어 다음과 같이 정리할 수 있다.

4.1 자료 구축 절차에서 비롯된 표기상의 특징

본 절에서 다루는 두 현상은 4.2 이하의 층위와 성격이 다르다. 번역문의 어휘와 통사가 목표어의 그것과 다른 것은 번역 행위와 관련되지만, 문장부호의 표기와 문장 길이의 분포는 번역자의 선택이 아니라 데이터를 구축하는 절차에서 결정된 것이다. 원천 자료를 어떤 기준으로 선별 정제하고 산출물을 어떤 형식으로 저장하는가에 따른 결과이며, 따라서 번역 보편소의 틀로는 설명되지 않는다. 그럼에도 이들을 별도의 절로 앞세우는 까닭은, 이 자료가 학습 자료로서 지니는 성격을 규정하는 데에서 차지하는 비중이 작지 않기 때문이다.

4.1.1 문장부호의 전각/반각

자료의 표기 상태를 점검하는 과정에서 세 말뭉치 사이에 문장부호의 전각 반각 사용이 일치하지 않음이 확인되었다. 그 결과는 <표 3>과 같다.

표 3
문장부호의 전각/반각 사용 양상(세 분야 합계)

말뭉치	전각 반점(,)	반각 반점(,)	전각 사용률	인용부호
AI 학습용 번역문	0	388,737	0.0%	반각" 사용
일간지 번역문	324,441	1,038	99.7%	전각“ ” 사용
비번역 중국어	504,107	1,051	99.8%	전각“ ” 사용

중국의 국가표준 『표점부호 용법(标点符号用法)』은 문장부호를 전각으로 표기하도록 규정하고 있으며(国家质量监督检验检疫总局과 国家标准化管理委员会, 2011), 반각 부호의 사용은 규범 일탈에 해당한다. 표기 규범은 목표어 공동체의 관습 가운데 가장 명시적이고 검증이 쉬운 층위이다. 번역 보

편소 논의가 주로 어휘와 통사에 집중되어 온 데 비해, 표기 층위는 번역문의 생산 절차가 남긴 흔적이 가장 직접적으로 드러나는 부분이기도 하다.

이 표기 불일치는 개별 번역자의 언어 능력 문제가 아니라 데이터 구축 절차의 후처리 단계에서 비롯된 것으로 보이며, 정규표현식 수준의 자동 검사로 전량 탐지 수정이 가능한 종류의 오류이다. 다만, 이 데이터로 학습한 모델이 중국어 텍스트에서는 쓰이지 않는 반각 문장부호를 생성할 가능성, 그리고 그렇게 생성된 텍스트가 다시 학습 자료로 유통될 가능성도 고려하면 짚고 넘어갈 필요가 있다.

방법론적으로도 이 발견은 중요하다. 전각 부호만을 표점 단위의 경계로 인식하는 절차를 적용할 경우 AI용 번역문의 경계는 대부분 탐지되지 않으며, 문장당 표점 단위 수가 실제보다 과소 추정된다. 본 연구의 예비 분석에서도 전각 부호만을 계수하였을 때 AI용 번역문의 문장당 표점 단위 수는 1.08개로 산출되어, 일간지 번역문 2.44개, 비번역 기사 3.24개와 극단적인 차이를 보였다. 4.3의 분석에서는 전각과 반각을 모두 경계로 인정하여 수치를 산출하였다.

4.1.2 문장 길이 분포와 짧은 문장의 소거

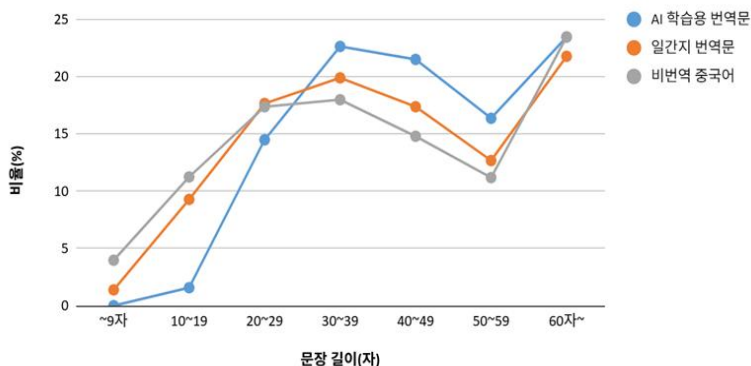
문장 길이의 구간별 분포는 <표 4>와 <그림 1>에 제시한다.

표 4

문장 길이 구간별 분포(세 분야 합계, 단위: 문장 수(%))

문장길이	AI 학습용 번역문	일간지 번역문	비번역 중국어
9자이하	6 (0.00)	3,315 (1.38)	9,517 (3.97)
10~19자	3,760 (1.57)	22,266 (9.28)	26,984 (11.24)
20~29자	34,754 (14.48)	42,350 (17.65)	41,690 (17.37)
30~39자	54,314 (22.63)	47,744 (19.89)	43,164 (17.98)
40~49자	51,612 (21.50)	41,679 (17.37)	35,491 (14.79)
50~59자	39,267 (16.36)	30,400 (12.67)	26,831 (11.18)
60자이상	56,287 (23.45)	52,246 (21.77)	56,323 (23.47)
합계	240,000 (100.00)	240,000 (100.00)	240,000 (100.00)

그림 1
문장 길이 구간별 분포 곡선



분포 곡선의 전체적인 모양은 세 말뭉치가 유사하나, 짧은 문장의 비중에서 뚜렷한 차이가 나타난다. 19자 이하의 문장이 비번역 중국어에서는 15.21%, 일간지 번역문에서는 10.66%를 차지하는 데 비해 AI용에서는 1.57%에 그친다. 특히 9자 이하 문장은 비번역 중국어 9,517문장, 일간지 번역문 3,315문장인 데 비해 AI용은 6문장에 불과하다. 그 결과 AI용 번역문의 분포는 30~49자 구간에 집중되어(44.14%), 비번역 중국어(32.77%)보다 11% 포인트 이상 높다.

AI 허브 데이터 구축 가이드라인은 한국어 원문을 평균 15어절, 5~30어절 범위로 정제하도록 규정하고 있다. 즉 극단적으로 짧은 문장은 애초에 원천 자료에서 배제되었다. 이 분포 차이는 번역 과정의 산물이 아니라 데이터 설계의 산물이며, 그런 뜻에서 번역투로 읽을 수 없다. 그러나 학습 자료로서의 성격을 논할 때에는 사정이 다르다.

앞서 본 짧은 문장의 결여는 학습 자료로서의 성격을 논할 때 다른 의미를 지닌다. 곧 이 말뭉치는 중국어 신문 문장의 한 유형 전체를 결여한 채 구성되어 있다. 학습 데이터의 분포적 특성이 모델 출력에 전이된다는 점이 여러 연구에서 확인되어 온 만큼(Daems et al., 2017; Toral, 2019; Vanmassenhove et al., 2021), 이 자료로 학습한 모델은 짧은 문장의 생성과 처리에서 편향을 보일 개연성이 있다. 다만 본 연구는 모델의 출력을 측정

하지 않았으므로 이는 자료의 구성에서 도출되는 예측이며, 그 검증은 별도의 과제이다.

한편 이 분포 차이는 이하의 분석에 방법론적 제약을 부과한다. 4.3에서 산출하는 구 의존관계의 분포 비율은 문장을 단위로 평균한 값이므로 문장 길이 분포에 민감하다.

4.2 어휘 층위: 품사별 어휘 다양성

4.2.1 표준화의 효과

세 하위말뭉치의 규모 차이가 TTR 결과를 왜곡하며, 왜곡의 크기가 작지 않다. 일반명사의 토큰 수는 AI용 1,460,939, 일간지 번역문 1,302,824, 비번역 1,297,011로 비번역 중국어가 최소이며, 원 TTR은 AI용 .0312, 일간지 번역문 .0388, 비번역 중국어 .0415이다. 공통 표본 크기를 비번역 중국어의 토큰 수로 맞추어 표준화하면 각각 .0331, .0389, .0415가 되어, 토큰 수가 가장 큰 AI용의 값만 6% 상승하고 비번역 중국어와의 격차가 줄어든다. 외래어처럼 토큰 수가 적고 말뭉치 간 편차가 큰 범주에서는 왜곡이 훨씬 커서, AI용의 값이 .2417에서 .4484로 배 가까이 달라진다.

규모 보정의 효과는 특히 일간지 번역문과 비번역 중국어의 관계에서 결정적이다. 원 TTR로는 일간지 번역문이 여러 범주에서 비번역 중국어에 근접하는 것처럼 보이지만, 이는 일간지 번역문이 가장 작은 말뭉치이기 때문에 생기는 상향 편향의 결과이다. 따라서 이하의 분석은 모두 표준화 값에 근거한다.

4.2.2 세 말뭉치의 3단 위계

표준화된 TTR을 품사별로 정리하면 <표 5>와 같다.

-
- 13) 판정은 비번역 중국어 대비 비의 95% 신뢰구간과 등가 한계 [0.900, 1.111]의 관계에 따른 것이다. 신뢰구간이 하한 아래에 온전히 놓이면 ‘낮음’, 실질적 차이 한계 안에 온전히 들면 ‘동등’, 상한 위에 온전히 놓이면 ‘높음’, 겹치되 포함되지 않으면 ‘보류’로 표시하였다.

표 5

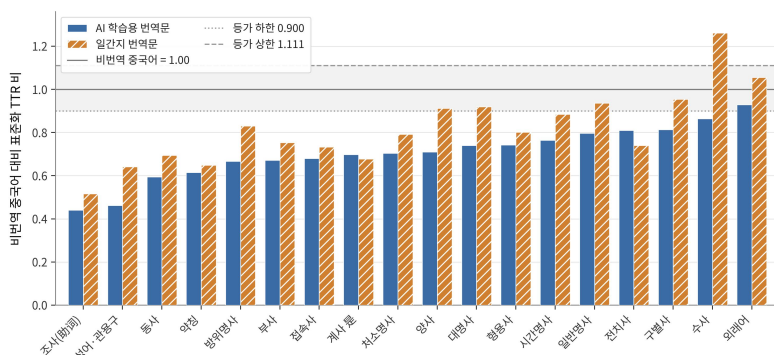
품사별 어휘 다양성: 공통 표본 크기에서 산출한 표준화 TTR⁽¹³⁾

품사		공통표본 m	원 TTR			표준화 TTR 및 판정		
대분류	세분류		AI용	일간지	비번역	AI용/비번역 [95% CI]	일간지/비번역 [95% CI]	AI용/ 일간지
명사	일반명사	1,297,011	.0331	.0389	.0415	0.797 낮음 [0.795, 0.798]	0.936 동등 [0.936, 0.937]	0.851
	방위명사	117,479	.00472	.00587	.00708	0.667 낮음 [0.655, 0.679]	0.830 낮음 [0.822, 0.838]	0.803
	시간명사	136,300	.0147	.0170	.0192	0.764 낮음 [0.755, 0.772]	0.884 낮음 [0.868, 0.899]	0.864
	처소명사	23,475	.0175	.0196	.0248	0.704 낮음 [0.685, 0.723]	0.792 낮음 [0.769, 0.815]	0.890
고유명사	인명	94,946	.274	.241	.249	1.100 동등 [1.093, 1.108]	0.969 동등 [0.964, 0.975]	1.136
	기관명	1,833	.0848	.0796	.0966	0.879 보류 [0.807, 0.950]	0.825 보류 [0.748, 0.901]	1.065
	지명	69,107	.0850	.0734	.136	0.624 낮음 [0.620, 0.628]	0.539 낮음 [0.530, 0.548]	1.157
	기타 고유명사	42,213	.342	.303	.238	1.436 높음 [1.428, 1.444]	1.272 높음 [1.263, 1.282]	1.129
대명사	대명사	131,234	.00486	.00604	.00656	0.740 낮음 [0.718, 0.762]	0.920 보류 [0.894, 0.945]	0.804
수량사	수사	189,664	.0817	.119	.0945	0.864 낮음 [0.859, 0.870]	1.261 높음 [1.251, 1.270]	0.686
	양사	124,063	.00528	.00678	.00743	0.710 낮음 [0.695, 0.725]	0.912 보류 [0.885, 0.939]	0.778
동사	동사	1,239,946	.0114	.0134	.0192	0.594 낮음 [0.591, 0.597]	0.695 낮음 [0.694, 0.697]	0.854
	계사	3,592	.163	.159	.234	0.698 낮음 [0.678, 0.718]	0.678 낮음 [0.654, 0.703]	1.029
형용사	형용사	176,024	.0209	.0226	.0281	0.743 낮음 [0.734, 0.753]	0.802 낮음 [0.795, 0.809]	0.927
	구별사	47,272	.0448	.0526	.0551	0.813 낮음 [0.801, 0.826]	0.954 동등 [0.946, 0.962]	0.852
부사	부사	283,970	.00586	.00658	.00874	0.671 낮음 [0.660, 0.681]	0.753 낮음 [0.742, 0.763]	0.891
허사	전치사	253,001	.00127	.00116	.00157	0.810 낮음 [0.783, 0.838]	0.740 낮음 [0.722, 0.757]	1.096

	접속사	148,454	.00181	.00195	.00267	0.680 낮음 [0.653, 0.707]	0.733 낮음 [0.727, 0.738]	0.928
	조사	413,237	.000246	.000287	.000557	0.441 낮음 [0.408, 0.475]	0.516 낮음 [0.499, 0.533]	0.855
기타	접두사	115	.187	.155	.148	1.265 보류 [0.893, 1.636]	1.048 보류 [0.723, 1.373]	1.207
	접미사	3,607	.00248	.00271	.00416	0.596 낮음 [0.446, 0.746]	0.651 낮음 [0.477, 0.825]	0.915
	성어, 관용구	22,290	.146	.203	.317	0.462 낮음 [0.456, 0.468]	0.642 낮음 [0.635, 0.648]	0.720
	약칭	25,162	.0735	.0776	.120	0.615 낮음 [0.605, 0.624]	0.649 낮음 [0.633, 0.664]	0.948
	외래어	2,163	.448	.509	.483	0.929 보류 [0.894, 0.964]	1.055 동등 [1.020, 1.090]	0.881

두 번역문의 값을 비번역 중국어 대비 비율로 바꾸어 품사별로 나타내면 <그림 2>와 같다.

그림 2
주요 품사별 표준화 TTR의 비번역 중국어 대비 비율



전체적인 양상은 명확하다. 24개 품사 범주 가운데 21개에서 AI 학습용 번역문의 어휘 다양성이 비번역 중국어보다 낮았으며, 이 가운데 19개는 비율의 신뢰구간이 등가 한계의 하한 아래에 온전히 놓여 ‘낮음’으로 판정되

었다(기관명과 외래어는 판정 보류). 주제 의존성이 큰 고유명사류와 추정이 불안정한 접사류를 제외할 때 격차가 가장 큰 범주는 조사(助詞, 비번역 중국어의 0.44배), 성어 관용구(0.46배), 동사(0.59배), 약칭(0.62배), 방위명사(0.67배), 부사(0.67배), 접속사(0.68배)이다. AI용이 비번역 중국어보다 높게 나타난 세 범주는 인명, 기타 고유명사, 접두사인테, 앞의 둘은 뒤에서 보듯 원천 텍스트의 주제 분포를 반영하며 접두사는 추정이 불안정하다.

이에 더해, 새롭게 드러난 것은 일간지 번역문의 표준화 TTR 역시 대부분의 범주에서 비번역 중국어보다 낮은 것으로 판정된다는 점이다. 동등성 판정 기준을 통과한 범주는 일반명사(0.936배), 구별사(0.954배), 인명, 외래어의 네 가지에 그치며, 동사 0.70배, 형용사 0.80배, 성어 관용구 0.64배, 조사 0.52배로 비번역 중국어와의 거리가 뚜렷하다.

즉 세 말뭉치는 ‘AI용 번역문 < 일간지 번역문 < 비번역 중국어’라는 3단 위계를 이룬다. 두 번역문 사이의 직접 비교에서도 같은 순서가 확인된다. 주제에 좌우되는 고유명사 네 범주와 추정이 불안정한 접사 두 범주를 제외한 18개 범주 가운데 16개에서 AI용이 일간지 번역문보다 낮았고(일반명사 0.85배, 동사 0.85배, 성어 관용구 0.72배), 역전된 것은 계사 是(1.03배)와 전치사(1.10배) 두 가지뿐이다¹⁴⁾.

이 결과는 두 가지 함의를 지닌다. 첫째, 어휘 다양성의 감소는 AI 학습용 번역문에만 나타나는 특성이 아니다. 앞서 보았듯 일간지 번역문 역시 대부분의 범주에서 비번역 중국어보다 낮다. 이는 베이커(1993)가 제시하고

14) 주제 구성이 결과에 미치는 영향을 가늠하기 위해 문화 분야를 제외하고 사회경제 두 분야만으로 표준화 TTR을 재산출하였다. 공통 표본 크기가 줄어들므로 세 말뭉치 전부를 새로운 m으로 다시 계산하였다. 그 결과 주제에 좌우되지 않는 다섯 범주에서 AI용과 일간지 번역문의 비는 조사 0.855에서 0.772로, 성어 관용구 0.720에서 0.637로, 부사 0.891에서 0.818로, 접속사 0.928에서 0.848로 낮아졌다. 전체 표본에서 유일하게 순서가 역전되었던 전치사도 1.096에서 0.951이 되어 다른 범주와 같은 방향을 보였다. 24개 범주 전체에서도 AI용이 일간지 번역문보다 낮은 범주가 17개에서 20개로 늘었다. 일간지 번역문과 비번역 중국어의 비 또한 함께 커져(조사 0.516→0.586, 부사 0.753→0.768) 두 번역문이 모두 비번역 중국어 쪽으로 이동하였다. 곧 문화 분야를 제외해도 세 말뭉치의 상대적 순서와 격차의 방향은 유지되므로, 본 절의 결과는 문화 분야 자료의 주제 구성에 좌우된 것이 아니다.

라비오사(1998)가 영어 자료로 실증한 단순화 가설과 부합한다. 다만 격차의 분포는 단순화만으로 설명되지 않는다. 비번역 중국어 대비 비율이 가장 낮은 두 범주가 조사(0.44배)와 성어 관용구(0.46배)라는 사실은 서로 다른 두 경향을 가리킨다. 조사의 다양도 저하는 문법 표지의 목록을 좁게 쓴다는 뜻이므로 단순화에 해당하지만, 성어 관용구의 저하는 목표어에 이미 마련되어 있는 관습적 표현 가운데 소수의 전형만을 되풀이해 쓴다는 뜻이므로 정규화(normalisation)의 지표로 읽는 편이 적절하다. 둘째, 그럼에도 AI 학습용 번역문은 일간지 번역문보다 한 단계 더 낮은 위치에 있으며, 그 격차는 분석상 핵심으로 삼은 다수 범주에서 일관되게 나타난다. 이 두 번째 층위의 차이는 두 번역문의 생산 조건과 관련된 것으로 보이며, 서로 배타적이지 않은 다음 세 가지 설명이 가능하다. (i) 상정된 수용자의 차이 — AI 학습용 번역문은 일반 독자에게 직접 게재 배포되는 기사문과 달리, 우선적으로 언어모델 학습을 위한 데이터셋 구성에 사용되도록 생산 정제된 텍스트이다. (ii) 편집 단위의 차이 — 문장 단위로 정제 저장되는 구성은 기사 전체를 단위로 하는 문체 조정과 문맥 결속 장치의 작동 범위를 제한한다. (iii) 품질 검수 기준의 차이 — 구축 절차의 검수는 출발 텍스트와의 대응 정확성을 중심으로 설계되어 있으며 목표어 문체의 다양성은 평가 항목이 아니다(한국지능정보사회진흥원, 2021a). 기계번역과 후편집이 어휘 다양성을 낮춘다는 후편집투 논의(Toral, 2019)는 이 세 설명과 양립 가능한 넷째 요인이나, 본 연구는 일간지 번역문의 생산 방식을 확인하지 못하였으므로 이를 단독 설명으로 제시하지 않는다. 세 말뭉치의 분포만으로는 네 요인의 몫을 분리할 수 없으며, 이를 구분하려면 출발 텍스트를 포함한 병렬말뭉치 분석이 필요하다.

한편, 이 결과는 영중 번역에서 번역 중국어와 원어인 중국어의 표준화 TTR에 유의한 차이가 없다고 보고한 샤오(2010)와는 차이를 보인다. 본 연구는 한중 번역에서 품사별로 측정할 경우 두 유형의 번역문 모두 원어인 중국어보다 일관되게 낮다는 상반된 결과를 얻었다. 이 불일치는 두 가지 요인으로 설명될 수 있다. 첫째, 지표의 차이이다. 샤오의 STTR은 텍스트 전체를 1,000단어 단위로 나누어 산출한 전역 지표¹⁵⁾인 반면, 본 연구는 품

15) 전역 지표란 개별 문장이 아니라 말뭉치 전체를 하나의 단위로 삼아 산출하는

사 범주별로 산출한다. 전체 TTR에서 상쇄되는 범주별 차이가 분해하면 드러날 수 있다. 둘째, 출발어의 차이이다. 한국어는 관형절과 연결어미로 정보를 응집시키는 반면 영어는 관계절과 접속사를 쓰므로, 중국어로 옮길 때 재구조화의 부담이 서로 다르다. 또한 한국어와 중국어는 한자어를 공유하므로 어휘 층위의 간접 양상도 영중 언어쌍과 같지 않다.

내용어와 기능어를 나누어 보면 해석의 신뢰도에 차이가 있다. 고유명사 네 범주에서는 AI용과 일간지 번역문이 오히려 비번역 중국어보다 높게 나타나는데(기타 고유명사 1.44배, 1.27배), 이는 한국 신문 기사를 원천으로 하는 두 번역문에 한국의 인명 기관명이 대량으로 등장하고 이들이 대개 저빈도 타입으로 산입되기 때문이다. 즉 이 결과는 번역 과정의 산물이 아니라 원천 텍스트의 주제 분포를 반영한다. 반면 조사, 접속사, 전치사, 성어 관용구 등 기능어와 관습적 표현 자원은 텍스트의 주제에 좌우되지 않으므로, 3.1.4의 해석 원칙에 따라 이 다섯 범주를 판정의 근거로 삼는다. 이 다섯 범주에서 AI 학습용 번역문은 모두 비번역 중국어보다 낮았고(0.44~0.81배), 일간지 번역문 역시 모두 낮았으며(0.52~0.75배), 두 번역문의 직접 비교에서도 전치사를 제외한 네 범주에서 AI용이 일간지 번역문보다 낮았다. 24개 범주 전체를 대상으로 한 앞의 서술은 기술적 요약이며, 3단 위계 주장의 근거는 이 다섯 범주에 둔다.

4.3 통사 층위

4.3.1 표점 단위와 술어 기반 교차 검증

평균 문장 길이(단어 기준)는 AI용 23.97, 일간지 번역문 21.64, 비번역 중국어 23.35로 세 말뭉치의 차이가 크지 않다(<표 6>). 문장 길이의 평균값만으로는 세 말뭉치를 변별하기 어렵다는 점은 샤오(2010)가 영중 번역에서 평균 문장 길이에 유의차가 없다고 보고한 것과 일치한다.

문장 표점 단위의 복잡도 지표는 <표 6>과 같다.

지표를 말한다. 품사별 표준화 TTR이 이에 해당하며, 문장별로 산출한 뒤 평균하는 의존관계 분포 비율과 구별된다.

표 6

문장 및 표점 단위 복잡도 지표(세 분야 평균)¹⁶⁾

지표	AI 학습용번역문	일간지 번역문	비번역 중국어
문장당 단어 수(MLS)	23.97	21.64	23.35
문장당 표점단위 수	2.72	2.45	3.24
표점단위당 단어 수	8.81	8.84	7.20
문장당 주술관계 수	1.91	1.80	2.34
문장당 병렬관계 수	1.74	1.61	2.24
문장당 술어보어 수	0.33	0.33	0.47
문장당 문장부호 관계 수	3.97	4.15	4.36

문장당 표점 단위 수는 비번역 중국어 3.24개, AI용 2.72개, 일간지 번역문 2.45개 순이다. 두 번역문 모두 비번역 중국어보다 단위의 수가 적으며, 그 결과 표점 단위당 평균 길이는 비번역 중국어 7.20단어에 비해 AI용 8.81단어, 일간지 번역문 8.84단어로 약 1.22배 길다. 두 번역문 사이에는 단위 길이의 차이가 거의 없다.

3.3.2에서 밝힌 대로 표점 부호에 기반한 이 지표는 절이 아니라 표점 단위를 세는 것이므로, 술어 기반 지표로 교차 검증하였다. 문장당 주술관계 수는 비번역 중국어 2.34, AI용 1.91, 일간지 번역문 1.80이고, 문장당 병렬관계 수는 비번역 중국어 2.24, AI용 1.74, 일간지 번역문 1.61이며, 문장당 술어-보어 구조 수는 비번역 중국어 0.47, AI용 0.33, 일간지 번역문 0.33이다. 세 지표 모두 표점 단위 수와 동일하게 ‘비번역 중국어 > AI용 번역문 > 일간지 번역문’의 순서를 보인다. 표점 부호에 의존하지 않는 지표가 같은 결론을 지지한다는 점에서, 두 번역문이 비번역 중국어보다 술어를 중심으로 한 절 연쇄를 적게 만든다는 관찰은 표기 방식에 의한 것은 아니라고 판단된다.

16) 표점 단위는 전각과 반각 반점(·, /), 쌍반점(; /), 쌍점(: /)을 모두 경계로 인정하여 산출하였다. 길이 단위는 단어이다. ‘문장당 주술관계 병렬관계 술어-보어 수’ 세 지표는 표점 부호에 의존하지 않는 술어 기반 교차 검증 지표이다. 맨 아래의 ‘문장당 문장부호관계 수’는 정의상 표점 부호에 의존하므로 교차 검증 지표가 아니며 참고로만 제시한다.

샤오와 다이(2014)에서는 영중 번역에서 번역 중국어의 평균 문절 길이가 원어인 중국어보다 길며, 그 원인을 명사가 관형 수식어를 취할 가능성이 높다는 데서 찾았다. 본 연구에서 관찰된 한중 번역의 양상도 이와 같은 방향이다. 중국어는 유수구(流水句)¹⁷⁾적 구성을 즐겨 쓰는 반면(呂叔湘, 1979), 한국어는 연결어미와 관형절을 이용해 하나의 문장 안에 정보를 응집시킨다. 한국어 원문의 응집적 구조가 중국어로 옮겨지는 과정에서 단위 분할이 충분히 이루어지지 않는 것은, 출발어의 구조가 목표어 텍스트에 남는 경우와 양립하는 분포이다(Teich, 2003). 다만 이 설명은 출발 텍스트와 번역문의 대응을 전제하므로 본 설계로는 판정할 수 없다.

한 가지 덧붙일 것은, 이 지표에서 일간지 번역문이 AI용 번역문보다 비번역 중국어에서 더 멀다는 점이다. 어휘 다양성에서 확인된 3단 위계가 여기서는 성립하지 않는다. 이는 표점 단위의 분할이 어휘 선택과 달리 생산 방식의 정교함에 비례하지 않는다는 것을 뜻한다. 곧 전문가 번역자라 하더라도 한국어 원문의 응집적 구조를 중국어에서 흔히 관찰되는 유수구적 구성으로 재분할하는 데에는 이르지 못하며, 오히려 원문의 문장 경계를 더 충실히 따르는 만큼 단위 수가 줄어들 수 있다. 이 관찰은 4.5에서 다시 다룬다.

4.3.2 구 의존관계 유형의 분포

구 단위에서는 의존 구문 분석 결과를 이용하여 14종 의존관계 유형의 분포 비율을 산출하였다(<표 7>, <그림 3>). 표본이 말뭉치당 24만 문장에 이르므로 아주 작은 차이도 통계적 검정에서는 구별되며, 3.3.2에서 밝힌 대로 관측치의 독립성 가정도 충족되지 않는다.

17) 유수구(流水句)는 접속사 없이 짧은 절을 반점으로 이어 붙여 확장하는 중국어 특유의 문장 구성 방식을 가리킨다.

18) <표 7>의 관계 유형은 '한국어 명칭(중국어 명칭)'으로 제시하였다. N-LTP가 부착하는 태그는 표의 순서대로 ATT, ADV, WP, VOB, RAD, SBV, COO, POB, HED, LAD, CMP, FOB, DBL, IOB이다.

19) 비율은 문장별로 산출한 값의 평균이며 각 말뭉치의 관측치는 240,000이다. d 는 3.3.2에서 밝힌 대로 차이의 방향과 상대적 크기를 읽는 데에만 사용한다.

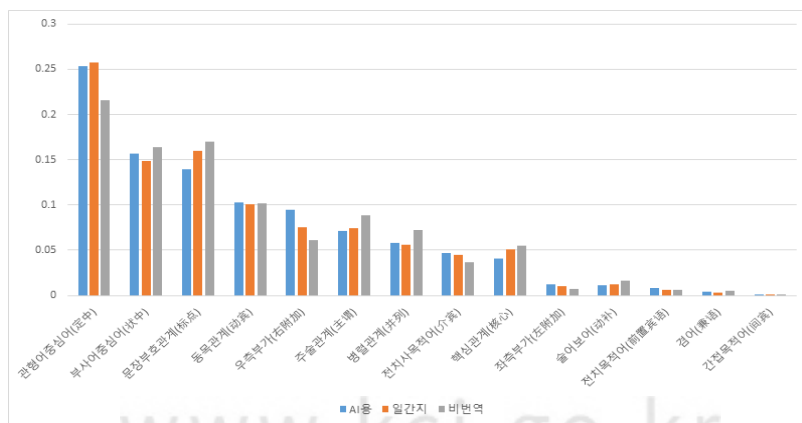
표 7

구 의존관계 유형별 분포 비율과 효과크기¹⁸⁾¹⁹⁾

구 의존관계 유형	AI-용	일간지	비번역	d(AI-비번역)	d(일간지-비번역)	d(AI-일간지)
관형어-중심어(定中)	.2538	.2576	.2155	+0.340	+0.357	-.035
부사어-중심어(状中)	.1563	.1485	.1642	-.097	-.184	+0.097
문장부호(标点)	.1393	.1594	.1702	-.406	-.132	-.324
동사-목적어(动宾)	.1030	.1009	.1015	+0.027	-.009	+0.038
우측 부가(右附加)	.0944	.0756	.0613	+0.644	+0.274	+0.359
주술(主谓)	.0715	.0740	.0883	-.355	-.289	-.058
병렬(并列)	.0581	.0555	.0719	-.276	-.317	+0.055
개사-목적어(介宾)	.0465	.0443	.0367	+0.257	+0.191	+0.057
핵심(核心)	.0412	.0512	.0550	-.335	-.082	-.367
좌측 부가(左附加)	.0118	.0102	.0069	+0.254	+0.171	+0.074
술어-보어(动补)	.0116	.0127	.0164	-.192	-.140	-.047
전치 목적어(前置宾语)	.0085	.0064	.0061	+0.120	+0.013	+0.107
겸어(兼语)	.0038	.0034	.0052	-.099	-.126	+0.032
간접 목적어(间宾)	.0002	.0002	.0008	-.129	-.128	+0.003

그림 3

구 의존관계 유형별 분포 비율



분포 비율은 문장별로 산출한 비율의 평균이므로 문장 길이의 분포에 민감하다. 문장이 짧을수록 문장당 한 번만 출현하는 관계가 전체에서 차지하는 비율은 커지고, 명사구를 확장하는 데 관여하는 관계의 비율은 작아진다. 그런데 4.1.2에서 보았듯 세 말뭉치의 문장 길이 분포는 데이터 구축 절차의 차이로 크게 다르며, 비번역 중국어에만 9자 이하 문장이 9,517개 존재한다. 이 절이 묻는 것은 번역 행위가 남긴 흔적이므로, 절차와 관련된 이 차이는 결과에서 분리되어야 한다.

d의 해석은 문장별 분포가 평균으로 요약되는 유형에 한정한다. 간접 목적어, 전치 목적어, 겹어, 술어-보어 구조, 좌측 부가 관계의 다섯 유형은 세 말뭉치와 세 분야 모두에서 중앙값이 0이어서 절반 이상의 문장에 나타나지 않는다. 이런 분포에서는 평균과 표준편차가 분포를 대표하지 못하므로 이하의 해석에서 제외한다.

비번역 중국어를 기준으로 할 때, AI 학습용 번역문에서 과다 사용된 의존관계 유형은 다음과 같다. 우측 부가 관계가 비번역 중국어 .0613에서 .0944로 가장 큰 차이를 보였고($d = +0.64$), 관형어-중심어 관계가 .2155에서 .2538로($d = +0.34$), 개사-목적어 관계가 .0367에서 .0465로($d = +0.26$) 뒤를 이었다. 반대로 과소 사용된 유형은 주술 관계($d = -0.36$)와 병렬 관계($d = -0.28$)이다.

이 관찰을 통해, 과다 사용된 유형은 모두 명사구의 확장에 관여하는 관계임을 알 수 있다. 관형어-중심어 관계는 수식 구조를, 우측 부가 관계는 ‘的’, ‘们’과 같은 후치 성분을, 개사-목적어 관계는 전치사구를 형성한다. 반면 과소 사용된 유형은 모두 절의 형성과 연쇄에 관여하는 관계이다. 주술 관계는 절의 골격을, 병렬관계는 단위와 단위의 연결을 반영한다. 요컨대 AI 학습용 번역문은 명사구를 상대적으로 길게 확장하고, 술어 중심의 절 연쇄는 상대적으로 적게 실현한다. 이는 4.3.1에서 확인한 표점 단위 결과와 정확히 맞물리며, 샤오와 다이(2014)가 영중 번역에서 보고한 관형 수식 구조의 과잉과도 같은 양상이다. 이 분포는 두 가지 설명과 모두 양립한다. 하나는 중국어 서면어 자체의 명사구 확장 경향이고, 다른 하나는 한국어의 관형절 중심 구조가 목표어에 남긴 흔적이다. 다만 본 연구의 설계는 출발 텍스트를 분석 대상으로 삼지 않았으므로 이 두 해석 가운데 어느 쪽이 타당

한지를 판정할 수 없다. 한편 과다 사용의 정도가 두 번역문에서 거의 같다는 사실($d = +0.34$ 대 $+0.36$)은 이 현상이 개별 번역 방식보다 언어쌍의 조건에 더 강하게 묶여 있을 가능성을 시사한다. 두 가설의 분리는 한국어 원문 자료를 함께 분석하는 후속 과제로 남긴다.

일간지 번역문의 관측 결과도 주목할 만하다. 관형어-중심어 관계에서는 일간지 번역문이 오히려 AI용 번역문보다 약간 높으며(.2576 대 .2538), 비번역 중국어와의 거리도 비슷하다($d = +0.36$). 즉 관형 구조의 상대적 고빈도는 본 자료의 두 번역문 집단에서 공통으로 관찰된 양상이다. 반면 우측 부가 관계에서는 AI용 번역문(.0944)이 일간지 번역문(.0756)보다 뚜렷이 높다($d = +0.36$). 요컨대 관형 구조의 과잉은 두 번역문이 거의 같은 크기로($d = +0.34$ 대 $+0.36$) 공유하는 반면, ‘的’류 부가 성분의 과다 사용은 AI 학습용 번역문에서 한층 심화된다. 앞의 양상은 언어쌍의 구조적 영향과, 뒤의 양상은 번역 생산 조건과 관련될 가능성을 각각 시사한다.

4.3.3 구 의존관계 유형별 어휘 다양성

4.3.2의 분포 비율이 어떤 구조를 얼마나 자주 쓰는지를 보여 준다면, 유형별 어휘 다양성은 같은 구조를 얼마나 다양한 어휘 조합으로 실현하는지를 보여 준다. 후런펀(胡撻奋, 2021)과 후 등(Hu et al., 2022)은 이 지표가 중국어 학습자의 숙달도를 예측한다는 점을 보인 바 있다. 본 연구에서는 의존관계 유형별로 서로 다른 어휘 쌍의 수를 해당 유형의 전체 출현 수로 나누어 산출하였다(<표 8>, <그림 4>).

- 20) ‘규모 편향’ 열은 토큰 수가 가장 많은 말뭉치가 어디인지에 따른 편향의 방향이다. ‘보수적’은 비번역 중국어의 값이 과소 추정되어 관찰된 격차가 실제보다 작게 나타남을, ‘반보수적’은 AI용 번역문의 값이 과소 추정되어 격차가 실제보다 크게 나타남을 뜻한다.

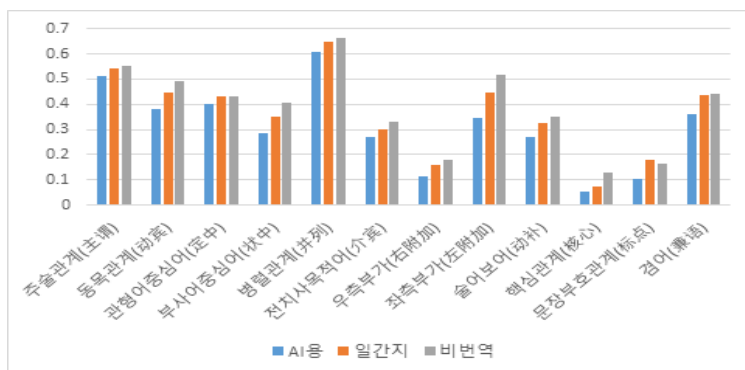
표 8

구 의존관계 유형별 어휘 다양성(토큰당 유형 비율)

구의존관계유형	AI용	일간지	비번역	AI용/비번역	일간지/비번역	규모편향
주술(主谓)	0.5118	0.5398	0.5519	0.93	0.98	보수적
동사-목적어(动宾)	0.3787	0.4447	0.4910	0.77	0.91	반보수적
관형어-중심어(定中)	0.3991	0.4311	0.4322	0.92	1.00	반보수적
부사어-중심어(状中)	0.2854	0.3500	0.4047	0.71	0.86	보수적
병렬(并列)	0.6084	0.6488	0.6647	0.92	0.98	보수적
개사-목적어(介宾)	0.2693	0.3025	0.3281	0.82	0.92	반보수적
우측 부가(右附加)	0.1133	0.1568	0.1784	0.63	0.88	반보수적
좌측 부가(左附加)	0.3478	0.4466	0.5160	0.67	0.87	반보수적
술어-보어(动补)	0.2693	0.3240	0.3512	0.77	0.92	보수적
핵심(核心)	0.0554	0.0747	0.1286	0.43	0.58	반보수적
문장부호(标点)	0.1019	0.1786	0.1631	0.62	1.10	보수적
겸어(兼语)	0.3618	0.4382	0.4409	0.82	0.99	보수적

그림 4

구 의존관계 유형별 어휘 다양성²¹⁾



21) 간접목적어와 전치 목적어는 출현 빈도가 낮아 어휘 쌍 유형 수의 추정이 불안정하므로 그림에서는 이 둘을 제외하고 12개 유형을 대상으로 한다.

12개 유형 전부에서 AI 학습용 번역문의 값이 비번역 중국어보다 낮으며, ‘AI용 번역문 < 일간지 번역문 ≤ 비번역 중국어’라는 세 말뭉치의 순서도 문장부호 관계 하나를 제외한 11개 유형에서 성립한다. 비번역 중국어 대비 AI용 번역문의 비는 핵심관계 0.43, 문장부호 관계 0.62, 우측 부가 관계 0.63, 좌측 부가 관계 0.67, 부사어-중심어 관계 0.71 순으로 낮다. 다만 앞서 밝힌 대로 핵심관계와 문장부호 관계는 문장 길이 분포에 의한 것이므로, 이 두 유형의 값은 해석에서 제외한다.

다만 이 지표에는 4.2에서 다룬 것과 동일한 규모 편향이 남아 있다. 편향의 방향은 유형마다 다르므로 <표 8>에 함께 표시하였다. 주술, 부사어-중심어 구조, 병렬, 술어-보어 구조, 문장부호, 겹어의 여섯 관계 유형은 비번역 중국어의 토큰 수가 가장 많아 비번역 중국어의 값이 과소 추정되며, 따라서 관찰된 격차는 보수적이다. 반면 동사-목적어, 관형어-중심어, 개사-목적어, 우측 부가, 좌측 부가, 핵심 등의 여섯 관계 유형은 AI용 번역문의 토큰 수가 가장 많아 AI용 번역문의 값이 과소 추정되므로, 이들 유형의 격차는 과장되었을 수 있다. 요컨대 결론을 뒷받침하는 증거는 전자의 여섯 유형 가운데 문장부호 관계를 뺀 다섯 유형, 곧 주술(0.93), 병렬(0.92), 겹어(0.82), 술어-보어(0.77), 부사어-중심어(0.71)에 국한된다. 이 다섯 유형에서 AI용 번역문의 비율은 일관되게 1을 밑돈다.

4.4 기능어 층위의 확인

양적 분석에서 드러난 경향이 개별 언어 형식으로 어떻게 실현되는지를 확인하기 위해 사회 분야 하위말뭉치의 조사(助詞) 사용 양상을 살펴보았다(22)(23). 조사는 텍스트의 주제에 좌우되지 않고, 문법 체계에 직접 관여하

- 22) 이 절의 결과는 사회 분야에 국한된 사례 검증이므로, 앞 절들의 결론을 뒷받침하는 독립적 증거가 아니라 그 결론이 개별 형식에서 어떻게 나타나는지를 예시하는 자료로 읽어야 한다.
- 23) 3.1.4의 해석 원칙에 따라 확인 대상을 기능어로 한정하였다. 실제로 사회 분야의 상위 2-그램 연어를 보면 이 위험이 확인된다. 비번역 중국어에서는 ‘工作人员(직원)’, ‘告诉记者(기자에게 말하다)’, ‘本报讯(본보 소식)’, ‘犯罪嫌疑人(피의자)’, ‘公安机关(공안 기관)’ 등 사건 사고 보도의 관습적 표현이 상위를 차지하는 반면, AI용 번역문에서는 ‘最低工资(최저 임금)’, ‘工作岗位(일자리)’, ‘雇佣劳动部

므로 생산 조건의 차이를 관찰하기에 적합하다. 결과는 <표 9>와 같다.

표 9

주요 조사(助詞)의 사용 빈도(사회 분야, 괄호 안은 만 단어당 빈도)

조사	기능	AI용 번역문	일간지 번역문	비번역 중국어
的	관형어표지	130,959 (657.4)	102,929 (563.0)	75,677 (447.3)
了	완료변화상	30,844 (154.8)	20,630 (112.8)	28,923 (170.9)
等	열거표지	16,640 (83.5)	9,239 (50.5)	2,881 (17.0)
着	지속상	1,262 (6.3)	1,647 (9.0)	5,522 (32.6)
地	부사어표지	2,353 (11.8)	1,551 (8.5)	2,433 (14.4)
过	경험상	686 (3.4)	1,101 (6.0)	2,189 (12.9)
得	보어표지	568 (2.9)	628 (3.4)	1,666 (9.8)

관형어 표지 ‘的’는 만 단어당 AI용 657.4회, 일간지 번역문 563.0회, 비번역 중국어 447.3회로, AI용이 비번역 중국어의 1.47배를 사용하였다.²⁴⁾ 4.3.2에서 확인한 우측 부가 관계의 과다 사용이 실제로는 ‘的’의 과다 사용으로 실현되고 있음을 알 수 있다. 세 말뭉치의 순서가 어휘 다양성에서와 동일하게 ‘AI용 번역문 > 일간지 번역문 > 비번역 중국어’로 나타난다는 점도 주목할 만하다. 이 결과는 샤오와 다이(2014)가 영중 번역에서 ‘的’를 번역 중국어의 최상위 키워드 가운데 하나로 보고하고 이를 관형 수식의 과잉과 문절 길이 증가로 연결한 것과 정확히 부합한다. 출발어가 다르고 생산 방식이 다른 자료에서 같은 현상이 관찰된다는 점에서, ‘的’의 과잉은 목표어가 중국어인 번역 텍스트의 두드러진 특징으로 보인다.

(고용노동부), ‘劳资双方(노사 양측)’ 등 노동 고용 관련 항목이 집중된다. 즉 같은 ‘사회’ 분류에 속하더라도 두 말뭉치가 다루는 하위 주제가 상당히 다르다. 이러한 조건에서 어휘 항목의 적절성을 개별적으로 판정하는 것은 신뢰하기 어렵다. 참고로 위에 든 AI용 번역문의 언어 항목들은 검토 결과 모두 한국 정부와 중국 매체가 실제로 사용하는 공식 내지 통용 형식이었으며, 축자역의 사례로 제시할 수 없는 것이었다.

- 24) N-LTP의 품사 체계는 구조조사 ‘的’와 어기조사 ‘的’를 구분하지 않으므로 <표 9>의 빈도에는 후자가 포함되어 있다. 다만 신문 보도문에서 어기조사 ‘的’의 비중은 낮으므로 세 말뭉치의 대조에 미치는 영향은 제한적이다.

열거 표지 ‘等(등)’의 경우 격차가 더 크다. AI용 번역문 83.5회, 일간지 번역문 50.5회, 비번역 중국어 17.0회로 AI용이 비번역 중국어의 4.9배에 이른다. 한국어의 열거 표현 ‘등’이 중국어 ‘等’으로 기계적으로 대응된 결과로 볼 수 있으며, 여기서도 세 말뭉치의 순서는 동일하다.

반대로 중국어 특유의 문법 범주를 표시하는 조사에서는 AI용의 사용이 현저히 적다. 지속상 표지 ‘着’는 AI용 6.3회로 비번역 중국어 32.6회의 19% 수준이고, 경험상 표지 ‘过’는 27%, 보어 표지 ‘得’는 29% 수준이다. 상(相) 표지와 보어 구문은 한국어에 대응 형식이 뚜렷하지 않은 중국어 고유의 문법 자원이다. 이들의 과소 사용은 4.3.2에서 확인한 술어-보어 구조의 과소 사용과 정확히 대응한다. 다이(2016)가 번역 중국어에서 경동사 구문과 ‘着’, ‘过’의 결합에 나타나는 문법 제약 위반을 출발어 투과의 사례로 보고한 것도 같은 맥락으로 읽을 수 있다.

요컨대 AI 학습용 번역문은 사회 분야의 조사 사용에서는, 한국어에 대응 형식이 있는 문법 요소가 상대적으로 고빈도로, 대응 형식이 뚜렷하지 않은 요소가 상대적으로 저빈도로 나타났다. 이 양상은 출발어의 형식 목록이 목표어 형식 선택에 영향을 미칠 가능성을 시사한다. 일간지 번역문도 같은 방향의 편향을 보이되 그 정도가 덜하다. 두 번역문 집단은 생산 목적 수용자 편집 단위 검수 방식이 동시에 다르므로, 이 격차가 어느 조건과 관련되는지는 본 설계로 특정할 수 없다.

경동사 ‘进行(진행하다)’의 경우는 해석에 신중을 요한다. 사회 분야에서 ‘进行’의 만 단어당 빈도는 AI용 51.4회, 일간지 번역문 35.3회, 비번역 중국어 13.4회로 AI용이 비번역 중국어의 3.8배, 일간지 번역문이 2.6배를 사용한다. 이를 한국어의 ‘~을 실시하다 진행하다’ 구성이 축자 대응된 결과로 보는 해석이 가능하나, 다이(2016)는 ‘进行+동사성 명사’ 구문이 번역 중국어 뿐 아니라 원어인 중국어에서도 통시적으로 확대되고 있음을 보고하였다. 즉 이 구문의 증가는 번역이 유발한 현상인 동시에 목표어 내부의 변화이기도 하다. 다만 본 연구의 세 말뭉치가 보이는 단조 증가 양상(비번역 중국어 < 일간지 번역문 < AI용 번역문)은 이 구문의 사용이 목표어 내적 변화만으로 설명되지 않으며 번역 방식에 따라 체계적으로 달라진다는 점을 시사한다.

4.5 소결

이상의 분석을 층위별로 정리하면 다음과 같다.

첫째, AI 학습용 번역문에서는 반각 문장부호의 일관된 사용과 짧은 문장 유형의 현저한 결여가 확인되었다. 이는 번역 선택보다 원자료의 정제 저장과 데이터 구축 절차에서 비롯된 특성으로 보인다.

둘째, 어휘 층위에서 AI용 번역문은 비번역 중국어보다 전반적으로 낮은 품사별 어휘 다양성을 보였으며, 일간지 중국어판 번역문과 비교해도 대체로 낮은 값을 나타냈다.

셋째, 통사 층위에서는 두 번역문 집단이 비번역 중국어보다 관형어-중심어 및 우측 부가 관계를 상대적으로 많이, 주술 및 병렬 관계를 상대적으로 적게 사용하는 경향을 보였다. 이 가운데 우측 부가 관계와 ‘的’의 사용에서는 AI용 번역문의 차이가 더 두드러졌다.

따라서 본 자료에서는 어휘 다양성의 차이가 두 번역문 집단 사이에서 비교적 선명한 반면, 일부 구조적 통사 지표의 차이는 두 집단에 공통적으로 나타났다.

5. 결론

본 연구는 AI 학습용 중국어 번역문과 일간지 번역문, 두 번역문 집단에 공통으로 나타나는 분포적 경향과, AI 학습용 번역문에서 상대적으로 두드러지는 특성을 분석하고자 하였다.

세 연구 문제에 대한 답은 다음과 같다.

연구 문제 1에 대하여, 두 번역문 집단은 비번역 기사와 공통으로 구별되었다. 어휘 영역에서는 품사별 어휘 다양성이 낮고, 통사 영역에서는 명사구를 확장하고 절의 연쇄를 줄이는 경향을 공유한다.

연구 문제 2에 대하여, 두 번역문 집단 사이에는 체계적 차이가 있었다. AI 학습용 번역문은 어휘 다양성에서 일간지 번역문보다 한 단계 더 낮은 위치에 있으며, 그 격차는 주제에 좌우되지 않는 문법적 관습적 자원에서 가장 크다. 통사 영역의 차이는 우측 부가 관계와 ‘的’의 사용에 국한된다.

표기 영역의 반각 문장부호는 번역 행위가 아니라 데이터 구축 절차에서 비롯된 특징이다.

연구 문제 3에 대하여, 공통 경향은 단순화와 정규화 가설과 정합적이며, 두 집단 사이의 격차는 수용자 편집 단위 검수 기준 번역 생산 방식이라는 네 요인과 양립하나 본 설계로는 그 몫을 분리할 수 없다.

본 연구의 의의는 다음의 세 가지로 정리할 수 있다.

첫째, 번역 중국어의 계량적 프로파일 연구를 한중 언어쌍으로 확장하였다. 이 분야의 연구는 ZCTC 계열(Xiao, 2010; Xiao & Dai, 2014; Xiao et al., 2010; Xiao & Hu, 2015)을 중심으로 영중 언어쌍에 집중되어 왔으며, 한국어를 출발어로 하는 번역 중국어에 대해서는 어휘 다양성 기능어 분포 의존관계 구조를 아우르는 대응말뭉치 연구가 확인되지 않았다. 본 연구는 ‘의’의 고빈도 사용과 관형 수식 구조의 확대라는 기존 발견이 한중 언어쌍에서도 재현됨을 보이는 한편, 어휘 다양성에 관해서는 샤오(2010)의 결과와 어긋나는 관찰을 제시하여 논쟁의 재료를 더하였다.

둘째, 비번역 중국어 참조물과 두 유형의 번역문을 함께 둔 3원 비교설계는 번역 텍스트와 비번역 텍스트의 차이를 생산 배경이 다른 번역문 사이의 차이와 함께 관찰하게 한다. 또한 어휘 다양성의 규모 보정과 의존관계 분포의 변이 고려는 관찰되는 차이의 크기와 해석에 영향을 주었다. 이는 번역 중국어의 계량적 연구에서 비교말뭉치 설계와 측정 방식의 명시가 중요함을 보여 준다.

셋째, 본 연구는 AI 학습용 번역 데이터의 품질 관리가 한국어 원문-번역문 쌍의 대응 정확성만으로 충분하지 않음을 시사한다. 반각 문장부호와 문장 길이 분포의 편중은 구축 절차에서 점검 완화할 수 있는 특성인 반면, 낮은 어휘 다양성과 일부 의존관계 분포의 편향은 번역문 전체의 목표어 적합성을 기준으로 별도로 관찰할 필요가 있다. 따라서 정서법 준수율, 품사별 표준화 TTR, 주요 의존관계 분포와 같은 말뭉치 수준의 지표는 AI 학습용 병렬말뭉치의 언어적 활용 조건을 점검하는 보조 기준이 될 수 있다. 다만 본 자료를 중국어의 실제 용법을 기술하는 참조 말뭉치로 사용하는 데에는 신중할 필요가 있으며, 번역투 기계번역 산출물 한중 대조언어학 연구를 위한 자료로 활용하는 편이 더 적절하다.

본 연구에는 몇 가지 한계가 있다. 세 하위말뭉치는 생산 시기, 매체 및 하위 주제가 완전히 일치하지 않으므로, 관찰된 차이를 번역 생산 조건의 순수한 효과로 해석할 수 없다. 또한 분석은 신문 기사와 표기 어휘 통사 층위에 한정되었고, N-LTP의 품사 및 의존구문 분석 결과에 의존하였다. 특히 분석기가 비번역 중국어 텍스트와 번역문에서 서로 다른 오류 양상을 보일 가능성을 표본 평가로 검증하지 못했다. 나아가 병렬말뭉치임에도 중국어 측만 분석하였으므로, 한국어의 관형절 구조가 중국어 번역문에 전이되었다는 설명은 가설적 수준에 머문다. 아울러 문장 단위로 산출한 지표의 신뢰 구간은 문장을 독립 관측으로 가정한 값이다. 세 하위말뭉치는 모두 문장이 기사 안에 군집된 위계적 자료이므로, 이 구간은 실제보다 좁게 추정되었을 수 있다. 기사를 재표집 단위로 한 재산출은 후속 과제로 남긴다. 후속 연구에서는 시기 매체 하위 주제를 더 엄격히 통제하고, 한중 대응쌍의 질적 분석과 분석기 성능의 표본 검증을 병행할 필요가 있다.

참고문헌

<1차 자료>

- 한국지능정보사회진흥원. (2020). 한국어-중국어 번역 말뭉치 2(사회과학) [데이터셋]. AI 허브.
- 황은하, 비립. (2023). 일간지 중국어 번역문 말뭉치 [미간행 원자료]. 중앙일보, 동아일보, 조선일보, 아주일보 중국어판 기사에서 수집 구축.
- 清华大学自然语言处理与社会人文计算实验室. (2016). THUCNews: 中文文本分类数据集 [THUCNews: A Chinese text classification dataset] [데이터셋]. <http://thuctc.thunlp.org>
- 孙茂松, 李景阳, 郭志芑, 赵宇, 郑亚斌, 司宪策, 刘知远. (2016). THUCTC: 一个高效的中文文本分类工具包 [THUCTC: An efficient Chinese text classifier]. 清华大学自然语言处理与社会人文计算实验室. <http://thuctc.thunlp.org>

<2차 자료>

- 김보영, 김연주, 서승희, 송신애, 이진현, 전경아, 최지수, 홍승빈, 정혜연, 허탁성. (2020). 번역자동평가에서 풀리지 않은 과제. 번역학연구, 21(1), 9-29.
- 신지연. (2017). 한중 번역에 나타나는 명시화 현상 고찰: 은희경의 『소년을 위로해줘』를 중심으로. 외국학연구, 41, 65-87.
- 웨이잉. (2019). 한중 번역 소설의 연결 표지 명시화 연구 [석사학위논문, 이화여자대학교]. RISS.
- 장현주. (2006). 중한 번역에서 번역투(translationese)와 정보 배열. 중국학연구, 37, 345-364.
- 정혜연. (2018). 번역의 자동평가: 기계번역 평가를 인간번역 평가에 적용해 보기. 통번역학연구, 22(4), 265-287.
- 정혜연, 박현일, 우경조, 서수영. (2021). 임베딩을 활용한 인간번역의 자동평가: 기계가 의미를 평가할 수 있을까. 통번역학연구, 25(3), 141-162.
- 정혜연, 최지수, 허탁성, 서수영. (2022). 인간 번역평가에서 재현도(recall)의 중요성. 번역학연구, 23(1), 81-100.
- 최지수. (2023). 인간번역의 자동평가에서 번역 텍스트 길이가 평가 타당성에 미치는 영향. 통번역학연구, 27(2), 145-164.
- 최효은, 이준호, 이청호. (2023). 자동화된 기계학습(AutoML)을 활용한 특허 특화 번역엔진의 영한번역 성능 평가. 번역학연구, 24(2), 101-130.
- 한국지능정보사회진흥원. (2021a). 인공지능 데이터 구축 활용 가이드라인: 전문분야 학습용 한-중/일 번역 말뭉치. AI 허브. <https://aihub.or.kr>
- 한국지능정보사회진흥원. (2021b). 한국어-중국어 번역 말뭉치 2(사회과학) 데이터 설명서. AI 허브. <https://aihub.or.kr>
- 황은하. (2016). 말뭉치 기반 한외(韓外) 대조언어학 연구에 대한 일고찰. 어문론총, 69, 39-62.
- 황은하. (2021). 대조분석을 위한 말뭉치의 타당성 연구: 한중 대조분석을 중심으로. 이중언어학, 82, 259-286.
- 国家质量监督检验检疫总局, 国家标准化管理委员会. (2011). 标点符号用法 (GB/T 15834-2011) [Rules for punctuation usage]. 中国标准出版社.

- 胡韧奋. (2021). 基于搭配的句法复杂度指标及其与汉语二语写作质量关系研究 [A study on collocation-based syntactic complexity indices and their relationship with L2 Chinese writing quality]. 语言文字应用, 2021(1), 132-144.
- 胡韧奋, 肖航. (2019). 面向二语教学的汉语搭配知识库构建及其应用研究 [Construction and application of a Chinese collocation knowledge base for L2 teaching]. 语言文字应用, 2019(1), 135-144.
- 黄立波. (2008). 英汉翻译中人称代词主语的显化: 基于语料库的考察 [Explicitation of personal pronoun subjects in English-Chinese translation: A corpus-based study]. 外语教学与研究, 2008(6), 454-459.
- 吕叔湘. (1979). 汉语语法分析问题 [Problems in the analysis of Chinese grammar]. 商务印书馆.
- Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233-250). John Benjamins.
- Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP and translation* (pp. 175-186). John Benjamins.
- Bielsa, E., & Bassnett, S. (2009). *Translation in global news*. Routledge.
- Che, W., Feng, Y., Qin, L., & Liu, T. (2021). N-LTP: An open-source neural language technology platform for Chinese. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 42-49).
- Chesterman, A. (2004). Beyond the particular. In A. Mauranen & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (pp. 33-49). John Benjamins.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.
- Daems, J., De Clercq, O., & Macken, L. (2017). Translationese and post-editese:

- How comparable is comparable quality? *Linguistica Antverpiensia, New Series — Themes in Translation Studies*, 16, 89-103.
- Dai, G. (2016). *Hybridity in translated Chinese: A corpus analytical framework*. Springer.
- Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In L. Wollin & H. Lindquist (Eds.), *Translation studies in Scandinavia* (pp. 88-95). CWK Gleerup.
- Heck, K. L., Jr., van Belle, G., & Simberloff, D. (1975). Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology*, 56(6), 1459-1461.
- Hermans, T. (1999). *Translation in systems: Descriptive and system-oriented approaches explained*. St. Jerome Publishing.
- House, J. (2008). Beyond intervention: Universals in translation? *trans-kom*, 1(1), 6-19.
- Hu, R., Wu, J., & Lu, X. (2022). Word-combination-based measures of phraseological diversity, sophistication and complexity and their relationship to second language Chinese proficiency and writing quality. *Language Learning*, 72(4), 1128-1169.
- Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (NCTE Research Report No. 3). National Council of Teachers of English. <https://eric.ed.gov/?id=ED113735>
- Hurlbert, S. H. (1971). The nonconcept of species diversity: A critique and alternative parameters. *Ecology*, 52(4), 577-586.
- Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, 43(4), 557-570.
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474-496.
- Teich, E. (2003). *Cross-linguistic variation in system and text*. Mouton de Gruyter.
- Toral, A. (2019). Post-editeese: An exacerbated translationese. *Proceedings of*

Machine Translation Summit XVII: Research Track (pp. 273-281).

- Toury, G. (1995). *Descriptive translation studies and beyond*. John Benjamins.
- Vanmassenhove, E., Shterionov, D., & Gwilliam, M. (2021). Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2203-2213).
- Xiao, R. (2010). How different is translated Chinese from native Chinese? A corpus-based study of translation universals. *International Journal of Corpus Linguistics*, 15(1), 5-35.
- Xiao, R., & Dai, G. (2014). Lexical and grammatical properties of translational Chinese: Translation universal hypotheses reevaluated from the Chinese perspective. *Corpus Linguistics and Linguistic Theory*, 10(1), 11-55.
- Xiao, R., He, L., & Yue, M. (2010). In pursuit of the “third code”: Using the ZJU Corpus of Translational Chinese in translation studies. In R. Xiao (Ed.), *Using corpora in contrastive and translation studies* (pp. 182-214). Cambridge Scholars Publishing.
- Xiao, R., & Hu, X. (2015). *Corpus-based studies of translational Chinese in English-Chinese translation*. Springer.

A comparable-corpus study of linguistic features in Korean-to-Chinese translations produced for AI training and published in Korean newspapers

Eunha Hwang

Department of Korean Language, Literature and Korean Language Education, Pai Chai University

Li Fei (corresponding author, feili0820@sookmyung.ac.kr)

Department of Business Administration, Sookmyung Women's University

Abstract

This study compares Korean-to-Chinese translations compiled for AI training and published in Korean newspaper Chinese editions with non-translated Chinese news. A three-way comparable corpus of 720,000 sentences, comprising 240,000 sentences per subcorpus, was constructed from the AI Hub Korean-Chinese parallel corpus, four Korean newspaper Chinese editions, and THUCNews. The subcorpora were balanced for news genre and the social, economic, and cultural domains; the AI Hub subset was restricted to files sourced exclusively from Korean news outlets where source information was available. Both translation sets showed lower size-adjusted lexical diversity than the reference corpus and a shared preference for expanded noun phrases over clause chaining. AI-training translations showed still lower lexical diversity, particularly in particles, idioms, adverbs, and conjunctions, as well as more frequent right-adjunct relations and use of the attributive marker *de* (的). Their consistent use of half-width punctuation resulted from the corpus-construction pipeline rather than from translation. The shared tendencies are consistent with simplification and normalization. Differences between the translation sets may be associated with intended readership, editorial unit, quality-control procedures, and translation production method, although their individual effects cannot be distinguished in the present design. Quality control for AI training corpora should therefore extend beyond source-target correspondence to target-language linguistic profiles and corpus-construction procedures.

Keywords: Korean-Chinese translation; translated Chinese; translationese; comparable corpus; AI-training parallel corpus

키워드: 한중 번역, 번역 중국어, 번역투, 비교말뭉치, AI 학습용 병렬말뭉치

황은하(제1저자, <https://orcid.org/0009-0000-5967-0808>)

배재대학교 국어국문한국어교육학과 조교수

yxhuang@pcu.ac.kr

비립(교신저자, <https://orcid.org/0000-0002-7366-1817>)

숙명여자대학교 경영학과 조교수

feili0820@sookmyung.ac.kr

논문 투고일: 2026년 8월 15일

1차 심사 완료일: 2026년 9월 1일

2차 심사 완료일: 2026년 9월 8일

게재 확정일: 2026년 9월 15일