http://dx.doi.org/10.15749/jts.2020.21.2.002

번역학연구
2020 여름호 제21권2호

영어 어휘 다양성과 어휘 정교성이 한영번역에 미치는 영향

김 훈 밀
(국제영어대학원대학교)

1. 서론

전 세계에 걸쳐 글로벌화가 가속화되고 있는 가운데 번역 수요는 꾸준히 증가하고 있으며(IbisWorld 보고서) 이에 따라 번역 품질 평가에 대한 관심도 크게 늘고 있다. 이는 학술지에 게재된 번역 관련 논문의 주제를 통해서도 알 수 있는데 2000년-2017년 사이 동료평가를 거친 학술지에 게재된 265편의 번역 관련 논문 중 번역품질평가를 주제로 한 것이 160편으로 전체 논문의 61%를 차지하는 것으로 나타났다(Akbari, 2018). 이는 제대로 된 번역, 우수한 번역사를 평가할 수 있는 객관적이고 실효성 높은 평가체제에 대한 관심과 필요성이 증대되고 있음을 보여준다.

번역사 인증을 위한 캐나다의 CTTIC와 호주의 NAATI, 번역물 평가를 위한 프랑스의 SEPT, 미국의 GTS, 일본의 히타치를 비롯하여 우수번역 지원 선정에 사용되는 국내 한국문학번역원의 평가 모델에 이르기까지 이미 구축되어 실제 사용되고 있는 평가 모델도 다양하다(유정화, 2016). 각각의 모델마다 사

용 목적과 용도에 따라 평가 기준과 방식에는 차이가 있으나 우수한 평가 모델의 요건으로 공통적으로 타당성(validity)과 신뢰성(reliability)을 꼽는다. 이와 더불어 언어 평가 모델을 위해 요구되는 또 다른 요건으로 자주 언급되는 것이 실효성(practicality)이다(Bachman and Palmer, 1996: 35).

타당성과 신뢰성, 실효성 세 가지 요건을 모두 만족시키는 평가 모델은 어떤 모델일까? 인간 평가자의 경우 타당성은 높은 것으로 평가되나 신뢰성과 실효성은 낮은 것이 문제인데 자동 평가 시스템은 높은 신뢰성과 실효성을 자랑한다. 과거에는 자동 채점 프로그램의 타당성이 낮아 상대적으로 낮은 신뢰성과 실효성에도 인간 평가를 선호하였다. 그러나 최근 들어 전산 언어학의 발달로 다양한 자동평가시스템이 개발되어 인간 평가에 대한 대안으로 사용되고 있으며 코메트릭스(Coh-Metrix) 등 일부 자동평가시스템은 타당성, 신뢰성, 실효성을 모두 갖춘 평가 방법으로 보고되고 있다(Graesser, McNamara, Louwerse and Cai, 2004; Graesser, McNamara & Kulikowich, 2011). 번역 평가와 유사점이 많은 영작문 평가 분야에서는 이미 TOEFL iBT나 IELTS와 같은 대표적인 영어인증시험 기관들이 자동 평가 프로그램을 이용해 영작문을 평가하고 있다.

번역 분야에서도 기계번역물의 평가를 위해 개발되기는 하였으나 BLEU(Bilingual Evalution Understudy)와 Meteor와 같은 자동 평가 도구가 점차 활발히 사용되고 있다. 아직까지는 인간 평가자에 비해 타당성이 낮은 것으로 평가되고 있으나, 평가 기준을 정교화함으로써 타당성 개선 여지가 있는 것으로 보고되었다(정혜연, 2018). 비단 BLEU나 Meteor와 같은 특정 자동평가 프로그램의 타당성 개선을 위해서가 아니더라도, 보다 높은 타당성을 이끌어낼 수 있는 번역품질평가 모델이 마련된다면 이는 우수한 자동 평가 프로그램 개발에 기여할 뿐 아니라 인간 평가에도 적용하여 인간평가의 신뢰성을 높이는 데도 기여할 수 있을 것이다.

본 연구에서는 BLEU와 Meteor와 같은 자동 번역 평가 프로그램에서 사용하는 평가 항목을 살펴보고 이를 인간의 평가 항목과 비교하여 인간 평가 항목에 비해 자동 번역 평가 프로그램에 덜 반영되어 있는 평가 항목을 파악하고자 한다. 부족한 항목으로 파악된 항목에 대해서는 해당 항목이 번역 평가에 통계적으로 유의미한 영향을 미치는지를 통계기법을 활용하여 검증하고, 이를 정량적으로 측정할 수 있는 지표를 제안하고자 한다. 번역 평가에 통계적으로 유의

미한 영향을 미치나 자동번역평가 모델에서 누락되어 있는 정량적으로 측정 가능한 항목은 자동번역평가 프로그램에 손쉽게 추가할 수 있어 자동번역평가 프로그램의 타당성 향상에 기여할 수 있다. 뿐만 아니라 인간 평가를 보완하는 데에도 활용될 수 있으며 이를 통해 인간평가의 신뢰성과 실효성을 증대할 수 있다. 본 연구 결과는 번역평가 모델의 타당성, 신뢰성과 실효성을 높일 수 있는 구체적이며 실천 가능한 방안을 제시함으로써 앞으로 더욱 요구될 객관적 번역평가 모델 구축에 활용될 수 있다. 또한 교육, 자격 인증 등 다양한 번역 평가 현장에서 해당 항목을 측정하여 평가에 활용함으로 번역평가품질을 개선에도 기여할 수 있다. 본 연구를 통해 조사하고자 하는 연구질문은 아래와 같다.

연구질문 1: 인간 평가자의 번역평가기준과 비교할 때 자동번역평가 프로그램에 누락된 항목은 무엇이며 이를 측정할 수 있는 계량화된 지표는 무엇인가?

연구질문 2: 연구질문 1에서 도출된 계량화된 지표는 번역 점수를 결정하는데 통계적으로 유의미한 영향을 미치는가?

연구질문 3: 통계적으로 유의미한 영향을 미친다면 그 영향은 어느 정도인가(번역 점수의 몇 %가 해당 지표의 영향으로 설명되는가)?

2. 선행연구 분석

2.1 인간 평가자 Vs.자동번역평가 프로그램의 번역 평가 항목

번역 평가는 크게 정량적 평가와 정성적 평가로 구분된다. 정량적 방식은 '의미', '문법', '표현' 등의 세부 평가 항목을 정의한 후 각각에 대해 점수를 부여하여 총점을 구하는 방식이며, 정성적 방식은 수치화하기 어려운 특성에 대해 그 내용을 기술함으로써 품질 척도를 나타내는 평가 방식이다. 정성적 평가에서도 평과 결과를 수치화하여 제시하지는 않더라도 평가의 대상이 되는 항목을 '독창성', '완성도', '가독성' 등과 같이 정의하여 사용한다. 따라서 정량적 평가와 정성적 평가 모두 일정한 평가 항목을 기준으로 삼아 평가가 이루어짐을 알 수 있다. 현재 다양한 번역평가 모델에서 사용하고 있는 평가 항목은 의

미, 표현, 문법, 누락, 첨가, 맞춤법, 구두법, 논리성, 가독성 등 다양하나 여러 평가항목과 직관에 의한 번역평가 점수와의 상관관계를 분석한 한국외대 번역평가인증팀(2016)의 연구를 토대로 정혜연(2018)은 인간 평가자들이 주로 보는 평가항목은 '의미와 표현'이라고 요약하였다.

BLEU와 Meteor와 같은 자동평가 프로그램에서 사용되는 평가 항목은 무엇일까? 간단히 말해 자동평가 툴은 모범번역답안과 평가대상 번역물 간의 일치도를 비교하여 일치도가 높을수록 높은 점수를 부여한다. BLUE의 경우 1그램에서 4그램까지의 n그램 일치도를 비교하는 것인데 여기서 말하는 1그램이란 단어 하나를, 2,3,4그램은 연속된 단어 2개, 3개, 4개를 한 단위로 보고 모범답안과 평가 대상간의 일치도를 비교하는 것이다. 1그램 일치도는 개별 단어의 일치도를 비교하는 것으로 이는 의미에 맞는 단어가 사용되었는지 (예: famous: 유명한)를 평가한다. 반면, 2그램-4그램 일치도는 구와 절 단위로 어순을 비교해 번역물의 통사구조와 논리를 평가한다. Meteor의 경우에도 모범번역답안과 평가대상 번역물 간의 일치도를 토대로 일치도가 높을수록 높은 점수를 부과한다는 기본 평가 방식은 BLEU와 유사하다. 다만 Meteor의 경우 단어의 일치도를 단어가 완벽히 일치하는 경우, 어근이 일치하는 경우(stem 일치도), 유의어를 사용한 경우 3가지로 보다 폭넓게 인정하며 어순에 대해서는 어순이 일치하지 않을수록 페널티를 매겨 점수를 산정한다. 각 BLEU와 Meteor의 평가 방식 대한 보다 상세한 설명은 파피네니 외(Papineni et al., 2002)와 라비 외(Lavie et al., 2009)에 나와있다.

위의 평가 방식을 고려할 때 자동평가 프로그램의 평가항목은 표현보다는 의미에 편중되어 있다고 볼 수 있다. 모범답안과 얼마나 동일한 어휘(어근, 유의어 포함)을 사용하는지는 의미를 평가하는 것이며, 어순을 비교하는 것도 의미 전달이 효과적으로 이루어질 통사구조를 가졌는지를 보는 의미 중심의 평가로 볼 수 있다. 다양한 어휘를 다채롭게 사용하는지, 선택한 어휘가 난이도가 높은 어휘인지 여부는 평가에 반영되지 않아 번역물의 표현력 차이가 정교하게 평가된다고 보기는 어렵다. 예를 들어 '유명한'에 대한 번역은 유의어인 'famous', 'popular', 'renowned', 'prominent'에 대해 모두 동일한 점수를 부여한다. 또한 원천 텍스트에 '유명한'이란 단어가 10회 나올 때 이를 10번 모두 famous로 번역한 것이나 famous, renowned, prominent 등의 다양한 어휘를 고

루 사용해 번역한 것이 같은 점수로 평가받고 있다.

번역문의 완성도를 결정짓는 요소 중에는 번역물이 얼마나 목표어 문화권의 기준에서 완결된 텍스트의 조건을 충족하는지, 목표어로 작성된 유사한 유형의 텍스트와 동일한 수준의 텍스트 품질을 가지는지가 포함된다. 따라서 한영번역의 경우 평가 기준에 영작문 평가 기준을 토대로 일정 수준 이상의 텍스트 특성과 품질을 확보하느냐가 포함되어야 한다. 한 텍스트 내에서 동일어 반복에 대한 거부감이 상대적으로 적은 한국어와는 달리 영어 텍스트는 동일어 반복 사용을 꺼린다. TOEFL, IELT를 비롯한 대부분의 영작문 평가 기준에는 어휘의 다양성과 어휘의 정교성이 필수 평가 항목으로 포함되어있다. 따라서 어휘 다양성과 어휘 정교성은 한영번역물의 품질에 영향을 주는 주요 요소로 판단되나 현 자동번역평가 프로그램에서는 어휘 다양성과 어휘 정교성에 대한 평가가 이루어지지 않고 있다.

요약하자면 인간평가와 자동평가의 평가 항목을 비교해볼 때 자동평가에서는 인간의 주요 평가항목인 의미와 표현 중 의미에 치중된 평가가 이루어지고 있으며, 표현에 대한 평가는 크게 반영되지 않고 있는 것으로 보인다. 이에 현 자동번역평가 프로그램에서 보완이 필요한 부분은 번역물의 표현력을 보다 정교히 평가할 수 있는 항목을 추가하는 것이며, 표현력에 대한 지표로는 영작문 평가 항목으로 많이 사용되며 정량적으로 측정할 수 있는 어휘 다양성과 어휘 정교성을 고려해볼 수 있다.

2.2 영작문 평가에서 표현력 측정에 사용되는 지표

영작문의 품질에 영향을 미치는 요인을 연구한 많은 논문에서 표현력을 나타내는 지표로 어휘 풍부성(lexical richness)이 사용되며, 어휘 풍부성과 작문 점수 간에는 높은 양의 상관관계가 보고되었다(Crossley, Salsbury, McNamara & Jarvis, 2010; Ferris, 1994; Kyle & Crossley, 2015; Laufer & Nation, 1995). 리드(Read, 2000)는 어휘 풍부성을 다시 어휘 다양성(lexical diversity), 어휘 밀도(lexical density), 어휘 정교성(lexical sophistication)과 에러율(proportion of error)로 구분하였는데 그 중 텍스트 분석과 관련하여 가장 자주 사용되는 지표는 어휘 다양성과 어휘 정교성이다(Crossley, Salsbury and McNamara, 2012;

Lee, 2003). 본 연구도 분석 대상을 글로 이루어진 텍스트인 번역물로 한정하고 있어 어휘 풍부성을 나타내는 지표로 어휘 다양성과 어휘 정교성을 살펴보고자 한다.

2.1.1 어휘 다양성(lexical diversity)

어휘 다양성은 영어 라이팅과 스피킹 평가의 주요 평가 항목으로 광범위하게 사용되고 있다. 대표적인 영어공인시험인 IELTS와 TOEFL iBT뿐 아니라 자동영작문 평가 프로그램에서도 가장 중요한 평가 항목 중 하나로 채택되어 활용되고 있다(Chodorow & Burstein, 2004). 폭넓게 사용되는 만큼 개념에 대한 정의도 다소 다양하나 본 연구에서는 듀란 외(Duran et al., 2004)의 “글 또는 말의 담화물을 통해 드러나는 어휘의 범위(range of the vocabulary)”라는 정의를 기본 개념으로 사용하고자한다.

어휘 다양성을 측정하는 데 광범위하게 사용되는 지수는 타입-토큰 비율(Type-Token Ratio, TTR)이다. TTR은 반복 등장을 포함하여 텍스트에 등장하는 모든 어휘 수인 토큰 대비 한번씩 만 등장하는 어휘 수인 타입의 비율을 나타낸다. 즉 전체 텍스트에 사용된 어휘 중 반복되지 않고 사용된 어휘의 비율을 나타낸다. TTR관 관련하여 지적되는 문제점은 텍스트의 길이에 따라 TTR이 크게 영향을 받는다는 점이다. 즉, 텍스트가 길어지면 전체 어휘수인 토큰수는 계속 증가하지만 타입수는 어느 지점 이후부터는 크게 늘지 않아 텍스트가 길어질수록 TTR 비율이 낮아지는 경향을 보인다. 최근에는 이러한 단점을 해결하기 위해 수정된 TTR 산출 공식을 사용되기도 하나 본 연구에서는 분석 대상이 모두 유사한 길이의 번역 텍스트이므로 TTR의 텍스트 길이에 대한 민감도 영향을 받지 않아 기존의 TTR 산출 방식을 이용하였다.

2.1.2 어휘 정교성(lexical sophistication)

어휘 정교성이란 “텍스트 내에 사용된 어휘 중 상대적으로 수준 높은 고급 어휘의 비율”을 의미한다(Read, 2000). 여기서 상대적으로 수준 높은 고급 어휘라 함은 사용 빈도가 낮은 어휘를 지칭한다(Laufer & Nation, 1995). 영어 능력과 어휘력 간의 상관관계를 연구한 많은 연구에서 어휘 빈도를 어휘 수준을 나타내는 지표로 사용하였으며 영어 능력이 우수할수록 빈도수가 낮은 어휘를 사

용하는 경향이 보고되었다.

영어 어휘의 사용빈도는 네이션과 베글러(Nation & Beglar, 2007)의 연구에 기초한다. 네이션과 베글러는 코퍼스 분석을 통해 영어 어휘군을 사용 빈도에 따라 가장 기초 수준에서 가장 높은 난이도까지 1,000개씩 분류하였는데 코퍼스 상 사용 빈도가 가장 높은 기초 어휘군 1,000개를 1K, 그 다음으로 많이 사용되는 어휘군 1,000개를 2K로 구분하는 방식으로 영어 어휘를 1K부터 10K까지 분류하였다. 네이션의 K 단위 분류와 더불어 많이 사용되는 어휘 빈도수 분류로는 콕스헤드(Coxhead, 2000)의 아카데믹 워드 리스트(Academic Word List, AWL)가 있다. AWL에는 학술 텍스트에 자주 사용되는 570개의 어휘군이 포함되며 네이션의 고빈도 어휘 1K와 2K에 포함된 어휘는 포함되지 않는다. AWL은 낮은 빈도의 고급 어휘를 대표하는 목록으로 자주 사용되며 1K, 2K 어휘군과 대조하여 특정 텍스트 내에 사용된 고급 어휘 비율을 나타내는데 자주 이용된다.

2.3 번역 품질과 어휘력

언어학과 영어교육 분야에서 어휘와 관련된 다양한 개념화와 연구가 활발한 것과는 달리 국내외 통번역분야 모두 어휘에 대한 연구가 드물다. 아래에선 국내 통번역 학술지에 비교적 최근에 발표된 어휘 관련 연구들을 간략히 살펴봄으로써 지금까지 보고된 통번역과 어휘와의 관계를 정리해보고자 한다.

김재희(2008)는 잡지 코리아나의 아랍어 번역을 분석하여 문화 관련 어휘의 번역 전략을 연구하였다. 연구 결과 문화어 번역에는 주로 이국화 전략과 최소 수정 원칙이 적용되며 번역사는 원천어의 풍부성이 보존되는 방향으로 번역을 한다는 점을 밝혔다.

김훈밀(2014)은 번역능력에 영향을 미치는 하위능력을 조사하며 출발어 어휘지식이 출발어 독해능력과 함께 학생들의 번역 점수에 영향을 미치는 주요 항목인 것을 보고하였다. 또한 보다 최근 연구에선 학생들의 B언어 어휘력이 AB순차통역 점수차이의 79%, BA순차통역 점수차이의 43%를 차지함을 보이며 B언어 어휘력이 학생들의 순차통역 수행력 차이의 큰 부분을 설명함을 보고하였다(김훈밀, 2019).

정혜연(2016)은 통역사와 비통역사 간의 어휘 능력을 비교한 연구에서 유의

어 능력은 통역 능력과 상관관계가 있으나 연어 능력은 통역 능력과 상관관계가 낮은 것을 발견하였다.

이창수(2020)는 기계번역물과 인간번역물의 어휘 사용을 비교 분석하여 그 차이점을 보고하였다. 분석 결과 기계번역물은 인간번역물에 비해 1) be동사의 사용율이 높고, 2) as, if 등 종속 접속사의 사용 비중이 높으며, 3) 인칭대명사 사용율이 낮고, 4)고빈도 일반명사의 사용의 반복 사용율이 높음을 발견하였다.

이상의 연구를 통해 보고된 내용은 인간 번역사는 번역 시 원천어 어휘의 풍부성을 보존하고자 노력하며, 유의어 지식이 뛰어나며, 기계번역에 비해 다양한 동사를 구사하고, 저빈도 일반명사를 사용하며, 어휘의 반복 사용을 지양한다는 것이다. 또한 번역사와 통역사의 어휘 능력은 번역 및 순차통역물의 품질에 영향을 미친다는 점도 발견되었다. 이를 종합하면 번역물에 사용된 어휘의 특성과 빈도를 통해 번역물의 성격이 영향을 받으며 따라서 번역물의 평가기준에 어휘의 풍부성, 어휘의 빈도 및 다양성이 포함되어야 함을 시사한다.

지금까지 통번역 평가에 있어 표현력/어휘력 항목은 주로 정성적 지표로 사용되어 '표현력/어휘력이 다채롭고 풍부하다-풍부하지 않다' 또는 '상, 중, 하'와 같은 이분법적 또는 삼분법적 평가에 의존해왔다. 이에 정량적으로 측정이 가능한 TTR과 어휘 빈도 정보를 어휘력/표현력 평가에 활용하면 보다 정교한 눈금으로 평가가 이루어져 번역 평가의 정교함과 신뢰성을 높일 수 있을 것으로 기대된다.

3. 연구 방법

3.1 참가자 및 데이터 수집

본 연구는 국내 통번역대학원 지원자 18명의 한영 번역물을 대상으로 이루어졌다. 번역에 사용된 출발어 한국어 텍스트는 2019년 국내 일간지에 실린 태풍 하기비스와 한일관계에 대한 기사를 수정한 것으로 총 122개 단어로 구성되었다(부록 참조). 참가자들은 통제된 공간에서 50분 동안 비슷한 분량의 영어 텍스트 하나와 한국어 텍스트 하나를 각각 영한, 한영으로 번역하였으며, 각 번

역에 대한 시간제한은 따로 두지 않았다. 참가자들은 수기로 번역을 수행하였으며 사전이나 인터넷 등 참고 자료 사용은 허용되지 않았다. 참가자 중 일부는 학부에서 통번역학을 전공한 경우도 있었으나 대다수는 비전공자로 통번역대학원 진학을 위해 사설 학원에서 통번역 수업을 수강해온 학생들이었다. 참가자들의 남녀 구성비는 5:13이었으며, 연령별로는 20대 11명, 30대 5명, 40대 1명, 50대 1명으로 구성되었다.

3.2 측정 도구

3.2.1 어휘 다양성 및 어휘 정교성

참가자가 수기로 작성한 번역물은 분석을 위해 연구자가 MS 워드로 타이핑하여 파일 형태로 만들었다. 분석 작업 중에는 사용된 어휘의 빈도수에 따라 어휘 정교성을 파악하는 작업이 포함되어 있어 오탈자로 인한 프로그램 어휘 인식 오류를 방지하기 위해 알파벳 2자 이하의 철자 오류는 타이핑 시 정정하여 입력하였다.

파일로 만들어진 번역물은 온라인 텍스트 분석 프로그램인 어휘 프로파일러(vocabulary profiler)를 사용하여 분석하였다(http://www.lextutor.ca/vp/eng/). 본 웹사이트는 캐나다 퀴벡 대학교의 콥(Tom Cobb) 교수가 운영하는 온라인 사이트로 어휘 프로파일러 이외에도 N-그램(N-Gram), 레인지(Range), 콘코던스(Corcordance) 등 어휘 관련 다양한 분석 툴을 제공한다. 어휘 프로파일러는 다양한 연구 활동에 활용되고 있으며(Cobb & Horst, 2001; Laufer & Nation, 1995; Meara & Fitzpatrick, 2000; Morris & Cobb, 2004) 분석 툴과 관련된 이론적 배경은 라우퍼와 네이션(1995)의 연구에 기술되어 있다.

어휘 프로파일러 창에 번역 텍스트를 하나씩 입력하면 아래 여섯 항목에 대한 분석 결과가 나온다.

1) 전체 어휘 수(token),
2) 반복 사용된 어휘를 제외한 어휘 수(type),
3) 전체 어휘 중 K1 어휘 비율,
4) 전체 어휘 중 K2 어휘 비율,
5) 전체 어휘 중 AWL 어휘 비율
6) 타입-토큰 비율(TTR)

3.3 데이터 분석

3.3.1 번역물 평가

수집된 번역물은 연구자와 제 2의 평가자가 각각 평가한 후 평균을 구해 최종 점수를 부여하였다. 제 2의 평가자는 20년 이상의 통번역 실무 경력을 가진 통번역대학원 강사로 통번역 수업을 3년 이상 한 해당분야 전문가이다. 번역 평가는 의미(충실성), 통사구조, 표현의 3가지 항목에 대해 각각 60, 20, 20을 만점으로 하는 정량적 평가 방식으로 이루어졌으며 각 항목별 점수를 합산하여 총점을 산출하였다.

3.2.2 어휘 다양성 및 정교성 측정

각 번역물을 온라인 텍스트 분석 툴인 어휘 프로파일러를 이용해 토큰, 타입, K1, K2, AWL어휘 비율과 TTR을 구하였다. 어휘 다양성을 측정하는 지표로는 프로파일러에서 산출된 TTR을 사용하였다. 어휘 정교성을 측정하는 지표로는 1K, 2K와 AWL 어휘 비율 중 어느 지표가 어휘 정교성을 가장 잘 대표하는지에 대한 선행 연구가 없어 1) 1K, 2) 2K, 3) AWL 어휘 비율, 4) 1K+2K 어휘의 합과 AWL어휘 간의 비율 총 4개의 지표를 산출하여 각각 변수로 사용하였다.

3.2.3 어휘 다양성 및 정교성과 번역 점수와의 관계 분석

어휘 다양성과 정교성을 위의 측정 지표를 통해 수치화 한 후 다중회귀분석을 통해 어휘 다양성 및 어휘 정교성과 번역 점수 간 관계를 분석하였다. 다중회귀분석은 종속변수와 독립변수와의 상관관계를 분석하는 통계 기법의 하나로 모집단의 종속변수에 대하여 다수의 독립변수가 존재할 때, 종속변수와 다수의 독립변수와의 관계를 규명하는데 사용하는 분석 기법이다. 상관분석이 변수 사이에 상관적 영향이 있는지를 분석하는 것이라면 회귀분석은 인과관계로서 독립변수가 종속변수에 얼마만큼 영향을 주는지를 분석한다. 이는 모집단의 종속변수 상의 편차가 독립변수를 통해 설명될 수 있는지 여부와 설명될 수 있다면 편차의 몇 퍼센트가 독립변수로 인해 설명되는가를 분석하는 것이다.

본 연구에서는 번역 성적이 종속변수로 어휘 다양성(TTR비율)과 어휘 정교

성(1K, 2K, AWL 어휘 비율)이 각각 독립변수로 사용되었다. 다중회귀분석을 통해 분석하고자 하는 바는 1)어휘 다양성과 어휘 정교성이 각각 또는 복합적으로 번역물 성적에 통계적으로 유의미한 영향을 미치는지 여부와, 2) 유의미한 영향을 미친다면 그 영향력이 어느 정도인지를 파악하는 것이다. 회귀분석은 IBM SPSS Statistics 22를 이용하여 이루어졌다.

4. 분석 결과

4.1 자동평가 프로그램에서 누락된 평가 항목 (연구질문 1)

인간 평가에 사용된 평가 항목과 자동평가 프로그램에 사용되는 평가 항목의 비교 및 현재까지 발표된 통번역과 어휘와의 관계에 대한 연구 결과를 종합해 볼 때 자동평가 프로그램에 누락된 평가항목은 '표현력'인 것으로 파악되었다. 이는 지금까지 번역 평가에 있어 표현력은 주로 정성적 항목으로 평가되어 왔기 때문에 자동평가 프로그램에 반영되지 않은 것으로 보인다. 표현력을 계량화하여 측정할 수 있는 지표가 제시된다면 자동평가 프로그램의 평가 항목으로 추가하여 평가 타당성을 개선하는데 도움이 될 것으로 보인다.

인접 학문인 언어학과 영어교육학의 선행연구를 통해 표현력을 측정하는 지표로 어휘 다양성과 어휘 정교성이 적합한 것으로 판단하였다. 이는 각각 텍스트의 TTR과 K1, K2, AWL어휘 비율을 통해 측정이 가능하여 표현력을 측정하는데 보다 정교한 눈금으로서의 역할을 할 수 있을 것으로 보인다.

4.2 어휘 다양성과 어휘 정교성과 번역 점수와의 관계 (연구질문 2)

4.2.1 번역물 점수

두 평가자가 각각 채점한 번역 점수에 대한 상관관계를 분석한 결과 높은 양의 상관관계를 보였다 (r = .93). 일반적으로 상관관계 분석에서 .9 이상의 상관관계는 두 변인 간 높은 상관관계가 있음을 보여준다. 즉, 두 평가자가 각각의 번역물에 부여한 번역 점수는 대체로 일치하며 이는 번역물에 부여된 점수가 일정 수준 이상의 평가자간 신뢰도(inter-rater reliability)를 가짐을 보여준다.

두 평가자의 번역점수를 평균한 최종 번역점수에 대한 기술통계는 아래와 같다. 최소값과 최대값 간 점수차이가 36점으로 비교적 큰 편이며 표준편차도 8.98로 높은 편인 점으로 볼 때 모집단 내 번역 능력(점수)은 어느 정도 편차가 있는 것으로 볼 수 있다.

〈표 1〉 번역 점수 평균치에 대한 기술통계

	최소값	최대값	평균	표준편차
평균 번역 점수 (100점 만점)	56	92	79.7	8.98

4.2.2 어휘 프로파일러 분석 결과

온라인 분석 툴인 어휘 프로파일러에 텍스트를 입력하여 얻은 각 번역물의 1) TTR, 2) K1 어휘 비율, 3) K2 어휘 비율, 4) AWL 어휘 비율은 아래 표2와 같으며 이에 대한 기술통계는 표3에 제시되어 있다.

〈표 2〉 번역물별 어휘 다양성 및 어휘 정교성

	TTR	K1 어휘 (%)	K2 어휘 (%)	AWL (%)
번역 1	75.19	73.68	5.26	5.26
번역 2	73.46	72.84	6.17	5.56
번역 3	72.85	68.87	8.61	9.27
번역 4	73.88	69.4	5.22	9.7
번역 5	66.89	67.55	6.62	7.95
번역 6	73.72	72.26	4.38	4.38
번역 7	69.13	74.5	4.7	6.71
번역 8	68.42	72.51	5.26	5.26
번역 9	67.78	72.78	5.56	9.44
번역 10	64.84	75.27	3.85	3.85
번역 11	73.72	71.79	8.97	6.41
번역 12	75.41	72.95	4.92	3.28

번역 13	58.23	77.22	7.59	3.16
번역 14	68.28	66.21	8.28	6.9
번역 15	70.81	73.91	6.21	8.07
번역 16	68.53	75.52	5.59	4.2
번역 17	68.85	78.69	6.01	2.73
번역 18	65.52	73.4	5.42	5.91

〈표 3〉 번역물의 어휘 다양성 및 어휘 정교성 요약

	최소값	최대값	평균	표준편차
TTR	58.23	75.41	69.8	4.39
빈도 1000	66.21	78.69	72.74	3.18
빈도 2000	3.85	8.97	6.03	1.46
AWL	2.73	9.7	6	2.21

어휘 다양성을 나타내는 TTR의 경우 최소값과 최대값 간 차이가 17.18이고 표준편차가 4.39로 번역문에 따라 어느 정도 편차가 있음을 보여준다. 즉, 번역물에 따라 동일한 단어가 여러 번 반복 사용된 텍스트와 동일한 단어의 반복 사용이 적은 텍스트간 어느 정도 차이가 있다는 것이다. 그에 비해 어휘 정교성을 나타내는 어휘 빈도 별 사용의 경우 3가지 모두 최소값과 최대값 간 차이가 크지 않고 표준편차 또한 TTR에 비해 상대적으로 낮아 번역물 간 차이는 뚜렷하지 않은 것으로 보인다.

4.3 회귀 분석 결과 (연구 질문 2&3)

4.3.1 단일 변수와의 상관관계: 단순 회귀분석

영어 어휘 다양성과 어휘 정교성이 한영번역 점수에 각각 미치는 영향을 회귀분석을 통해 분석한 결과는 아래 표4와 같다.

〈표 4〉 번역 점수와 단일 변수와의 단일회귀분석 결과

종속변수	독립변수	R^2	p 검증	계수
번역 점수	TTR	.25	p = .03	1.02
	K1 어휘	.00	p = .80	-0.17
	K2 어휘	.12	p = .15	-2.10
	AWL	.00	p = 95	0.06

회귀분석에서 p 검증은 각 분석이 통계적으로 유의미한 결과인지 여부를 나타낸다. p 값이 0.05보다 적은 경우, 분석 결과가 95% 신뢰 수준에서 통계적으로 유의미한 결과임을 의미하며 p 값이 0.01보다 적은 경우, 99% 신뢰수준에서 분석 결과가 통계적으로 유의미한 결과임을 나타낸다. 단순 회귀분석에 사용된 4가지 독립변수 중 p 값이 .05보다 적게 나온 것은 TTR이 유일하다. 이는 TTR에 대한 분석 결과만 통계적으로 유의하며 나머지 3개 독립변수에 대한 분석 결과는 통계적으로 유의미하지 않음을 의미한다. 이는 4개의 독립변수 중 TTR만 번역 점수에 통계적으로 유의미한 영향을 가짐을 보여준다.

회귀 분석에서 R^2은 독립변수가 종속변수에 미치는 영향의 정도를 나타낸다. 앞에서 p 값을 살펴보았을 때 p >.05 인 K1, K2, AWL에 대해서는 분석 결과가 유의미하지 않으므로 R^2 값도 의미가 없다. 종속변수 중 p 값이 .05보다 큰 TTR만 R^2 값을 통해 어휘 다양성이 번역 점수에 미치는 영향을 가늠할 수 있다. 위의 <표 4>에서 종속변수 TTR의 R^2 값이 .25라는 것은 번역 점수 중 25%는 TTR 차이로 인한 것임을 의미한다. 이는 다시 말해 번역 점수의 25%는 해당 번역물에 동일 어휘의 반복 사용 정도에 따라 영향을 받는다는 것을 뜻한다.

계수(coefficient)는 각 종속변수가 독립변수에 미치는 영향의 정도를 나타내는 상관계수이다. 특정 종속변수의 계수가 1이라는 것은 해당 종속변수가 1단위 변화할 때 독립변수도 1단위 변화함을 의미한다. 본 분석에서는 p 값이 .05보다 작은 TTR의 계수만 의미를 가진다. 위의 <표 4>에서 TTR의 계수는 1.02이다. 이는 TTR이 1.02% 높아질 때마다 번역 점수가 1점씩 올라감을 의미한다.

요약하면, 각 독립 변수의 영향을 개별적으로 분석한 단순 회귀분석에서는 신뢰수준 95% 범위에서 TTR만이 번역점수에 통계적으로 유의미한 영향을 미

치는 것으로 나타났으며, TTR이 번역 점수에서 차지하는 비중은 25%이고, TTR이 1.02% 상승할 때 번역 점수가 1점 올라가는 영향력을 가지는 것으로 나타났다.

4.3.2 복합 변수와의 상관관계: 다중 회귀분석

다중회귀분석은 다수의 종속변수가 동시에 집합적으로 독립변수에 미치는 영향을 보여준다. 다중 회귀분석에서 종속변수의 개수가 늘어난다고 해서 독립변수에 대한 예측력이 올라가는 것은 아니다. 종속변수에 영향을 미치지 않는 변수를 추가하면 오히려 종속변수에 대한 설명력이 떨어지게 된다. 본 연구에서는 어휘 다양성과 어휘 정교성의 영향이 복합적으로 작용할 때 번역점수에 미치는 영향이 어느 정도인지 파악하기 위해 다중회귀분석을 실시하였으나 단일회귀분석 결과 4가지 변수가 번역점수에 미치는 복합적인 영향력은 통계적으로 유의미하지 않은 것으로 나타났다($F(4, 17) = 1.832$, $p > .05$, $R^2 = .36$). 즉, 4가지 독립변수를 모두 이용하는 것이 TTR 하나의 변수만을 이용하는 것보다 번역 점수를 예측하는 데 있어 예측력이 떨어지는 것으로 나타났다. 이는 어휘 다양성과 어휘 정교성 두 가지 변수가 복합적으로 작용할 때 보다 어휘 다양성이 단독으로 작용할 때 번역 점수에 미치는 영향이 통계적으로 의미있음을 뜻한다. 4가지 독립변수를 모두 넣고 실행한 다중회귀분석 결과는 아래 표 5에 제시되어 있다.

〈표 5〉 번역 점수와 복합 변수와의 다중회귀분석 결과

종속변수	독립변수	ß	t	*p*
번역 점수	(상수)			
	TTR	.959	1.99	**.06**
	K1 어휘	-.157	-.19	**.84**
	K2 어휘	-.177	-1.35	**.19**
	AWL	-2.01	-.13	**.89**

5. 논의 및 결론

5.1 분석 결과에 대한 논의

본 연구는 시중에서 사용되고 있는 자동번역평가 프로그램의 타당성을 향상시키기 위해 추가할 수 있는 평가 항목을 파악하기 위해 진행되었다. 인간 평가와 자동 평가 항목 간 비교 및 언어학과 영어교육학, 통번역학의 선행연구에 통해 분석한 결과 '표현력'이 자동평가 프로그램에서 빠져있는 것으로 평가 항목으로 도출되었다. 선행연구에 근거해 '표현력'을 구체화한 측정 가능한 평가 지표로 어휘 다양성과 어휘 정교성을 제시하였으며, 어휘 다양성은 TTR을 통해, 어휘 정교성은 어휘 빈도수를 통해 측정할 수 있음을 제시하였다.

위에서 도출된 평가 항목 및 그 측정 지표가 번역 점수에 영향을 미치는지 실증적으로 검증하기 위해서는 통번역대학원 지원자 18명의 번역물을 분석하였다. 통계기법을 통해 각 번역물의 어휘 다양성(TTR), 어휘 정교성(어휘 빈도수)과 번역 점수 간의 관계를 분석한 결과 어휘 다양성은 번역 점수와 통계적으로 유의미한 관계를 가지는 것으로 나타났다. 보다 구체적으로는 번역 점수의 25%는 어휘 다양성의 차이에 기인하며, 어휘 다양성 지표(TTR)가 1.02% 증가할 때 번역 점수는 1점 증가하는 것으로 나타났다. 반면 어휘 정교성은 번역 점수와 통계적으로 유의미한 관계를 가지지 않는 것으로 나타났다.

어휘 다양성이 번역 점수와 통계적으로 유의미한 관계를 가지는 것은 선행연구와 일치하는 결과로 예상되었던 결과이다. 다수의 영작문 평가 연구에서 어휘 다양성은 텍스트 품질에 통계적으로 유의미한 관계를 가지는 것으로 나타났다. 번역과 관련된 선행 연구에서도 어휘 능력이 번역 점수 편차의 큰 부분을 차지하고(김훈밀, 2014), 유의어 능력은 통역사와 비통역사를 구분 짓는 변별 항목인 점은(정혜연, 2016) 본 연구의 결과와 결을 같이 한다. 한편 본 연구를 통해 어휘 다양성이 번역 점수에 대해 25%의 예측력을 가진다는 점이 드러난 것은 새로운 발견이며 이는 여러 번역 점수의 평가 항목 중 '표현력'에 부여할 가중치의 적정수준을 제시하는 역할을 한다. 이는 향후 번역평가 모델을 구축하는데 유용한 자료로 사용될 수 있으며 통번역 교육에도 시사점을 가진다.

어휘 빈도가 번역점수와 통계적으로 유의미한 관계를 가지지 않는다는 결과는 선행연구의 결과와 일치하지 않는 결과이다. 그 이유는 세 가지 정도로 생

각해볼 수 있는데 첫째는 어휘 빈도가 영작문 성적에 미치는 영향이 어휘 다양성에 비해 크지 않다는 선행연구에서 찾을 수 있다(Gonzales, 2017). 어휘 다양성은 영작문 성적에 통계적으로 유의미한 영향을 미치며, 우수 영작문과(4,5등급)과 중간등급(2,3등급)을 구분하는 주요 요인으로 나타났다. 이를 통해 영작문 평가에 덜 중요하게 작용하는 어휘 정교성이 한영번역 평가에서도 상대적으로 낮은 영향력을 가지는 것은 아닌지 유추해볼 수 있다. 둘째로는 번역 주제의 영향을 들 수 있다. 즉, 선행연구와 본 연구에서 사용한 번역 텍스트의 주제가 동일하지 않은 것이 결과의 차이를 가져온 것일 수 있다. 어휘 다양성 및 어휘 정교성과 영작문과의 관계를 분석한 류(Ryoo, 2018)의 연구에서는 영작문의 주제에 따라 어휘 정교성과 어휘 다양성이 영향을 받는 것으로 나타났는데 본 연구에서 사용한 번역물의 주제가 가지는 특수성 때문에 어휘 정교성이 유의미하지 않은 결과를 나타났을 가능성이 있다. 마지막으로 본 연구의 모집단의 수가 18명으로 작은 데에 원인이 있을 수 있다. 일반적으로 두 개의 종속변수로 분석을 하기에는 18명의 모집단이 작은 규모이며, 모집단이 작을수록 통계분석에서 유의미한 결과를 도출하기는 어려운 경향이 있다. 이는 추후 보다 큰 모집단을 대상으로 동일한 분석을 다시 시행하여 확인할 필요가 있을 것으로 보인다.

5.2 연구의 한계와 의의

본 연구는 18명이 번역한 200단어 이하의 번역물을 분석한 것으로 모집단의 크기가 매우 작은 소규모 연구이다. 따라서 본 연구 결과를 일반화 하는데는 주의가 필요하며 보다 큰 모집단을 통해 연구 결과를 강화할 필요가 있다.

그럼에도 불구하고 본 연구는 번역 점수에 영향을 미치는 항목으로 어휘 다양성과 어휘 정교성을 살펴본 국내 첫 연구라는 점에서 의의가 있으며 연구 결과 도출된 TTR은 비교적 손쉽게 평가 프로그램에 반영할 수 있는 항목으로 자동번역평가 모델 보완에 즉각 활용될 수 있다는 점에서도 의미가 있다.

또한 본 연구결과는 어휘 다양성이 번역 점수에 미치는 영향력을 특정함으로써 교육 및 학생 지도에 있어서도 다양한 어휘력 강화 프로그램을 도입하는데 이론적 근거로 활용될 수 있으며 이를 통해 학생들의 번역 능력 향상에 기여할 수 있을 것으로 기대된다.

참고문헌

김재희 (2008)「문화 관련 어휘 번역 방법 연구-코리아나 아랍어 번역 텍스트 분석을 중심으로」,『통역과 번역』10(1): 25-42.

김훈밀 (2014)「학부생 번역능력의 하위능력 검증 －출발어 어휘지식, 출발어 독해능력과 주제지식을 중심으로」,『통역과 번역』16(1): 1-24.

김훈밀 (2019)「다층위 어휘지식의 순차통역 수행 예측력」,『통역과 번역』21(2): 31-54.

유정화 (2016)「번역인증제도(이론편)」,『한국외국어대학교 통번역연구소 학술대회』, 13-21.

이창수 (2020)「한-영 신문사설 번역에서 나타나는 인간번역과 기계번역 간의 어휘 사용 차이 연구」,『통역과 번역』22(1): 245-262.

정혜연 (2016)「전문통역사 머릿속 사전의 구조-통합어에 대한 경험연구」,『번역학연구』18(1): 147-169.

정혜연 (2018)「번역의 자동평가: 기계번역 평가를 인가번역 평가에 적용해보기」,『통번역학 연구』22(4): 265-287.

한국외대 번역평가인증 연구팀 (2016)「번역인증제도 (실무편)」,『한국외대 통번역연구소 학술대회 <언어, 통번역의 평가 및 인증> 발표집』: 22-33.

Akbari, Alireza (2018) 'Translation Quality Research A data-driven collection of peer-reviewed journal articles during 2000-2017', *Babel* 64(4): 548-578.

Bachman, Lyle and Adrian Palmer (1996) *Language Testing in Practice,* Oxford: Oxford University Press

Chodorow, Martin and Jill Burstein (2004) 'Beyond Essary Length: Evaluating e-rater's Performance on TOEFL Essays (Research Report No. 73)', Educational Testing Service.

Coxhead, Averil (2000) 'A New Academic Word List', *TESOL Quarterly* 34(2): 213-238.

Crossley, Scott, Tom Salsbury and Danielle McNamara (2012) 'Predicting the Proficiency Level of Language Learners using Lexical Indices', *Language Testing* 29(2): 243-264.

Crossley, Scott, Tom Salsbury, Danielle McNamara and Scott Jarvis (2010) 'Predicting Lexical Proficiency in Language Learner Texts using Computational Indices', *Language Testing* 28(4): 561-580.

Duran, Pilar, David Malvern and Brian Richards (2004) 'Developmental Trends in Lexical Diversity', *Applied Linguistics* 25(2): 220-242.

Gonzales, Melanie (2017) 'The Contribution of Lexical Diversity to College-Level Writing', TESOL Journal 8(4), 899-919.

Graesser, Arthur, Danielle McNamara, Max Louwerse and Zhiqiang Cai. (2004) 'Coh-Metrix: Analysis of text on cohesion and language', *Behavior Research Methods, Instruments & Computers* 36(2): 193-202.

Graesser, Arthur, Danielle McNamara, and Jonna Kulikowich (2011) 'Coh-metrix: Providing multilevel analyses of text characteristics', *Educational Researcher* 40(5): 223-234.

Kyle, Kristopher and Scott Crossley (2015) 'Automatically Assessing Lexical Sophistication: Indices, tools, findings and application', *TESOL Quarterly* 49(4): 757-786.

Laufer, Batia and Paul Nation (1995) 'Vocabulary Size and Use: Lexical Richness in L2 Written Production', *Applied Linguistics*, 16(3): 307-322.

Lavie, Alon and Michael Denkowski (2009) 'The METEOR metric for automatic evaluation of machine translation', *Machine Translation* 23: 105-115.

Lee, Siok H (2003) 'ESL Learners' Vocabulary Use in Writing and the Effects of Explicit Vocabulary Instruction', *System* 31(4): 537-561.

Morris, Lori and Tom Cobb (2004) 'Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees', *System* 32(1): 75-87.

Nation, Paul and DAvid Beglar (2007) 'A Vocabulary Size Test', The Language Teacher 31(7): 9-13.

Papineni, Kishore, Salin Roukes and Wei-Jing Zhu (2002) Proceedings of the 40th Annual Meeting of the *Association for Computational Linguistics*

(ACL), Philadelphia: 311-318.

Read, John (2000) *Assessing vocabulary*, Cambridge: Cambridge UP

Ryoo, Young-sook (2018) 'Comparing Lexical Diversity and Lexical Sophistication in Korean EFL Writing: Topic and Text Length', *Multimedia Assisted Language Learning* 21(3): 63-87.

internet source

IbisWorld report

https://www.ibisworld.com/united-states/market-research-reports/translation-services-industry/

부록

한영 번역에 사용된 출발어 한국어 텍스트

일본 덮친 태풍, 위로하는 성숙함 보여야

지난 주말 일보을 강타한 태풍 하기비스로 일본 열도는 사망·실종자만 50명이 넘는 타격을 입었다. 일부 지역에는 48시간 동안 약 1,000mm의 비가 내리는 등 관측 사상 최대 강수량을 기록하면서 21개 제방이 무너졌고 142개 강이 범람했다. 또한 2011년 동일본 대지진 때 원전사고가 일어난 후 후쿠시마현에서는 오염 제거 작업을 통해 수거한 방사성물질 폐기물 자루들이 불어난 강물에 유실되는 사고가 있었다.

천재지변이란 인간의 힘으로는 어쩔 수 없는 재해다. 그래서 피해자를 힘닿는 한 돕고 고통을 나누는 게 동서고금 인간의 도리다. 국가 간에도 아웃나라가 재난을 입었을 때 함께 안타까워하고 하루빨리 피해에서 벗어나도록 격려해 주는 게 도리일 것이다. 일본의 수출 규제 후 대일 여론이 매우 부정적인 것은 어쩔 수 없는 현실이라 해도, 정치와 무관한 재난 소식에 대해 일부 누리꾼이 증오가 섞인 반응을 보이는 것은 우려스럽다.

[Abstract]

The Effects of Lexical Diversity and Lexical Sophistication of English on Korean-English Translation

Kim, Hoonmil
(International Graduate School of English)

This study investigates ways to improve the validity of automatic translation assessment tools. Comparison of assessment criteria used by human raters against those employed by automatic assessment tools show that automatic assesment tools are geared towards evaluating ‘meaning’ and not so much towards ‘expression’. Lexical diversity, as measured by type-token ratio, and lexical sophistication, as measured by lexical frequency, are identified as potential criteria that can improve validity of automatic translation assessment when added to automatic assessment tools.

To empirically test the effects of lexical diversity and sophistication on translation scores, Korean-English translation of 18 applicants for an interpreting and translation graduate school in Korea were collected and analyzed using an online lexical profiling tool followed by multiple regression analyses designed to validate their significance.

Analyses show that lexical diversity has statistically significant correlation with translation scores, explaining 25% of variance in translation scores. On the other hand, lexical sophistication showed no statistically significant correlation with translation scores. Implications and limitations of the study are discussed.

▸ Key Words: lexical diversity, lexical sophistication, automatic translation assessment, translation evaluation criteria
▸ 주제어: 어휘 다양성, 어휘 정교성, 자동번역평가, 번역 평가 항목

김훈밀
국제영어대학원대학교 조교수
milliek@hanmail.net
관심분야: 통번역 평가, 통역교육, 번역교육, 텍스트 분석

논문투고일: 2020년 5월 5일
심사완료일: 2020년 5월 21일
게재확정일: 2020년 5월 25일