http://dx.doi.org/10.15749/jts.2019.20.2.007

번역학연구
2019 여름호 제20권2호

# 4차 산업혁명 시대에서 번역학의 방향: 빅데이터로서의 코퍼스의 역할과 기능*

조 준 형
(고려대)

## 1. 서론

상이한 문화와 역사를 지닌 민족을 온전히 이해한다는 것은 매우 어려운 일이며, 그 민족의 사상을 고스란히 담고 있는 언어를 완벽하게 해석해 낸다는 것은 더더욱 힘든 일이다. 신화 속의 바벨탑(Tower of Babel) 그리고 현실 속의 바벨탑인 에스페란토어(Esperanto)는 이러한 문제를 극복하기 위한 인간의 노력이었다. 그리고 바벨탑이 무너지며 번역은 태어났다.[1)]

번역은 지난 세기에 걸쳐서 '번역이란 무엇인가?'라는 개념적·철학적 질문

* 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017S1A5B5A07058079).

1) 심층적으로는 차이가 있지만, 일반론적 차원에서 '통역'도 '번역'과 같은 맥락에서 이해할 수 있을 것이다. 따라서 본 논문에서 '통번역'이 아닌, '번역'이라고 표현하였더라도, 넓은 의미에서 '통역'의 개념도 포함되어 있다.

과 ‘번역을 어떻게 할 것인가?’라는 실천적 질문을 중심으로 수많은 학자들이 자신의 주장을 내세운 논의의 장이었다. 인문학적 사유의 대상이었던 번역은 20세기에 들어서 새로운 환경을 맞이하게 된다. 바로 기계번역의 등장이다. 기계번역은 1950년대 러시아어와 영어 사이의 자동 번역에서 시작되어(윤애선 2019; Guidère 2008), 번역을 인문학 연구에서 공학적 연구의 대상으로까지 확대했다. 그렇지만 수준 낮은 번역 결과물은 기계번역을 무시하는 결과를 초래했다. 번역 연구자들이 기계번역에 대해 본격적인 관심을 갖기 시작한 것은 1980-90년 컴퓨터 능력이 획기적으로 발전하면서 언어 데이터, 즉 코퍼스를 활용한 통계기반 기계번역 방식이 등장하면서부터이다. 이때부터 기계에 의한 번역 품질이 상당히 개선되었고, 이와 관련된 연구도 활발하게 진행되었다.

21세기의 제4차 산업혁명은 이러한 흐름을 가속화시켰다. ‘인공지능(Artificial Intelligent)’과 ‘빅데이터(Big Data)’라는 4차 산업혁명의 핵심 개념은 거의 인간과 다름없는 번역 기계의 등장을 예고했고, 실제로 특정 분야에서는 상당한 수준의 결과물을 내놓았다. 그리고 이것이 인문학적으로도 실무적으로도 번역학의 위기라는 두려움을 가져왔다. 비약적인 발전을 이룬 기계번역이 인간 통번역가를 대체할지 모른다는 두려움이 바로 그것이다. 그렇지만, 이 위기감은 곧 인간과 기계의 상생을 위한 논의로 바뀌었다. 기계가 인간을 대체하는 것이 아니라, 번역지원도구(Computer-Assisted Translation Tools)와 같이 기계가 인간의 번역 활동에 도움을 줄 수 있으며, 반대로 인간의 번역 활동은 번역 기계의 수준을 높이기 위한 학습 자료를 제공하기 때문이다. 4차 산업혁명의 두 핵심 개념들 중 ‘인공지능’보다 ‘빅데이터’가 중요한 이유가 바로 여기에 있다. 인공지능 기계번역이라고 할지라도 근본은 ‘빅데이터’, 다시 말해서 대규모 코퍼스를 바탕으로 하기 때문이다. 코퍼스를 통한 학습 없이 현재의 기술로는 완벽한 기계번역은 불가능하다. 그리고 인간 연구자나 실무자에게도 이 데이터는 연구와 실무를 위해서도 매우 중요하다. 즉 코퍼스는 번역학에 있어서 인간과 기계를 이어주는 공통분모가 될 수 있다.

이러한 관점에서 본 논문은 1960년대 이래로 언어학을 비롯한 언어 관련 연구에서 보편적 도구가 된 코퍼스가 데이터의 가치가 높아진 4차 산업혁명 시대에서 빅데이터로 어떻게 재정립할 수 있으며, 이를 위해 10여 년 전부터 새롭게 조명받기 시작한 웹 코퍼스가 하나의 대안이 될 수 있음을 보이고자 한다.

## 2. 4차 산업혁명과 통번역학

21세기 들어 "4차 산업혁명"은 사회 전반에서 가장 많이 들리는 화두 중 하나이다. 앞서 세 차례의 산업혁명이 있었지만, 4차 산업혁명은 우리에게 전혀 다른 모습으로 다가오고 있다. 현재의 산업혁명을 포함해서 모두 4차례의 산업혁명의 특징을 요약하면 다음과 같다(하원규와 최남희 2016: 31; 최희섭 2017: 2).

- **1차 산업혁명**(18세기) - 농업혁명 : 증기기관, 농업의 산업화, 인간의 정주화
- **2차 산업혁명**(19세기-20세기 초) - 산업혁명 : 전기, 공업의 산업화, 인간의 육제노동 대체
- **3차 산업혁명**(20세기 후반) - 정보혁명 : 컴퓨터와 인터넷, 정보처리와 커뮤니케이션혁명, 인간의 두뇌 노동 대체
- **4차 산업혁명**(현재) - 만물초지능혁명 : 인공지능과 빅데이터, IoT·CPS, 인공지능 혁명, 인간의 두뇌 노동을 사물, 기계, 공간으로 외부화

1-2차 산업혁명은 사회 전반에 걸쳐 커다란 영향을 미쳤지만, 현재의 우리는 이 산업혁명의 결과의 시대에 살고 있어서, 우리는 그 충격을 간접적으로 느낄 수밖에 없다. 이에 반해 3-4차 산업혁명은 우리가 실제로 체험하였고 여전히 진행되고 있기 때문에, 그 영향이 직접적일 수밖에 없다. 넓은 의미에서 3-4차 산업혁명은 컴퓨터 및 정보화를 바탕으로 하고 있으며, 인간 노동력의 질적 발전(1-2차 산업혁명)을 넘어서 지식 정보와 관련된 기술의 진화와 맞물려 있다. 지식과 정보의 교환이 획기적으로 늘어나고 그 영역도 확대되면서, 새로운 사회 변화를 창출했다. 정보화 사회는『제3의 물결(*The Third Wave)*』에서 앨빈 토플러(Alvin Toffler)가 예견했듯이, 인간 노동력의 상실과 빈부 격차 등 많은 사회적 문제점을 초래하기도 했지만, 좋은 의미든 나쁜 의미든 대량 생산과 자동화를 통한 인간의 편의성을 끌어내는 방향으로 나갔다. 그리고 이 시대부터 인간의 언어는 공학자들의 연구 대상이 되기 시작했다.

통번역학과 직접적인 관련이 없어 보였던 제3차 산업혁명은 앞으로 다가올 기계번역 시대의 문을 열어놓았다. 구글 번역기(Google Translation),[2] 네이버

2) 구글 번역기(https://translate.google.com)

파파고(Papago),[3] 한글과 컴퓨터의 지니톡(GenieTalk)[4]과 같은 번역기는 이 시대부터 이미 토대가 마련되기 시작했다고 할 수 있다. 그리고 제4차 산업혁명은 이 번역기의 전성기를 가져왔다. 크로닌(Cronin 2013)이 말하듯이, 우리는 이제 디지털의 시대에 살고 있으며, 번역에 대한 시각도 이전과 달라질 수밖에 없다. 특히 기계번역과 인터넷의 발전은 경계의 개념과 실시간 번역 그리고 번역 후처리(post-editing)와 관련된 새로운 번역 화두를 제기한다. 또한 이전 시기부터 시작된 자동화는 인공지능과 빅데이터를 무기로 더욱 강력하고 더욱 지능화된 자동화의 모습을 갖게 되었다. 통번역은 더 이상 인간만의 전유물, 다시 말해서 인문학적 사유의 대상이 아닌 시대가 되었다. 이것은 번역뿐만 아니라 언어와 관련된 모든 연구 분야에서 마찬가지다.

1960년대 언어 분석에 통계 기법이 활용되면서 언어에 관한 정량적 연구가 본격화되었지만, 인간의 언어를 수치로 표상하는 것이 불가능하다는 이론 언어학자들의 반박에 통계학자와 언어학자들 사이에 논쟁이 있었다(Guiraud 1960). 그렇지만 지금은 언어의 통계적 분석이 매우 보편화되고 있으며, 이러한 경향은 언어뿐만 아니라 인문학 전반에 걸쳐 나타나고 있다. 이제 여기에 공학의 영역이 중첩되고 있다. 아니 이미 인문, 통계, 공학은 서로 떨어질 수 없는 협력 관계에 있으며, 앞으로 그 관계는 더욱 굳건해질 것이다.

이를 가장 분명하게 확인할 수 있는 것이 바로 기계번역과 언어서비스업체(Language Service Provider)이다. 가장 최근의 한국번역학회 봄 학술대회(2019년 4월)를 비롯해서 최근 몇 년간 통번역학 관련 학회를 보면 학술연구자뿐만 아니라, Flitto, Saltlux, SDL Korea 등 다수의 언어서비스업체가 참여하여 서로 간의 정보를 공유하고 협력을 모색하고 있다. 연구자들은 업체들을 통해서 최근 번역공학에 관한 정보를 수집할 수 있고, 업체들은 통번역 대학(원) 출신의 학생들을 채용할 기회가 되고 있다. 이것은 현재 시기에 통번역학 연구가 새로운 모습으로 변모할 기회가 된다. 고대 시기부터 이어져 온 통번역학의 모습은 순수한 이론 연구였지만, 그 속에는 실무적인 경향도 분명히 내재되어 있다. 번역이란 무엇인가?의 출발점은 번역 결과물의 관찰에서 출발하기 때문이다. 번역은

3) 파파고(https://papago.naver.com/)
4) 지니톡(http://www.interfree.com/if/productService/genieApp.do)

추상적인 성찰의 대상이 아니다. 인간의 활동이기 때문이다. 번역이라는 인간의 활동 없이 어떻게 번역을 논할 수 있을 것인가. 이것은 언어에 관한 순수 연구와는 다르다. 따라서 타 학문에 비해 실무적인 측면이 강하게 나타나며, 실제로 통번역 교육 기관에서 진행되는 교과과정에는 실무차원의 교육이 두드러진다. 그렇지만 또한 번역은 추상적인 성찰의 대상이다. 번역 역시도 궁극적으로 언어이기 때문이다. 정용목(2018)에 따르면, 번역은 읽기와 번역가의 글쓰기이다. 핵심이 언어이기 때문에, 인문학 연구일 수밖에 없다. 이향(2019)의 '번역학의 과제'에 대한 논쟁은 이러한 점에서 매우 중요한 주제이다. 단순히 기술적인 차원의 영역이 아니라, 온전히 인문학적인 연구로서 번역학에 대한 문제 인식은 상대적으로 타 학문 연구자들에게 번역학이 가지는 약한 위상을 강화시킬 수 있는 근본적인 주제이다. 그리고 이러한 성찰이 이론과 실재의 번역학을 온전한 학문 영역으로 자리매김할 수 있게 하는 토대가 된다. 학술적 차원의 번역 연구 그리고 번역 공학, 번역학 내에서 서로 대립이 되면서도 협치를 해야 하는 두 영역이다. 4차 산업혁명은 이를 분명하게 표면화시켰다. 이전부터 기계번역은 존재했지만, 번역학 내에서 크게 문제시되지 않았다. 몇몇 소수의 연구자를 제외하고 서로 별개의 영역인 것처럼 행동해왔다. 언어서비스업체도 마찬가지였다. 완전히 독립된 실체인 것처럼 보였다. 4차 산업혁명은 서로 외면했던 이 시선들이 서로를 바라보게 했다. 그렇지만 그 시선은 처음에는 불균등해 보였다.

1977년에 처음 개봉된 영화 스타워즈(Star Wars)에는 우주의 모든 언어를 통역할 수 있는 로봇이 등장한다. C-3PO라는 이름을 가진 이 로봇은 우주에 존재하는 대부분의 언어를 통역할 수 있다. 가장 이상적인 인공지능 통역가라고 할 수 있다. 현재의 기술로는 아직 불가능하지만[5] 앞으로 자동번역 분야의 연구자들이 목표로 하는 최종 결과물이 바로 이 로봇일 것이다. 그리고 이 이상적인 기계번역의 모습이 인간 통번역가들에게 위기감을 가져왔다. 학술 연구자가 반대의 위치에 있었던 번역공학과 언어서비스업체에 관심을 갖게 되는 순간이다. 자연스러운 학문적 발전이라기보다는 어쩔 수 없는 시대에 대한 순응인 것인가.

5) 우주가 아니라 지구상에 존재하는 언어를 통역 혹은 번역할 수 있는 기계도 아직 존재하지 않는다. 현재 가장 많은 언어를 자동 번역할 수 있는 구글 번역기조차도 104개의 언어를 번역할 수 있을 뿐이며, 지구상에 존재하는 모든 언어를 대상으로 하지 않는다.

### 3. 통번역학의 위기인가? 통번역가의 위기인가?

앞서 언급한 영화 스타워즈Star Wars의 로봇 C-3PO는 인간형의 로봇으로 영화의 여러 등장인물과 함께 우주를 누비면서 외계 종족을 비롯해서 다른 로봇과 인간과의 의사소통을 돕는다. 만일 지금 이러한 로봇이 존재한다면, 아마도 인간 통번역가는 설 자리를 잃게 될지도 모른다. 그리고 이러한 내용이 4차 산업혁명이 시작되고 '인공지능'과 '빅데이터'가 쟁점으로 부각되면서, 인간 통번역 연구자들이 학술대회 및 글을 통해서 표출한 두려움이었다. 그러나 그 두려움은 바로 협업의 필요성을 불러왔다. 최희섭(2017)은 인공지능에 의해 다수의 일자리가 사라지고, 번역가도 이러한 문제를 생각해 보아야 하지만, 인공지능이 아무리 발전한다고 하더라도 결코 대체할 수 없는 인간의 창조적 영역이 있다고 주장한다. 여기서 우리는 분명히 인식해야 할 부분이 있다. '인공지능'의 발달이 '통번역학의 위기'인지 '통번역가의 위기'인지를 분명히 하는 것이다. 현대적 개념의 번역학이 20세기 후반, 무넹(Mounin 1963), 스타이너(Steiner 1988), 투리(Toury 1995), 베이커(Baker 1993, 1995) 등의 일련의 학자들의 논저들과 함께 시작되었다고 하지만,[6] 번역에 대한 사유는 고대로 거슬러 올라간다(Ballard 1992 참조). 따라서 학문적 연구로서 통번역학은 아무리 인공지능이 발달한다고 하더라도 기계가 이러한 영역을 대체할 수는 없다.

그렇다면 4차 산업혁명은 통번역가의 위기인가? 위에서 언급한 로봇이 등장한다면 이 질문에 대해서 그렇다라고 답할 수도 있을 것이다. 윤애선(2019), 송연석(2018), 장애리(2017)에서 볼 수 있듯이, 기계번역은 규칙기반에서 통계기반으로 그리고 신경망으로 발전을 거듭하면서 상당한 수준의 번역 결과물을 내놓게 되었다. 그리고 실용적인 영역에서 번역은 고품질의 번역 결과물을 만들어낸다. 2017년 2월에 세종대학교에서 국제통역번역협회 주최로 인간과 인공지능의 번역 대결이 있었다. 네이버 파파고와 기계번역 기업인 시스트란(Systran)이 참여한 이 대회에서 결과는 인간의 승리로 끝났다.[7] 비록 속도는

6) 해석이론 및 비교문체론을 비롯한 번역의 현대적 이론 및 개념들은 대부분이 이 시기에 정립되었다고 할 수 있다.

7) 「인간과 AI의 번역 대결은 전문 번역사 승리 유력」, 『연합뉴스』, 2017.02.16, 2019년 3월 17일 검색.

인공지능이 빨랐지만, 정확도 및 표현에 있어서 기계가 인간을 앞설 수는 없었다. 더욱이 문학 텍스트에서 기계는 훨씬 저조한 수준을 보였다. 기계가 번역을 아무리 잘한다고 하더라도, 기계가 상대적으로 가지는 장점은 속도일 뿐이다. 번역의 정확성이 기존보다 획기적으로 개선되었다고 하지만, 우리가 주목해야 할 부분은 바로 '그럴 듯한 번역', 인간이 한 것과 '거의 유사한' 번역이라는 것이다. '이해할 수' 있는 번역, '상당히 자연스런' 번역, 이 모든 표현은 기계가 한 번역 결과물이 완벽하다는 것을 의미하는 것이 아니다. 그리고 비교 대상은 항상 '인간'이다. 언어는 수학 공식이 아니며, 바둑과 같은 게임이 아니다.

이러한 관점에서 '인공지능'이 아무리 발달한다고 하더라도 C-3PO와 같은 로봇의 출현은 바벨탑의 신화를 만들어내지 않는 이상 불가능하다고 할 수 있다. 적어도 지금으로선 그렇다. 앞으로 기술의 발전이 어떻게 변할지 모르지만, 남기춘(2019)이 말하였듯이 기술적인 측면에서 인공지능의 수준은 여전히 답보 상태에 있으며 이런 로봇의 출현은 신화의 영역에 속할지도 모른다. 물론 언어서비스업체에서는 이와 반대되는 시각을 가지고 있으며, 인공지능 번역기에 대한 매우 희망적인 태도를 보인다. 그런데 업체들의 주장을 보면 인공지능 번역기의 발전에는 데이터가 중요하다고 강조하고 있다. 데이터 다시 말해서 '빅데이터'의 축적이다. 언어서비스업체 플리토(Flitto)의 이정수(2019)는 인공지능의 능력을 향상시키기 위한 필수 요건이 빅데이터라고 하였다. 4차 산업혁명에서 새로운 혁신을 창출하는 것이 바로 데이터의 가치를 만들어내는 것(하원규와 최남희 2016: 53-58)이란 주장과 일맥상통한다. 현재 가장 많은 언어를 번역할 수 있으며, 기술을 선도하고 있는 구글 번역기의 가장 큰 무기는 구글 검색엔진을 통해서 확보할 수 있는 방대한 웹(Web) 문서들과 구글 번역기를 사용하는 사용자들이 제공하는 번역 결과물이라고 할 수 있다. 이것은 인공지능 번역기라고 하지만, 결국 인공지능 단독의 기술이 아닌, 빅데이터와의 결합에 의해서만 능력이 향상될 수 있다는 것을 의미한다.

신경망기반 기계번역이라고 하더라도 이전 개념을 무시하고 완전히 새롭게 탄생한 것이 아니라, 빅데이터를 학습할 수 있는 딥러닝(deep learning) 방법론을 도입하여 기계번역의 새로운 도약이 마련되었다(윤애선 2019). 그리고 이 부분이 바로 4차 산업혁명 시대에 인간과 기계의 접점이 될 수 있는 부분이다. 인공지능이 빅데이터로 활용할 수 있는 것은 결국 인간 통번역가가 생산한 고

품질의 번역 텍스트이기 때문이다. 인간은 번역 활동을 통해서 수많은 번역 텍스트를 만들어낸다. 그리고 연구자는 이를 활용하여 연구하고, 교육하고, 기계는 이 자료를 바탕으로 학습을 하고 인간의 번역 기술을 습득한다.

인간과 기계의 대립은 각자의 역할을 인정하지 않은 데 있다. 번역학과 번역공학 사이의 대립이 이러한 결과를 만들어내었다(이영훈 2018). 통번역학의 위기 혹은 통번역가의 위기가 아니라, 통번역학의 틀에서 인간과 기계의 협력을 모색해야 한다. 실제로 근래 몇 년 사이에 번역학 관련 학회에서 제4차 산업혁명과 번역에 대한 주제로 발표되는 연구 및 학술대회가 많이 늘어났다. 다양한 주장들이 있었지만, 핵심은 곧 인간과 기계가 어떻게 협력해서 이론과 실재의 조화를 이룰 수 있는지에 관한 주장이 대다수였다. 기존의 산업혁명에 의한 대량 실업과 빈부 격차와는 다른 문제이다. 실무 번역자도 단순히 주어진 텍스트를 정형화된 사전만 가지고 번역하는 시대는 이제 사라지고 있다. 굳이 4차 산업혁명을 언급하지 않더라도 세계화 속에서 급속도로 발전하는 환경은 번역가에게 새로운 문명의 이기에 관한 기술 습득을 요구한다. 철학, 인문학 번역은 보류한다고 하더라도, 실무 번역 시장은 새로운 지식의 확대와 탈지역화를 요구하고 빠른 정보처리를 요구한다. 그래서 번역가는 이에 신속하게 대처해야 한다. 지식의 확산 및 습득에 있어서 인터넷과 기계는 매우 중요한 수단이다. 그리고 그 지식은 이제 번역 대상이 되는 텍스트 자체에 관한 것뿐만 아니라 번역 활동에 있어서 필요한 도구들에 대한 지식도 마찬가지다(Fouad 15-16).

빅데이터는 그 이전보다 훨씬 중요해졌다. 그래서 코퍼스의 개념과 역할은 바뀌어야 한다. 전통적인 개념의 코퍼스는 제한된 크기와 텍스트 종류로 빅데이터의 개념에 적절하지 않다. 이러한 관점에서 웹 코퍼스는 전통적인 코퍼스를 보완할 수 있는 해결책이 될 수 있다.

## 4. 코퍼스의 새로운 역할

코퍼스를 기반으로 한 번역 연구를 말할 때, 일반적으로 떠오르는 개념은 기디언 투리(G. Toury)로 대표되는 기술번역학(Descriptive Translation Studies)의 규범(norm)과 모나 베이커(M. Baker)의 번역 보편소(translation universal)이

다. 사실 이 개념은 코퍼스의 관찰을 바탕으로 한다. 실제 번역 결과물을 살펴보고 이를 통해서 번역의 특징들을 찾아내고 규정하는 것이기 때문이다. 이후의 많은 연구자들이 현실적인 자료의 관찰을 통한 연구를 기술할 때 이들 두 연구자는 거의 빠짐없이 언급된다.

그래서 다양한 언어 텍스트를 바탕으로 구축되는 코퍼스는 언어 연구뿐만 아니라 번역 연구에도 중요한 원천 자료가 된다. 언어 집합체(collection of language)라는 초기의 코퍼스 정의는 시간이 지날수록 기술적인 발전과 연구의 전문화로 인해서 보다 전문화되었고, 이와 더불어 다양한 형태의 코퍼스가 등장하게 되었다. 단순한 언어 집합체를 넘어서 통계와 전산 개념이 추가되어 단순한 일반 텍스트 형태의 코퍼스가 아닌, 여러 가지 통계적 지표가 포함된 데이터베이스 형태의 코퍼스가 연구에 사용된다. 코퍼스 형태의 진화는 기술적인 발전에 힘입은 바가 크다. 컴퓨터 능력의 획기적인 발달로 우리는 매우 많은 자료를 이전보다 훨씬 빠르고 효율적으로 처리할 수 있게 되었으며, 인간 연구자는 연구 목적에 맞는 결과물을 체계적으로 관찰할 수 있게 되었다.

그렇지만 아무리 현대적 개념의 전문 코퍼스라고 하더라도 그리고 코퍼스의 형태가 어떤 것이라고 하더라도 관찰 대상은 언어이며[8], 이 언어가 바로 연구의 주제이다. 번역학의 관점에서는 번역어라고 할 수 있다. 일반적으로 코퍼스는 '용례(sample)', '대표성(representation)' 그리고 '실제 사용(usage)'으로 정의되는데, 그 중에서 '실제 사용'이 가장 핵심적인 개념이라고 할 수 있다. 대시와 아룰모자이(Dash & Arulmozi 2018)가 정리한 코퍼스에 관한 기존의 다양한 정의에서 반복되어 나타나는 표현 중의 하나가 바로 실제 언어(real-life language)이다. 기존의 전통적인 연구에서 볼 수 있는 연구자에 의해 만들어진 사례가 아니라, 실제 언어 사용자에 의해 산출된 살아있는 언어를 포함하고 있는 것이 바로 코퍼스이다. 이를 바탕으로 한 연구가 코퍼스 연구이다(McEnery & Wilson 2001: 1). 여기에는 규범적인 것뿐만 아니라 비규범적인 것도 함께 존재한다. 코퍼스에서 비규범적인 언어 사용이 많이 보인다면, 신뢰성(fidelity)의 문제가 나타날 수 있지만, 다수의 언어 화자가 범용적으로 용인하는 사용이

8) 단일어 연구에서는 한 언어의 보편적인 특징을 관찰할 수 있어야 하며, 번역학에서 다양한 번역 현상을 관찰할 수 있는 도구가 되어야 한다.

라면, 이것은 언어 연구뿐만 아니라 언어 교육에 있어서도 유의미한 자료가 될 수 있다. 코퍼스 연구가 이론을 입증하기 위해 잘 가공된 자료를 제시하는 것이 아니라, 실제를 바탕으로 이론을 검증하는 것이기 때문에, 코퍼스는 다양한 언어의 모습을 포함하고 있어야 한다. 다시 말해서 하향식(Top-down) 연구가 아닌, 상향식(Bottom-up) 연구의 특성을 가진다(Loock 2016). 연구의 목적이 어떤 것이든, 연구 방법이 어떤 것이든, 코퍼스 연구는 사례를 관찰하는 것에서 출발하기 때문이다. 다시 말해서, 어떤 특정 원리가 정해진 것이 아니라, 사례 관찰을 통해서 보편적 규칙성을 규명하는 것이다. 물론 원론적인 관점에서 보면 기존의 이론 연구도 이와 같은 방식을 가진다고 볼 수 있다. 그렇지만, 기존의 연구는 가설로 설정된 개념을 설명하기 위해서 적절한 사례를 '만들어내는' 것이라면, 코퍼스 연구는 규범이든 아니든 언어의 '실제 사용'을 바탕으로 가설을 정립해 나간다는 점에서 차별화된다고 할 수 있다.

이러한 코퍼스 기반 연구의 특성은 번역 연구와도 부합된다. 번역은 그 자체로 존재하는 것이 아니라, 번역가에 의해 실행된 번역 행위의 결과물이기 때문이다. 따라서 번역 자체를 연구하는 번역학은 다분히 경험적 연구 태도를 보일 수밖에 없다. 물론 초기의 번역 연구는 번역 자체의 개념에 대한 논의가 주를 이루었다. 그렇지만 이후 원문과 번역문 사이의 여러 수준에서의 등가 및 대응 관계를 살펴보는 연구가 점차 확산되었다(Baker 1993: 235-236). 번역학 연구 초기의 유명 저서인 『프랑스어-영어의 비교문체론(*Stylistique comparée du français et de l'anglais*)』도 비네와 다르벨네(Vinay & Darbelnet)의 수년에 걸친 번역 텍스트 분석의 결과물이라고 할 수 있다. 이론적 연구가 아닌 이상, 언어 현상의 일반화는 수많은 사례의 관찰에서 출발한다. 베이커(1993, 1995)와 라비오사(Laviosa 2002)가 코퍼스 언어학의 방법론을 번역 연구에 응용하기를 주장한 이유도 코퍼스 언어학의 방법이 다른 언어 연구 방법보다 뛰어났기 때문이 아니라, 코퍼스 연구의 특징이 번역 연구에 적합했기 때문이라고 할 수 있다. 다시 말해서, 번역어의 보편적 특징을 이해하기 위해서는 번역 텍스트에 대한 광범위한 조사가 필요할 것이다. 번역어가 가지는 다양한 특징들을 이해하기 위해서 투리는 대규모 코퍼스가 필요하다고 주장했지만, 대규모 코퍼스의 활용 방법에 대해서는 별다른 언급이 없었다. 기술번역학 관점에서 이루어진 많은 연구들(Laviosa 2002)은 개별 언어 간의 비교 연구와 다름없으며, 이 또한

넓은 의미에서 개인이 구축한 DIY 코퍼스(Do It Yourself corpus) 분석을 토대로 한 것이다.

컴퓨터 능력의 향상과 인터넷의 발달은 언어 및 번역 연구를 위한 전자문서의 양적인 증가를 이루었다. 인터넷 상에서 만들어지는 수많은 웹 문서는 우리가 언제든 그리고 거의 무조건적으로 접근할 수 있는 언어 텍스트다. 실제 많은 연구자들이 웹 문서를 코퍼스의 원천으로 활용하고 있다. 이와 더불어 개인 블로그(blog), 트위터(Twitter) 및 페이스북(Facebook)과 같은 소셜 네트워크 서비스(Social Network Services, SNS)는 규범적인 언어 양태를 넘어서 비규범적인 언어 양태를 관찰할 수 있는 코퍼스가 될 수 있다. 이처럼 다양한 언어의 모습을 담고 있기 때문에, 웹 문서는 전통적인 언어 연구뿐만 아니라, 한 언어 공동체의 통시적 언어 연구, 사회적·문화적 변화에 따른 언어 습관, 비표준 언어 연구 등 다양한 응용 연구를 위한 참조 자료가 되기도 한다. 번역의 관점에서도 웹은 새롭게 만들어지는 신조어의 번역어를 찾고, 계층과 문화적 범주에 따른 번역의 다양성을 연구하고, 교육에 활용할 수 있는 바탕이 될 수 있다.

물론, 학술적인 관점에서 코퍼스는 일정한 기준에 따라서 체계적이고 신중하게 선별된 텍스트를 활용해야 하지만, 이러한 개념이 오히려 개인 연구자들이 코퍼스 기반 연구를 하는 데 있어서 상당한 제약이 될 수 있다. 예를 들어, BNC(British National Corpus)[9]나 COCA(Corpus Of Contemporary American English)[10]와 같은 참조 코퍼스(reference corpus)는 개인 연구자들이 아니라, 협업 작업의 결과이다. 국내의 21세기 세종말뭉치[11]도 국립국어원의 재정 지원으로 1998년부터 2007년까지 여러 연구자가 작업한 결과물이다.[12]

기계번역의 질적 향상과 인간과의 협업적 연구를 위해서 코퍼스는 빅데이터로서의 면모를 갖추기 위해서 크기뿐만 아니라 포함된 텍스트의 다양성도 요구된다. 그리고 현실 언어의 모습을 반영하기 위해서는 데이터의 주기적인 갱

9) British National Corpus(http://www.natcorp.ox.ac.uk/)
10) Corpus Of Contemporary American English(https://www.english-corpora.org/coca/)
11) 국립국어원 말뭉치(https://ithub.korean.go.kr/user/guide/corpus/guide1.do)
12) 불행히도 세종 말뭉치 이후 새로운 대규모 코퍼스는 구축되고 있지 않으며, 세종 말뭉치조차도 2007년 이후 재정 지원 중단으로 더 이상 갱신이 되고 있지 않다(이성규 2016).

신도 필요하다. 개인이 구축하는 DIY 코퍼스는 이러한 측면에서 한계가 있다. 소규모일 수밖에 없는 개인 코퍼스는 언어의 다양한 모습을 담고 있지 못하기 때문에, 다양한 번역 등가를 찾는 데 어려움이 있다. 그리고 특정 장르의 텍스트로 구축되는 경향이 있기 때문에, 번역의 보편적 연구에도 적합하지 않다. 새로운 환경의 언어 및 번역 연구를 위해서는 코퍼스도 기존의 틀에서 탈피해서 새로운 변화를 받아들여야 할 것이다. '언어 용례', '선별된 텍스트', '전자 문서', '대표성'으로 정의되는 코퍼스는 이제 '다양성', '보편성', '접근성', '현실성'의 개념을 포함하는 보다 넓은 의미의 코퍼스가 되어야 할 것이다.

## 5. 빅데이터로서의 코퍼스

### 5.1. 웹 코퍼스

코퍼스 연구에서 가장 중요한 것은 원천 자료로서의 코퍼스를 어떻게 구축하는가이다. 신뢰할 수 있는 코퍼스를 구축하기 위해서는 '다양한 장르'의 텍스트(다양성)와 '많은' 텍스트를 포함할 필요가 있다.[13] 더구나 기술번역학에서의 규범과 베이커의 번역 보편성을 설명하기 위해서는 특정 장르에 국한되지 않은 대규모의 범용 코퍼스(보편성)가 요구된다. 또한 언어는 항상 변하고 시대에 따라서 새로운 신조어가 등장하고 기존의 표현 방식이 새로운 표현으로 대체되기도 한다. 이러한 언어 사용의 변화를 제대로 반영하기 위해서는 코퍼스의 유지·보수(현실성)도 중요한 부분이라고 할 수 있다. 앞서 참조 코퍼스 사례로 든 BNC와 COCA도 최신 언어의 모습을 반영하기 위해서 갱신 작업이 지속적으로 이루어진다(Loock 2016: 30). 그렇지만 개인 연구자의 DIY 코퍼스는 이러한 작업이 쉽지가 않을 것이다. 더욱이 BNC나 COCA 같은 대규모 코퍼스의 보수 및 유지를 위해서는 막대한 비용이 든다. 앞서 언급한 것처럼 세종코퍼스가 더 이상 유지되지 못하고 있는 이유이다.

13) 일반적으로 참조 코퍼스의 대표적인 예로 인용되는 BNC 코퍼스 혹은 COCA 코퍼스 같은 경우는 포함하고 있는 텍스트 종류뿐만 아니라, 크기에 있어서도 언어 연구에 충분한 신뢰를 보장할 수 있다.

이러한 점에서 인터넷에 존재하는 웹 사이트는 코퍼스의 원천 자료로서 가치가 있는 도구가 될 수 있다. 웹 코퍼스(영어로 Web corpus 혹은 Web as corpus)라고 일컫는 코퍼스는 최근에 많은 관심을 받는 코퍼스의 한 형태라고 할 수 있다. 사실 웹 코퍼스를 완전히 새로운 형태의 코퍼스라고 말하기는 어렵다. 웹 코퍼스도 엄밀한 의미에서, 기존의 정의인 단일어 코퍼스, 다국어 코퍼스, 비교 코퍼스, 병렬 코퍼스에서 벗어난 형태가 아니기 때문이다. 단지 폐쇄된 오프라인의 물리적 전자 저장장치에 있지 않을 뿐이다. 반대로 말하면 공개되어 누구나 접근 가능한 코퍼스이다(접근성).

랭커스터 대학(Lancaster University) 저널에 발표한 킬가리프(Kilgariff 2010)의 짧은 논문에서 Web as corpus(웹 코퍼스)라는 표현이 등장한 이후, 이 명칭은 일종의 고유명사처럼 사용되고 있다. 그렇지만 언어 연구에서 웹이 활용되기 시작한 것은 1990년대부터이다(Kilgariff 2010: 343). 웹 코퍼스의 바탕은 구글(Google)[14]이나 빙(Bing)[15]과 같은 검색엔진이다. 기존의 코퍼스 연구와 마찬가지로 이 검색엔진을 사용하여 키워드 검색을 한 후 나온 결과물이 곧 웹 코퍼스가 된다. 키워드 검색의 결과물은 WordSmith[16]나 AntConc[17]와 같은 코퍼스 전문분석 도구와 마찬가지로 어휘색인목록(Concordance)화 된다. 특정 코퍼스가 아닌, 접근 가능한 웹 문서 전체를 대상으로 한다는 점에서 기존의 코퍼스와 차이가 있다.

규모 면에서도 웹 코퍼스는 기존의 코퍼스와 차이가 있다. BNC가 영국식 문어 및 구어 영어 1억 어절을 포함하고 있으며, COCA는 미국식 문어 및 구어 영어 5억 3천만 어절을 포함하고 있다. 이들 코퍼스가 특정 언어, 다시 말해서 영국 영어와 미국 영어를 대상으로 한 코퍼스라면 웹 코퍼스는 이러한 한계를 넘어선다. 로렌스와 질(Lawrence & Gile 1999: 107, Meyer *et al*., 2016: 243 재인용)에 따르면, 거의 8십만 웹 페이지에 해당하는 코퍼스를 분석한 결과 포함된 어절 수가 8천억 이상이었다. 그리고 언어의 다양성도 웹 코퍼스의 특징이다. 검색엔진이 특정 언어로 작성된 문서를 검색 대상으로 하지 않기 때

14) Google(https://www.google.com/)
15) Bing(https://www.bing.com/)
16) WordSmith Tools(https://www.lexically.net/wordsmith/)
17) AntConc(https://www.laurenceanthony.net/software/antconc/)

문에 해당 키워드가 포함된 다국어 코퍼스를 생성할 수도 있다. 그렇지만, 웹 코퍼스도 특정 언어 그리고 특정 장르에 편중되어 있다고 한다. 플레젠트(Pleasants 2001, Meyer *et al*. 2016: 243 재인용)에 따르면 영어로 작성된 웹 문서가 반 이상(68.4%)을 차지하고 나머지 언어들이 3-4% 내외로 분포한다. 장르에 있어서도 대다수의 웹 문서가 상업(83%)과 관련된다(Lawrence & Gile 1999, Meyer *et al*. 2016: 244 재인용).

자네틴(Zanettin 2014a, 2014b)에 따르면, 웹 코퍼스는 언어 및 번역 연구를 위해 학술적 관점에서 구축된 것이 아니기 때문에, 체계적이지 못하고 코퍼스의 정량적 분석도 상당히 모호할 수 있다. 그렇지만, WebCorp Live[18], WebAsCorpus[19]와 같이 AntConc와 같은 전문 도구 수준에는 미치지 못하지만, 일반 검색엔진보다는 훨씬 정교한 분석이 가능한 도구를 활용할 수 있기 때문에(Zanettin 2014b: 57-68), 기존의 코퍼스를 보완할 수 있는 도구가 될 수 있다. 또한 요한슨과 게바라(Johannessen & Guevara 2011)에 따르면, 웹 코퍼스는 문어뿐만 아니라 구어도 동시에 포함하고 있기 때문에, 주로 문어 텍스트를 대상으로 한 기본의 코퍼스보다 더 많은 이점이 있으며, 언어 교육을 위해 언어 패턴을 살피는 데에도 장점이 있다(Kvashnina & Sumtsova 2018).

### 5.2. 웹코퍼스 분석 사례

#### 5.2.1. habit/습관

키워드 검색을 이용하여 기존의 코퍼스와 동일한 어휘목록색인을 웹 코퍼스를 대상으로도 할 수 있다. 웹 코퍼스를 기반으로 habit과 한국어 '습관'을 검색하면 다음과 같은 어휘목록색인을 얻을 수 있다. Webcorp Live 도구를 이용한 키워드 검색은 웹 페이지에서 habit과 '습관'을 포함한 페이지들을 사이트별로 보여준다. 문제는 이러한 결과물이 번역 대응 관계가 아니기 때문에, habit과 '습관' 사이의 번역 대응을 살펴보기는 어렵지만, 최소한 이들 어휘를 중심으로 한 영어와 한국어의 어휘 결합 형태를 확인할 수 있다.[20]

---

18) WebCorp Live(http://www.webcorp.org.uk/live/)
19) WebAsCorpus(https://kwicfinder.com/searchwac.html)
20) Webcorp Live와 같은 웹 코퍼스 도구를 구체적으로 소개하는 것이 옳지만, 본 논문

〈그림 1〉 Webcorp Live에 의한 *habit* 어휘색인목록


〈그림 2〉 Webcorp Live에 의한 '습관' 어휘색인목록


일종의 비교 코퍼스를 활용한 연구와 비슷하다고 할 수 있는데, 사례를 면밀하게 살펴보면, 'habit formation/습관형성', 'bad habit/나쁜 습관'처럼 서로 번

---

은 빅데이터로서의 코퍼스에 대한 개념적 논의를 목적으로 하기 때문에, 도구에 대한 설명은 생략하고자 한다. 이 도구에 대한 자세한 내용은 자네틴(2014b) 참조.

역 대응이 있을 수 있는 표현들을 찾을 수 있다. 여기서는 사례를 간단히 보였지만, 검색 결과 전체를 살펴보면 훨씬 많은 대응 관계를 발견할 수 있을 것이며, 해당 표현이 포함된 맥락도 함께 비교할 수 있을 것이다. 개인 코퍼스는 크기가 작아서, habit/습관이 포함되어 있지 않다면, 이 어휘의 번역 관찰이 불가능하지만, 웹 문서는 이러한 문제점은 없다고 할 수 있다.

5.2.2. window/windows

좀 더 명확한 비교를 위해서는 웹에서 번역 병렬코퍼스로 구축된 도구를 사용하는 것도 하나의 방법이다. 예를 들어, Linguee[21] 혹은 Reverso[22] 같은 사이트는 일종의 인터넷 병렬코퍼스 사이트라고 할 수 있다. 검색하고자 하는 키워드를 중심으로 번역 대응되는 맥락을 병치하여 보여준다. 문제는 한국어가 빠져있기 때문에, 한국어와 다른 외국어 사이의 직접적인 번역 대응어 관찰은 불가능하다. 그렇지만, 영어-프랑스어와 같은 다른 외국어 사이의 검색은 가능하기 때문에, 간접적인 방식으로의 연구는 가능하다.

예를 들어, Linguee에서 영어 window의 프랑스어 역을 비교하기 위해서 window를 키워드로 검색하면 다음과 같은 결과물을 얻을 수 있다.[23]

(1) a. We therefore offer a complete range of products which allow you to clean a wide variety of **window** surfaces quickly and without streaking.
b. C’est pourquoi nous proposons un assortiment complet permettant de nettoyer rapidement et sans traces les **vitres** les plus diverses.
(2) a. The definition and input of this number through a “single **window**” into the logistic and supply chain should happen only once.
b. La définition et la saisie de ce numéro par un “**guichet** unique” dans la chaîne logistique et d’approvisionnement ne doivent avoir lieu qu’une seule fois.
(3) a. Closes the **window** of the selected circuit or the selected parts list.

21) Linguee(https://www.linguee.com/)
22) Reverso(https://context.reverso.net/translation/)
23) Linguee는 훨씬 많은 사례를 제시하지만, 번역의 다양성을 보여주기 위해 서로 다른 의미를 가지는 몇 사례만 인용하였다.

b. Ferme les **fenêtres** du circuit sélectionné ou de la liste des pièces sélectionnée.

위 사례에서 영어 window는 프랑스어에서 fenêtre를 비롯해서 vitre, guichet라는 다양한 프랑스어로 번역되고 있음을 확인할 수 있다. 예문 (1)에서는 '창문'의 의미로 예문 (2)에서는 '창구'로, 그리고 마지막으로 예문 (3)에서는 정보학의 개념으로서의 '창'을 의미한다. 이를 통해서 우리는 영어 window와 프랑스어 vitre/guichet/fenêtre의 의미 범주가 서로 다르다는 것을 확인할 수 있으며,[24] 만일 번역에서 영어와 프랑스어 어휘가 동일한 의미 범주를 가진다고 생각하면, 번역 오류가 생길 수 있다.

복수형 windows는 이와는 조금 다른 양상을 보인다.

(4) a. Last week, I was in one of those areas and was told about the black soot that lands on people's washing lines and **windows** and affects people's health.
b. [...] j'ai entendu parler de la suie noire qui se dépose sur les cordes à linge et les **fenêtres** et affecte la santé des citoyens.

(5) a. [...] energy efficiency perspective so they could change their **windows** and doors, put more insulation in or buy more efficient furnaces.
b. [...] l'efficacité énergétique, par exemple en remplaçant les portes et les **fenêtres**, en isolant davantage ou en achetant des appareils [...]

(6) a. If you have a house with no doors, people will come in through the **windows**.
b. Si notre maison n'a pas de porte, les gens entreront par la **fenêtre**.

(7) a. The new façade will be composed of a patchwork of traditional wood-**frame windows** from different European countries.
b. La nouvelle façade sera composée d'un patchwork de vieux **châssis** de bois, en provenance de différents pays européens.

24) 영어는 winodw라는 하나의 어휘로 '창문', '창구', '창(전산학)'을 모두 표현할 수 있지만, 프랑스어에서는 맥락에 따라서 서로 다른 어휘를 사용해야 한다는 것을 알 수 있다.

(8) a. With the arrow navigation you can fade in all contents **windows** one after another and thus click through the entire content.
b. Les touches de direction vous permettent de sélectionner les **fenêtres** l'une après l'autre et afficher ainsi tous les contenus.
(9) a. As **Windows** starts, you will see it find your new hardware and software.
b. Quand **Windows** se mettra en route, il retrouvera automatiquement votre nouveau périphérique et votre logiciel.
(10) a. In **Windows**, open the Fast Tranck Ultra control panel by double-clicking on the M-Audio icon in the system tray, or from Start [...]
b. Sous **Windows**, ouvrez le panneau de configuration de la Fast Track Ultra en double-cliquant sur l'icône M-Audio de la barre [...]

위 사례를 보면 영어 windows도 단수형처럼 여러 프랑스어 어휘(fenêtres, châssis, Windows)로 번역되고 있음을 확인할 수 있다. 그런데 의미는 조금 다르다. 예를 들어, windows가 일반적인 창 혹은 창틀을 가리킬 때, 프랑스어는 fenêtres와 châssis를 번역어로 선택하고 있다. 그런데 (8b)를 보면 프랑스어가 fenêtres로 번역되어 있지만, 이것은 건물의 창이 아니라 컴퓨터 상의 프로그램 창을 가리킨다. 그런데 마지막 예(9b, 10b)를 보면 대문자로 쓰인 Windows가 나타나는데, 이것은 운영체제인 마이크로소프트사(Microsoft)의 윈도우를 가리킨다. 이것으로 우리는 영어는 대문자와 소문자로 구별하여 윈도우와 창을 구별하지만, 프랑스어는 창을 가리킬 때는 fenêtre를 사용하지만, 윈도우를 가리킬 때는 영어를 차용하고 있음을 알 수 있다.[25] 윈도우 자체가 일종의 고유명사이기 때문에, 영어를 그대로 사용하고 있다.[26]

---

25) 프랑스어와 달리 한국어는 이런 맥락에서 '창' 혹은 '윈도우'를 사용한다. 그리고 운영체제 자체를 의미할 때는 '창'이 아닌 '윈도우'만을 사용한다. 반면 프랑스어는 '창'을 가리킬 때, fenêtre만을 사용한다는 것을 확인할 수 있다.

26) 유사한 사례로 운영체제를 뜻하는 영어 Operating system은 프랑스어로 Système d'exploitation으로 번역된다. 그런데 약어로 표기할 때 영어가 OS라면, 프랑스어는 SE가 되어야 하는데, 실제로 코퍼스를 검색해보면 영어를 그대로 사용하여 OS라고 한다(Loock 2016: 135-138). 그런데 코퍼스가 컴퓨터 관련 텍스트를 포함하지 않는다면, 이러한 조사가 쉽지 않을 수 있다.

5.2.3. petit prince/어린 왕자

〈표 1〉『어린 왕자』의 고빈도어 목록[27]

| 원문 | 민희식 | 이진구 | 정소성 | 김화영 |
|---|---|---|---|---|
| le | 어린(189) | 어린(172) | 어린(208) | 어린(225) |
| de | 말했다 | 말했다 | 나는 | 나는 |
| il | 나는 | 그 | 말했다 | 그 |
| je | 그 | 나는 | 그 | 말했다 |
| et | 그는 | 그는 | 왕자는(106) | 왕자는(100) |
| un | 왕자는(87) | 왕자가(87) | 그는 | 왕자가(92) |
| est | 왕자가(77) | 내 | 내 | 한 |
| les | 내 | 난 | 거야 | 거야 |
| la | 난 | 내가 | 이 | 있는 |
| Petit(203) | 거야 | 그가 | 이렇게 | 그는 |
| à | 내가 | 한 | 그러나 | 내 |
| pas | 그러나 | 그러나 | 왕자가(60) | 이 |
| l | 이 | 거야 | 있는 | 내가 |
| ne | 수 | 왕자는(53) | 내가 | 하고 |
| Prince(173) | 한 | 게 | 수 | 것이다 |

<표 1>은 프랑스 소설『어린왕자(Petit Prince)』에서 빈도가 높은 순서로 단어를 나열한 것이다. 소설 제목이 '어린 왕자'라서, 가장 빈도가 높은 단어가 기능어(관사, 전치사, 접속사 등)를 제외한 실사로서의 단어는 소설의 주인공을 가리키는 petit와 prince가 되는 것은 당연할 것이다. 원문에 대한 4종류의 한국어 번역문에서도 '어린'과 '왕자'가 가장 빈도가 높은 어휘들이다.[28] 프랑스어 단어 prince는 맥락에 따라서 '군주' 혹은 '왕자'로 번역될 수 있지만, 생텍쥐페리의 소설에서는 고유명사처럼 사용되는 '어린왕자'로 번역된다. 그리고 petit는 좀 더 다양한 의미로 번역이 될 수 있지만, '작다' 혹은 '어린'이 일반적으로 많이 쓰이는 의미라고 할 수 있다. 그런데 소설에서가 아닌, 일반적인 맥락에서 petit prince가 '어린 왕자'가 아닌 '작은 왕자'로는 번역될 수 없느냐는 의문이 생길 수 있다. Petit garçon이 '아이'에 해당한다면, petit가 모두 '나이가 어리다'를 의미한다고 생각할 수도 있다. 그렇다면 이때 이 프랑스어 표현이 '(키 혹은

27) 표 내의 (숫자)는 해당 어휘의 빈도를 가리킨다.

28) 한국어 번역 텍스트는 형태소 분석을 실행하지 않은 원시 텍스트를 사용했기 때문에, 명사+조사의 결합 형태로 prince의 번역어가 분산되어 나타나고 있다. 따라서 prince와 '왕자'의 빈도수가 다르다.

몸이) 작은 아이'로는 해석될 여지는 없는가?

인터넷 병렬코퍼스인 Linguee에서 프랑스어 petit garcon에 해당하는 영어 표현을 찾아보면, 다음과 같은 대응 관계를 확인할 수 있다.

(11) a. Puis le **petit garçon**, il pleurait, pleurait vraiment beaucoup.
b. But the **little boy** cried, he really cried so much.
(12) a. Au cours des recherches, la CBSARA a enrichi ses connaissances sur l'autisme et en a appris davantage sur ce **petit garçon**.
b. In the course of the search, we found out more about autism and this particular **young boy**.
(13) a. Voilà qu'un jour, de ce baobab naît un oeuf et, de cet oeuf, un **petit garçon**.
b. One day the baobab lays an egg and out of the egg comes a **small boy**.
(14) a. À la suite du divorce de ses parents le **petit garçon** vécut en Louisiane avec sa mère.
b. Following the dissolution of the parents' marriage the **boy** lived with the mother in Louisiana.

사례 대부분이 (11b)처럼 little boy로 번역되었지만, 드문 사례로 young boy(12b), small boy(13b), 심지어 boy(14b)로만 번역된 경우도 발견된다. 그런데 구글에서 다음과 같은 한국어 사례를 찾아볼 수 있다.

> 생텍쥐페리의 책에 나오는 **작은 왕자**(Little Prince)가 자신이 기른 작은 장미 한 송이와 여행 도중에 사귄 여우는 서로가 길들었기 때문에 소중한 것이라고 대답했던 생각이 났다.
> 어렸을 때는 **작은 왕자**가 사막에서 겪는 일과 그의 여행사건에 초점을 맞춰서 읽었던 기억이 있는데, 지금은 **어린왕자**의 표현 하나하나가 [...]

위 사례를 보면 little이 '작은'으로 표현이 되어 있다. 이 경우에 우리는 '작은'이라는 표현을 어떻게 해석할 수 있을 것인가? '몸이 작다'라는 말인가? '나이가 어리다'는 말인가? 첫 번째 사례에서는 '작은 장미'라는 표현도 확인할 수 있는데, 한국어에서 '작다'라는 표현은 '나이'보다는 '크기'와 관련된다고 할 수

있다.[29] 왕자를 중심으로 구글 검색으로 '어린', '작은', '꼬마'와의 결합 형태를 찾아보면 다음과 같은 결과를 확인할 수 있다.

어린 왕자 : 407,000
작은 왕자 : 18,500
꼬마 왕자 : 12,300

구체적인 출현 맥락을 살펴보면, 어린 왕자는 생텍쥐페리 소설 관련 텍스트였으며, 작은 왕자는 부분적인 연관성(212개)을 보였다. 반면 꼬마 왕자는 소설과 관련 없는 맥락에서 출현하였다. 이러한 분포 상황을 고려하면 어린 왕자는 생텍쥐페리 소설에서만 나타나는 고유명사화된 표현이라고 생각할 수 있다.

5.2.4. 거짓말/mensonge/lie

앞서 살펴본 것처럼 번역어는 맥락이나 언어에 따라서 여러 대응어가 있을 수 있다. 우리가 일상에서 좋은 의도로 하는 거짓말을 '선의의 거짓말'이라고 한다. 이에 대한 프랑스어는 mensonge officieux와 pieux mensonge 두 가지 표현을 생각할 수 있다. 그리고 영어는 white lie가 사용된다. Linguee나 Reverso에서 white lie에 대응되는 프랑스어 표현을 찾아보면 앞서 언급한 두 프랑스어 표현이 나온다. 그런데 이중 어떤 표현을 더 많이 사용하는지 확인하기 위해서 구글 검색을 활용할 수 있다. 그 결과는 다음과 같다.

Mensonge officieux - 139,000
Pieux mensonge - 29,300

단순히 빈도가 모든 것을 말하는 것은 아니지만, 어쨌든 빈도가 높다는 것은 그만큼 프랑스어 화자들이 많이 선호한다는 것을 의미한다. 절대적 지표가 될 수는 없다고 하더라도 충분히 참조할 수 있는 수치라고 할 수 있다. 더욱이 위의 빈도수에서 알 수 있듯이 mensonge officieux가 pieux mensonge보다는 십

29) 한국어는 '나이'-'어리다'/'적다', '길이, 넓이, 부피'-'작다'와 같은 호응관계를 가진다면, 프랑스어 petit와 영어 little은 두 경우 모두 가능하다.

만 이상의 빈도 차이를 보인다. 따라서 후자보다는 전자가 한국어 '선의의 거짓말'로 선택하는 것이 타당할 수 있다.[30]

그런데 여기서 '선의의 거짓말'과 대응되는 표현으로서 '좋은 거짓말'을 고려할 수 있고, 이에 대해서 프랑스어 'bon mensonge'와 영어 'good lie'를 번역어로 생각할 수 있다. 이러한 표현들이 실제로 각 언어 화자들이 많이 사용한다면 이 표현들도 '선의의 거짓말'에 대한 번역어로 고려할 수 있을 것이다.

같은 방식으로 구글 검색을 하면 다음과 같은 빈도 결과를 얻을 수 있다.

선의의 거짓말 - 58,500
좋은 거짓말 - 10,600

White lie - 2,850,000
Good lie - 680,000

Bon mensonge - 7,670

세 언어 모두에서 좋은/good/bon과 거짓말/lie/mensonge와의 결합은 낮은 빈도를 보였다. 특이한 점은 다른 두 언어와 달리 영어의 good+lie 조합은 상대적으로 비교적 높은 빈도를 보였는데, 구체적인 맥락을 살펴본 결과 이 조합은 한국어의 '선의의 거짓말'의 맥락이 아니라, 영화 제목이었다.[31] 따라서 매우 특이한 사용으로 볼 수 있다. 어쨌든 이러한 검색 결과 좋은/good/bon과 거짓말/lie/mensonge의 결합은 의미적으로는 수용할 수 있지만, 사용에서는 세 언어 모두에서 낮은 빈도를 보이며, 따라서 이러한 유형의 결합은 관용적으로 사용이 되지 않는다고 생각할 수 있다. 외국어 학습 및 번역에서 가장 어려운 점이 바로 이런 부분이라고 할 수 있다. 문법적으로도 의미적으로도 전혀 문제될 것이 없지만, 해당 언어권의 화자들이 이를 사용하지 않는다면 그 어휘 혹은 표현은 지양해야 할 것이다.

---

30) 이 결과는 officieux가 공식적이고 pieux는 비규범적이라는 것을 가리키는 것이 아니라, officieux가 실제 사용(usage)에서 더 선호된다는 것을 의미한다.
31) 특이한 점은 good lie의 한국어 제목은 '뷰티풀 라이'로 되어 있다.

5.2.5. 인사표현

처음 프랑스어에 입문할 때, 프랑스어 인사법으로 안부를 묻는 표현에서 다음과 같은 표현을 배운다.

comment allez-vous? - 1,910,000
comment ça va? - 12,400,000
ça va? - 50,500,000

영어의 how do you do?/how are you?에 해당하는 이 표현들은 존대법[32]의 차이일 뿐 의미는 동일하다. 그런데 프랑스어 입문자 혹은 영어 입문자도 마찬가지겠지만, 학습자는 교과서에 나와 있는 표현을 그대로 사용하는 경향이 있기 때문에, 실제 언어 사용과는 괴리감이 있을 수 있다. 위 예에서 볼 수 있듯이 comment allez-vous?와 나머지 두 표현 사이에는 매우 큰 빈도 격차가 있다. 물론 맥락에 따라서 매우 공식적인 표현인 comment allez-vous?를 사용해야겠지만, 대화 상황이 매우 제한적이고 나머지 두 표현이 더 일반적이라는 것을 간접적으로 확인할 수 있다. 이러한 관점에서 교육적인 차원에서 웹 코퍼스는 많은 장점이 있다고 할 수 있다.

반면에 영어는 조금 다른 모습을 보이는데, 구글 검색을 하면 다음과 같은 빈도 양상을 보인다.

how do you do? - 282,000,000
how are you? - 161,000,000

프랑스어와 달리 how do you do?가 how are you?보다는 훨씬 많은 빈도를 보이는데, 물론 웹 사이트에서 이들 표현의 구체적인 맥락을 면밀하게 분석을 해야 하겠지만, 영어보다 프랑스어의 공식적인 표현의 사용이 훨씬 제한적이라고 생각할 수 있다.

32) 프랑스어에서 존대법은 한국어의 그것과 다른 의미이며, 상하 관계라기보다는 친밀도에 가깝다고 할 수 있다.

## 5. 결론

지금까지 우리는 4차 산업혁명이라는 새로운 환경에서 빅데이터로서의 코퍼스, 정확히 말하면 웹 코퍼스가 번역학에서 어떻게 사용될 수 있는지를 살펴보았다. 인공지능과 기계 번역의 영향력이 점차 확장되는 분위기에서 인간 번역가는 분명 기존의 것과는 다른 시각을 가져야 할 것이다. 코퍼스 기반 연구도 마찬가지라고 할 수 있다. 베이커가 번역 연구에서 코퍼스의 활용을 주장한 이래로 많은 시간이 흐른 것은 아니지만, 그 짧은 시간 동안 연구 환경은 급속하게 변했다. 그리고 이러한 환경적 변화는 코퍼스의 범주를 더욱 확장시켰다.

서양과 달리, 국내 코퍼스 연구 환경은 사실상 열악하다고 할 수 있다. 연구자 개인의 역량이 부족하다는 것이 아니다. 국내 연구자도 다양한 주제로 코퍼스를 기반으로 한 연구 결과물을 내놓고 있기 때문이다. 문제는 근본적인 문제, 즉 코퍼스 자체에 있다. 언어별로 다양한 형태의 코퍼스가 구축된 것도 아니며, 몇몇 연구소 혹은 연구원 차원에서 구축된 코퍼스를 제외하면, 사실상 국내 코퍼스 환경은 열악하다고 할 수 있다.

이러한 관점에서 웹 문서는 하나의 대안이 될 수 있다. 앞서 살펴본 것처럼 기존의 대규모 코퍼스는 영어를 중심으로 한 코퍼스이기 때문에, 한국어가 포함된 코퍼스를 찾기가 어렵고, 개인용 코퍼스는 크기 및 포함된 텍스트 종류의 한계 때문에 특정한 연구를 위해 구축되는 코퍼스일 수밖에 없다. 그래서 찾고자 하는 표현이 없을 수도 있고, 출현 빈도도 매우 낮을 수 있다.

이러한 관점에서 웹 코퍼스는 크기 및 텍스트의 종류에 제한이 없기 때문에, 전통적 코퍼스를 보완하는 대안이 될 수 있다. 물론 웹 코퍼스는 학술적 관점에서 구축된 코퍼스가 아니기 때문에, 엄밀한 언어 혹은 번역 분석을 위한 코퍼스가 되기에는 제약이 있을 수 있다. 그리고 인터넷 병렬코퍼스라고 할 수 있는 Linguee에서도 알 수 있듯이 한국어를 포함한 번역 웹 문서는 상대적으로 부족하기 때문에, 한국어를 중심으로 한 번역 연구는 한계가 있을 수밖에 없다.

그러나 기존의 코퍼스가 특정 기준에 맞추어 구축되어, 언어의 특정 사용 환경만을 반영할 가능성이 높다면, 웹 코퍼스는 일종의 언어 빅데이터로서 크기 및 텍스트의 종류에 제한을 두지 않기 때문에, 기존의 코퍼스에 비해 분석 대상이 되는 언어를 매우 다양한 사용 환경에서 관찰할 수 있는 기회를 제공할

수 있다. 이러한 관점에서 비교 코퍼스의 관점에서 웹 코퍼스는 전통적 코퍼스의 보완적 수단으로서 많은 사례를 관찰할 필요가 있는 번역 연구뿐만 아니라 번역 교육에도 도움이 될 수 있을 것이다.

끝으로 본 연구에서 웹 코퍼스를 활용하기 위한 다양한 도구들을 소개하였지만, 몇몇 사례만을 제시하고 보다 다양한 활용법을 제시하지는 못하였다. 이에 관해서는 향후 연구를 통해서 보다 구체적으로 소개하고자 한다.

## 참고문헌

김진욱 옮김 (2014)『제3의 물결』, 서울: 범우사. (Alvin Toffler, *The Third Wave*, Bantam Books, 1980).

남기춘 (2019) Translation and Brain,『한국번역학회 2019년 봄 학술대회 발표논문집』1-10.

송연석 (2018)「기계번역 담론에 대한 비판적 고찰」,『번역학연구』19(1): 119-145.

윤애선 (2019)「『디지털 바벨탑』 세우기, 어디까지 왔나? －프-한 기계번역의 현황과 전망」,『불어불문학연구』117: 157-199.

이영훈 (2018)「번역공학적 이성 비판: 디지털 시대 번역학의 과제」,『2018 한국외국어대학교 통번역연구소 국제학술대회』120-132.

이정수 (2019)「언어데이터의 중요성과 전문통번역가의 역할」,『한국번역학회 2019년 봄 학술대회 발표논문집』155-160.

이향 (2019)「번역학의 과제」,『한국번역학회 2019년 봄 학술대회 발표논문집』89-92.

장애리 (2017)「국내 기계 통번역의 발전 현황 분석 －한·중 언어 쌍을 중심으로」,『번역학연구』18(2): 171-206.

정영목 (2018)『완전한 번역에서 완전한 언어로』, 서울: 문학동네.

조영임 (2013)「빅데이터의 이해와 주요 이슈들」,『한국지역정보화학회지』16(3): 43-65.

조준형 (2012)「병렬코퍼스에서 맥락 탐색의 의미와 한계」,『번역학연구』13(5): 223-246.

조준형 (2019)「웹코퍼스 기반 번역 교육」,『한국번역학회 2019년 봄 학술대회 발표논문집』113-119.

최희섭 (2017)「인공지능 시대를 맞이하는 번역가의 자세」,『한국번역학회 2017 봄 학술대회 발표논문집』1-4.

하원규, 최남희 (2016)『제4차 산업혁명』, 서울: ㈜콘텐츠하다.

Baker, Mona (1993) 'Corpus Linguistics and Translation Studies: Implications and Applications', in *Text and Technology*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 233-250.

Baker, Mona (1995) 'Corpora in Translation Studies : An Overview and Some Suggestions for Future Research', *Target* 7(2): 223-243.

Ballard, Michel (1992) *De Cicéron à Benjamin. Traducteurs, traductions, réflexions*, Lille: Presses universitaires de Lille.

Cronin, Michael (2013) *Translation in the Digital Age*, New York: Routledge.

Dash, Niladri Sekhar & Arulmozi, Selvaraj (2018) *History, Features, and Typology of Language Corpora*, Singapore: Springer.

Fouad, Maali Tewfic (2011) *Enseignement de la traduction spécialisées par corpus bilingues alignés*, Saarebruck: Éditions universitaires européennes.

Guidère, Mathieu (2008) *Introduction à la traductologie*, Bruxelles: De Boeck.

Guiraud, Pierre (1960) *Problèmes et méthodes de la statistique linguistique*, Paris: Presses universitaires de France.

Johannessen, Janne Bondi & Guevara, Emiliano Raul (2011) 'What kind of corpus is a web corpus?', in Bolette Sandford Pedersen, Gunta Nešpore & Inguna Skadina (eds.), *NODALIDA 2011 Conference Proceedings*, 122-129.

Keromnes, Yvon (2016) 'La comparaison de traductions et de textes parallèles comme méthode heuristique en traductologie', in Jörn Albrecht & René Métrich (éds.), *Manuel de traductologie*, Berlin/Boston: De Gruyter, 99-117.

Kilgarriff, Adam (2001) 'Web as Corpus', *Proceedings of Corpus Linguistics 2001 conference*, Lancaster University, 342-344.

Kvashnina, Olga S. & Sumtsova, Olga V. (2018) 'Using Google to Search Language Patterns in Web-Corpus: EFL Writing Pedagogy', *International Journal of Emerging Technologies in Learning* 13(3): 173-179.

Laviosa, Sara (2002) *Corpus-based Translation Studies: Theory, Findings, Applications*, Amsterdam/New York: Rodopi.

Lawrence, Steve & Giles, C. Lee (1999) 'Accessibility of Information on the Web', *Nature*, 107-109.

Loock, Rudy (2016) *La traductologie de corpus*, Lille: PU. Septentrion.

McEnery, Tony & Wilson, Andrew (2001) *Corpus Linguistics*, Edinburgh: Edinburgh University Press.

Meyer, Charles F., Grabowski, Roger, Han, Hung-Yul, Mantzouranis, Konstantin & Moses, Stephanie (2016) 'The World Wide Web as Linguistic Corpus', *Corpus Analysis* 46: 241-254.

Mounin, Georges (1963) *Les problèmes théoriques de la traduction*, Paris: Gallimard.

Resnik, Philip (1999) 'Mining the Web for Bilingual Text', *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 527-534.

Steiner, George (1975) *After Babel: Aspects of language & translation*, New York: Oxford University Press.

Toury, Gideon (1995) *Descriptive Translation Studies—and Beyond*, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Tymoczko, Maria (1998) 'Computerized Corpora and the Future of Translation Studies', *Meta* 43(4): 652-660.

Vinay, Jean-Paul et Darbelnet, Jean (1958) *Stylistique comparée du français et de l'anglais*, Paris: Didier.

Zanettin, Federico (2014a) 'DIY Corpora: The WWW and the Translator', In Belinda Maia, Jonathan Haller & Margherita Urlrych (eds.) *Training the Language Services Provider for the New Millennium*, Porto: Facultade de Letras, Universidade de Porto, 239-248.

Zanettin, Federico (2014b) *Translation-Driven Corpora*, New York: Routledge.

- 인터넷 자료

이성규 (2016.7.27)「인공지능 씨앗 한글 말뭉치, 2007년 멈춰선 까닭」,『블로터』, http://www.bloter.net/archives/260569, 2019년 3월. 2일 검색.

연합뉴스 기사 (2017.02.16)「인간과 AI의 번역 대결은 전문 번역사 승리 유력」, https://www.yna.co.kr/view/AKR20170216122600017?input=1195m, 2019년 3월 17일 검색.

- 어린왕자 코퍼스

Antoine de Saint-Exupéry (1946) *Le Petit Prince*, Paris: Gallimard.

민희식 역 (1986)『어린왕자』, 문학출판사.

이진구 역 (1986)『어린왕자』, 범조사.

정소성 역 (2003)『어린 왕자』, 투영.

김화영 역 (2007)『어린왕자』, 문학동네.

[Abstract]

**The direction of translation studies in the era of the fourth industrial revolution: The role and function of corpus as big data**

Cho, Joon-Hyung
(Korea University)

Translation has been a human activity since ancient times. It carries out an important role for communication and information exchange in many domains these days. We meet and exchange information with other people through translation. These activities produce a large number of written or oral translation texts.

The results of these human activities, as a corpus, become valuable resources for the study of translation in academic, pedagogical and practical fields because of many real translation facts. Therefore, we can carefully observe translation processes with a translation corpus and compare language differences between two languages in question and learn from their principles.

However, the traditional form of corpora has shown its limits as it has not immediately accepted language changes and only gives typical examples of translation. Since the 2010s, some researchers have become interested in a new form of corpus, called “web as corpus.” The web as corpus is a collection of texts from millions of web pages. Because the Internet is the largest existing repository of texts, we can investigate written and spoken languages easily. In addition, the web as corpus contains a sizable amount of normal and abnormal linguistic information, which can be looked at in different ways as translation correspondences between two languages. As such, web as corpus performs an important role in the field of translation studies and translation education in the era of big data.

▸ Keywords: corpus, web as corpus, 4th Industrial Revolution, Big Data, translation corpus
▸ 주제어: 코퍼스, 웹코퍼스, 4차 산업혁명, 빅데이터, 번역 코퍼스

조준형
고려대학교 BK21Plus 번역불문사업팀 연구교수
chojh4net@naver.com
관심분야: 코퍼스 번역학, 번역평가

논문투고일: 2019년 4월 30일
심사완료일: 2019년 5월 25일
게재확정일: 2019년 5월 28일