번역학연구
2012년 겨울 제13권5호

병렬코퍼스에서 맥락 탐색의 의미와 한계

조 준 형
(고려대)

1. 서론

컴퓨터의 발달은 자연과학뿐만 아니라 인문학에서도 현상의 분석을 위한 새로운 접근법을 창출하였다. 연구자는 자료의 수집과 처리, 분석 그리고 결과의 도출로 이어지는 일련의 연구 과정에서 컴퓨터의 도움을 받아 이전보다 능률적인 작업 수행이 가능해 지면서 연구의 효율성을 극대화 시킬 수 있게 되었다. 이러한 변화는 특히 언어학에서 '코퍼스 언어학'이라는 새로운 연구 학제가 탄생할 수 있었던 바탕이 되었다.

1980년대부터 서로 다른 언어 간의 유사성과 차이점을 설명하고자 하는 대조언어학 연구자들이 '코퍼스'에 관심을 가지면서 여러 언어로 작성된 텍스트로 구성된 '다국어 코퍼스'(corpus multilingue)라는 새로운 형태의 코퍼스가 등장하였다. 다국어 코퍼스는 유사한 주제를 가진 서로 다른 언어로 작성된 텍스트로 구축된 '비교코퍼스'(corpus comparable)와 번역 텍스트를 기반으로 하는 '병렬코퍼스'(corpus parallèle) 혹은 '번역코퍼스'(corpus de traduction), 두 가지

형태로 존재한다. 이중에서 번역 연구자들은 번역의 실제 모습을 관찰할 수 있다는 점에서 '병렬코퍼스'를 주목하기 시작하였다. 전문 혹은 아마추어 번역가에 의해 생산된 번역 텍스트로 작성된 '병렬코퍼스'는 실제 번역 사례들을 담고 있는 저장소이자 보고라고 할 수 있다. 오늘날 '병렬코퍼스'는 번역 연구에서뿐만 아니라, 번역 교육 및 자동번역과 같은 다양한 연관 분야에서 응용될 수 있는 유용한 도구로 인식되고 있다. 이러한 이유로 현대 번역학에서 병렬코퍼스는 언급한다는 것 자체가 진부하다고 느껴질 정도로 보편화된 개념이 되었다.[1)]

코퍼스를 기반으로 번역을 연구하는 것은 코퍼스로 사용된 원문과 번역문 사이에서 번역어, 다시 말해서 번역 등가어를 찾아내고, 그 원리를 이해하고 설명하는 데에 일차적인 목적이 있다.[2)] 그런데 병렬코퍼스에서 추출된 결과물은 일반적으로 번역 등가어들을 단순히 나열하는 형태로 제시되는 경우가 많은데, 이 경우에 원문과 번역문 사이에서 해당 등가어의 번역 연관성을 입증할 수 있는 실질적인 증거는 통계 수치와 단어의 형식적인 유사성 이외에는 찾아보기 어렵다. 이를 위해서 병렬코퍼스 기반 연구는 인간 연구자가 컴퓨터의 도움을 받아 해당 등가어가 속한 맥락을 탐색할 수 있는 여러 가지 기법을 제안하고 있다. 그렇지만 번역 연구에서 병렬코퍼스 자체의 범용적인 활용과 달리, 맥락 탐색 기법의 활용은 질적인 측면에서도 양적인 측면에서도 번역학 분야에서 여전히 어떤 한계를 가지고 있는 것 같다.

코퍼스의 현대적인 개념은 컴퓨터와 전문 분석 도구의 활용을 전제한다. 이러한 도구의 활용은 텍스트 자료 처리에 관한 전문 지식[3)]과 형식적인 분석 방

---

1) 넓은 의미에서 번역 연구를 위해 사용된 텍스트가 곧 코퍼스이기 때문에 코퍼스 기반 번역 연구는 매우 오래된 보편적인 연구 방식이라고 할 수 있다. 그러나 컴퓨터학 분야에서 논의되는 코퍼스는 정보화 처리가 된 매우 좁은 의미의 전자텍스트를 가리킨다. 이 분야에서 인문학과 컴퓨터학의 대립은 이러한 코퍼스에 대한 시각적인 차이에서 비롯된다고도 할 수 있다. 예를 들어 정호정(2003)은 코퍼스 기반 번역 연구 동향을 기술하고 있지만, 컴퓨터학과 관련된 기술적인 측면은 전혀 언급하고 있지 않기 때문에 인문학적인 관점의 논의라고 할 수 있다.

2) 발라르(Ballard 2007: 24-25)에 따르면, '코퍼스를 기반으로 한 번역 연구는 원문과 번역문, 두 텍스트의 대조에서 시작한다. 이로 인해 연구는 말하는 방식(manières de parler), 다시 말해서 등가(équivalence)의 비교를 지향하게 된다'.

3) 병렬코퍼스의 정보처리 과정은 단순히 종이 텍스트를 전자 문서화하는 것으로 그치

법[4]을 요구한다. '의미'를 가진 언어의 '형식'적인 처리 그리고 컴퓨터의 전문적인 활용[5]은 인문학 연구자의 관점에서는 쉽게 적응하기 힘든 측면이 있다고 할 수 있다. 그러나 현대 인문학에서도 컴퓨터라는 도구 사용은 매우 보편화되고 있기 때문에, 번역 연구에서도 이를 적극 활용할 필요가 있다.

이러한 관점에서 본고는 인문 번역 연구자들이 맥락 탐색을 위해 활용할 수 있는 기계적인 기법들을 살펴보고, 그러한 기법들에서 나타나는 문제점과 인문학적인 연구 관점에서 어떻게 이를 극복할 수 있을 것인지를 살펴보고자 한다. 이를 위해서 먼저 코퍼스에서 번역 등가어를 추출하고 이를 검증하기 위해서 사용하는 맥락 탐색이 전통적인 인문학적인 방법과 기계적인 분석을 전제하는 병렬코퍼스 기반 연구 방법 사이에서 어떤 차이점을 보이는지를 설명할 것이다.[6] 다음으로 병렬코퍼스 연구에서 맥락 탐색을 위해 많이 활용되는 두 가지 기법인 '키워드 검색'(KWIC, Keyword in context)과 '공기어'(cooccurrence) 분석을 소개하면서, 인문학적인 관점에서 이 기법들의 유효성과 한계를 되짚어 보고, 이를 통해 번역 연구에서 기존 연구 방식과 현대적인 의미의 코퍼스 기반 연구와의 협력 가능성을 살펴보고자 한다.

는 것이 아니라 컴퓨터 및 분석 도구가 텍스트 자료를 제대로 인식할 수 있도록 하기 위해서 *XML* 혹은 이에 준하는 구조 표기(structural markup)와 같은 정보 처리 과정을 요구한다.

4) 코퍼스에 포함된 자료 분석을 위해서 활용되는 가장 일반적인 방식이 바로 어휘들의 출현 빈도수(frequency)와 같은 통계 기법들이다.

5) 인문학과 컴퓨터학의 연계가 부족한 국내의 현실에서 컴퓨터의 활용과 전산화 작업은 인문학 연구자들이 보다 적극적으로 코퍼스에 접근하는 데에 커다란 장벽이 될 수 있다. 그러나 코퍼스 분석을 위한 전문 분석 도구에는 미치지 못하지만, 마이크로소프트(Microsoft)의 엑셀(EXCEL)과 같은 프로그램을 사용하면 코퍼스에서 번역 등가어 검색 같은 기초적인 작업은 쉽게 이루어질 수 있다. 엑셀은 회계 및 통계 분석을 목적으로 하는 프로그램이지만, 수치 자료뿐만 아니라 언어 자료 분석도 가능하기 때문에 훌륭한 데이터베이스 프로그램이기도 하다. 이러한 프로그램을 활용한다면 인문학 연구자들도 코퍼스를 활용한 번역 연구를 충분히 진행할 수 있다.

6) 본 논문에서 '병렬코퍼스 기반 연구'는 종이 형태로 된 고전적인 텍스트를 전자텍스트로 전환하고, 일정한 전처리 과정을 거쳐 코퍼스를 처리할 수 있는 전문 분석 프로그램을 통해서 원하는 자료를 추출하고 이를 인간 연구자가 분석하는 과정을 총칭한다. 전자텍스트라도 분석 도구의 도움 없이 컴퓨터상에서 인간 연구자가 이를 직접 관찰한다면 기존의 연구 방법으로 간주하였다.

2. 번역학과 병렬코퍼스

2.1. 번역학에서 병렬코퍼스의 유용성

번역은 이질적인 언어, 문화, 사상의 만남이 이루어지는 복합적인 공간이라고 할 수 있다. 번역학은 '번역'이 내포하고 있는 이러한 다면적인 만남의 현상을 체계적으로 설명하고자 한다. 이를 위해서 번역학은 특정한 사유에만 의존하는 것이 아니라, 다양한 언어적, 인식론적, 해석적 방법들을 활용한다. '번역학'을 특징짓는 다중학제적 성격은 바로 이러한 이유에서 비롯된다고 할 수 있다. 그렇지만 발라르(2007: 17-18)가 말하듯이 번역 연구에 있어서 가장 핵심적인 것은 바로 언어, 보다 엄밀하게 말해서 언어와 텍스트라고 할 수 있다. 발라르(2006)는 번역학을 '관찰의 학문'이라고 말한다. 번역 연구는 이미 나와 있는 번역 결과물을 면밀하게 관찰하는 것에서 출발하기 때문이다. 이러한 관점에서 번역학은 관념적인 사유의 학문이기 이전에, 번역 현상을 탐구하는 경험적 학문이라고 할 수 있다. 홈스(Holmes 1972)와 투리(Toury 1995)는 번역학의 경험적 특성에 주목하여 기술론적인 방법론을 주장하는데, 그들은 실제 번역 텍스트의 관찰을 통해서 번역 현상을 설명하고 이를 규범화시키고자 한다.

번역 연구의 이러한 경험적 특성에 주목한다면, 병렬코퍼스는 이 분야에 있어서 최적의 분석 자료가 될 수 있다. 병렬코퍼스는 실제로 출판된 번역 텍스트들로 구성되기 때문이다. 번역 연구에서 본격적인 병렬코퍼스의 활용을 주장하고 구체적인 분석 기법을 소개한 베이커(Baker 1995) 그리고 자동번역에서 병렬코퍼스의 중요성을 주장한 이자벨과 워릭-암스트롱(Isabelle & Warwick-Armstrong 1993)은 모두 병렬코퍼스가 가지는 그러한 특성에 주목했기 때문이다.[7)]

---

7) 기존의 자동번역은 원문에서 번역문으로 문법 및 어휘 구조의 기계적인 전환에만 의존했기 때문에, 완전히 다른 언어적인 특성을 가진 언어들 사이의 번역은 매우 조악한 결과를 제시하였다. 특히 서양어와 한국어, 중국어, 일본어와 같은 아시아 언어 사이의 자동번역이 이에 해당하였다. 그런데 현재의 자동번역은 형식적인 언어 구조와 더불어 실제 번역 등가어들에 해당하는 일종 번역사전을 보조 수단으로 하여 번역품질을 개선하려는 시도를 하고 있다. 이때 병렬코퍼스가 번역 등가어들의 원천이 될 수가 있기 때문에, 특히 자동번역 연구자들이 병렬코퍼스의 유용성에 주목하고 있다.

많은 번역 용례와 함께 번역 현상의 다양한 유형을 기술하고 있는 비네와 다르벨네(Vinay & Darbelnet)의 『불어와 영어의 비교문체론(*Stylistique comparée du français et de l'anglais*)』(1977)은 넓은 의미에서 코퍼스 기반 번역 연구의 전형이라고 할 수 있다. 병렬코퍼스에서 번역 등가어들을 탐색하고 그 결과를 목록화하는 과정이 마치 비네와 다르벨네가 수많은 번역 자료의 연구 분석을 통해서 결과물로 내놓은 『비교문체론』에 상응하기 때문이다. 만일 그들이 연구 작업을 하던 시기에 현대적 의미의 병렬코퍼스와 분석 기법이 존재했다면,[8] 오늘날 전혀 새로운 『비교문체론』이 출판되었을지도 모른다.[9]

2.2. 병렬코퍼스에서 텍스트 탐색의 의미

번역 텍스트에서 번역어의 '관찰'이라는 것은 형식적인 언어의 모습을 관찰한다는 것을 의미하는가? 번역 차원에서 언어 문제를 논의하기 위해서는 단순히 형식적인 측면만이 아니라 여러 가지 복합적인 요소를 고려해야만 한다. 번역 텍스트에서 언어의 관찰은 텍스트라는 맥락 속에서 원문과 번역문 사이에서 일어나는 어휘, 의미, 문화 등의 이행과정을 살피는 것이다. 텍스트 내에서 언어 요소는 독립적으로 존재하는 것이 아니며, 다른 언어 요소와의 상호작용에 의해서 존재 이유를 가진다. 다시 말해서 번역 텍스트에서 한 어휘의 번역을 관찰하는 것은 해당 어휘만을 고려하는 것이 아니라 다른 어휘들과의 관계를 살피는 것이고, 따라서 그 어휘들이 속한 텍스트라는 거대한 맥락(contexte)을 탐구하는 것이다. 이 맥락 속에서 어휘와 어휘, 어휘와 텍스트 간의 관계를 복합적으로 탐구하는 과정이 곧 번역 연구이다. 전성기(2009:135)가 언급하듯이,

---

자동번역에 관해서는 Fuchs(1993), Loffler-Laurian(1996) 참조.

8) 비네와 다르벨네의 저작은 1958년에 처음으로 출판되었다. 반면에 현대적 의미의 최초의 병렬코퍼스는 1980년대에 등장한다.

9) 다른 관점에서 보면, 만일 비네와 다르벨네가 현대적 코퍼스를 기반으로 연구를 했다면 현재 수많은 번역 관련 연구논문에서 인용되는 『비교문체론』이 빛을 보지 못했을 수도 있다. 코퍼스 기반 연구는 분석 대상이 코퍼스에 포함된 텍스트에 한정된다는 단점이 있기 때문이다. 또한 『비교문체론』은 다양한 분야에 속하는 여러 가지 예들을 제시하고 있는데, 이를 위해서는 병렬코퍼스가 매우 다양한 장르에 속하는 텍스트를 포함해야하는데, 이러한 코퍼스를 구축한다는 것은 결코 쉬운 일이 아니기 때문이다.

“텍스트가 언어로 이루어진 이상 텍스트의 언어에 대한 탐구는 불가피하다. 맥락과 불가분의 관계에 있는 이 언어의 탐구는 기본적으로 맥락을 비롯한 다양한 요소들을 고려하는 부단한 문답 과정이다.”[10] 이러한 문답 과정을 통해서 연구자는 원문과 번역문 사이에서 어휘 혹은 표현의 정확한 번역 관계를 설명할 수 있다.

코퍼스 기반 연구에서도 번역 텍스트, 다시 말해서 맥락은 번역어들의 면밀한 관찰을 위해 궁극적으로 회귀할 수밖에 없는 출발점이기 때문에, 끊임없는 문답 과정을 통한 탐구과정은 병렬코퍼스 기반 연구에서도 필연적인 과정이라고 할 수 있다. 그러나 문답의 과정이 번역학의 전통적인 연구 방식과 코퍼스 기반 번역 연구 방식이 동일하다고 말하기는 어려운 것 같다. 전통적인 연구 방식과 달리 코퍼스 기반 연구는 방법론적인 성격 때문에 언어자동처리(Traitement automatique des langues) 분야와 밀접한 관련을 가지고 있다. 언어자동처리는 텍스트의 기계적인 처리와 전문 도구에 의한 언어자료 분석을 전제한다. 따라서 인간 연구자가 텍스트 혹은 맥락에 접근하는 태도가 달라질 수밖에 없다. 고전적인 연구 방식과 코퍼스 기반 연구 방식은 다음과 같이 차별화할 수 있을 것이다.

〈표 1〉 전통 연구와 코퍼스 기반 연구에서 맥락 관찰의 차이

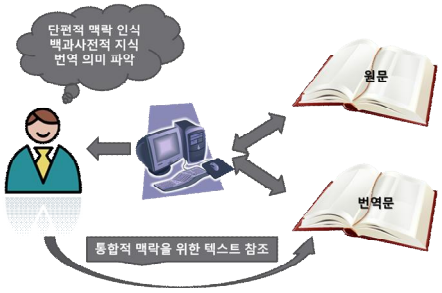
	전통적 연구 방식	코퍼스 기반 연구 방식
분석 과정	텍스트의 순차적 관찰	텍스트 표본 관찰
텍스트 접근성	선 접근성	후 접근성
텍스트 해석	인간의 사유	기계의 분석+인간의 사유
맥락 분석	맥락의 직접적 접근성	맥락의 간접적 접근성

전통적인 연구 방식에서 연구자는 번역 텍스트를 맥락에 따라서 순차적으로 강독을 하면서 번역 현상을 관찰하고 필요한 용례를 선별하게 된다. 따라서 연구자는 원칙적으로 번역과 번역 맥락을 동시에 인식하면서 연구를 진행하게

10) 전성기(2009)는 ‘탐구번역’을 통해서 번역 연구에서 지향해야 할 연구 태도를 제시하는데, 그의 ‘탐구번역론’은 한정적인 의미의 번역 연구뿐만 아니라 인문학 번역 전체에 적용되는 보편적이고 인식론적인 개념이다.

된다. 반면 코퍼스 기반 연구에서 연구자는 일차적으로 텍스트에 직접적으로 접근하기가 어렵다고 할 수 있다. <그림 1>에서 보듯이 연구자는 텍스트에 직접적으로 접근하는 것이 아니라, 컴퓨터를 매개로 해서 간접적으로 텍스트에 다가가기 때문이다.

〈그림 1〉 코퍼스 기반 연구에서 텍스트 접근성



보다 정확히 말하면, 연구자는 텍스트 자체가 아닌, 코퍼스 분석 도구에 의해 추출된 결과들을 먼저 접하게 된다. 그리고 그 결과물은 흔히 단어 형태의 언어 자료와 빈도수 형태의 통계 자료들로 구성된다. 이 결과물은 비록 번역 등가어 목록이라고 하더라도 어떤 맥락적인 근거를 동시에 제시하고 있지 않기 때문에, 연구자가 이 목록만을 가지고 해당 어휘의 번역 관계를 판단하는 것은 매우 위험한 일이다. 따라서 코퍼스 기반 연구는 개별 단어와 빈도수의 조합으로 된 결과물에서 번역어들의 의미를 보다 확실하게 파악하기 위해서 다시 맥락으로 돌아가는 회귀적인 방식을 취하게 된다.

참조하려는 맥락 역시도 전통적인 연구 방식에서의 그것과는 다른 특성을 보인다. 기존의 방식에서 연구자는 텍스트를 직접적으로 관찰하기 때문에 상당히 넓은 범위의 맥락을 참조할 수 있다. 반면에 코퍼스에서 맥락 탐색에 의해 연구자가 확인하는 맥락들은 매우 단편적이며 각 맥락과의 연계성도 확인할 수 없다. 왜냐하면 분석 도구에 의해 추출된 맥락은 특별한 기준에 의해 선정된 매우 작은 범위, 일반적으로 문장 혹은 문단에 해당하는 범위로 한정되기 때문이다. 또한 이 맥락들 간의 연계성도 보장되지 않는데, 분석 도구는 분석 대상이 된 어휘가 포함된 맥락만을 번역 등가어 목록과 같은 형식으로 보여주기 때문

이다. 물론 코퍼스 기반 맥락 검색에서 맥락의 범위를 어떻게 설정하느냐에 따라서 연구자가 확인할 수 있는 맥락 텍스트의 크기는 가변적이지만, 연구의 기초 단계에서 텍스트 전체에 접근하는 것이 불가능하다는 근본적인 한계는 코퍼스 기반 연구에서 분명하게 고려해야 할 점이다.[11]

3. 번역코퍼스에서 맥락 기준

번역을 이야기할 때 우리는 흔히 '맥락'이라는 말을 많이 사용한다. 그렇다면 '맥락'이란 무엇인가? 접근 방식에 따라서 여러 가지 정의가 가능하겠지만, 언어학 및 커뮤니케이션의 관점에서 맥락은 '특정 언어 요소가 속해 있는 텍스트'(Dubois *et al*. 2001: 116)이며, '하나의 발화가 창출되는 사회적·문화적 배경'(*Ibid*.)이라고 할 수 있다. 텍스트는 의미를 가진 최소 자립 단위인 단어로 이루어져 있다. 각 단어들은 앞뒤에 출현하는 또 다른 단어들과의 관계 속에서 '구(句)', '문장', '문단'이라는 보다 상위의 텍스트 구성단위를 차례로 형성한다. 이러한 구성단위들은 각각이 서로 다른 구성단위들과 상호작용을 하면서 텍스트 전체의 의미를 생성하게 된다. 인간은 순차적으로 나타나는 텍스트 구성단위들을 받아들임과 동시에 뇌에서 이들의 관계망, 다시 말해서 맥락 속에서 텍스트가 창출하는 의미를 인식론적으로 이해하게 된다. 텍스트 구성 요소들 간의 상호관계를 정확하게 파악하지 못한다면 텍스트의 의미를 명확하게 이해하

11) 코퍼스의 규모가 크지 않다면 사용된 텍스트의 종류 및 크기도 매우 한정적이기 때문에 번역어들의 검토 과정에서 텍스트 전체를 관찰하는 것은 그리 어려운 작업은 아니라고 할 수 있다. 더욱이 대부분의 일반적인 코퍼스 분석은 하나의 문학작품, 한 종류의 텍스트를 대상으로 하는 경우가 많기 때문에 원텍스트를 참조하면서 연구를 진행하는 것이 가능하다. 반면 다양한 종류의 텍스트로 작성된 대규모 균형 병렬코퍼스를 기반으로 한다면 텍스트의 일차적인 접근성은 상대적으로 상당히 낮아질 수밖에 없다. 모든 텍스트를 동시에 참조할 수는 없기 때문이다. 따라서 분석 대상 어휘가 어떤 맥락에 속하고, 이 맥락이 어떤 텍스트에 해당하는지를 알려주는 특별한 표지가 없다면 연구자는 단편적인 맥락 속에서만 번역어들을 관찰하게 되고, 이들의 번역적 판단을 위해서는 통계 수치와 언어의 형식적인 유사성에 의존할 가능성이 그만큼 더 증가할 수밖에 없다.

지 못하게 되고, 이러한 상태에서 텍스트를 다른 언어로 옮기게 되면 충실하지 못한 번역의 결과물에 이르게 된다. 따라서 번역 연구에서 맥락은 번역 판단을 위한 바탕이 되는 것이라고 할 수 있다.

반면 컴퓨터의 도움을 받는 코퍼스 기반 연구에서 텍스트는 형식적인 차원에서 인식될 수밖에 없다. 기계의 관점에서 맥락의 개념은 '어휘들의 집합'이라는 지극히 형식적인 태도를 취하게 된다.[12] 물론 언어자동처리 분야에서 '의미망', '어휘망'과 같은 개념을 활용하면서 코퍼스의 맥락 탐색에 관한 많은 연구가 이루어지고 있지만 궁극적으로 기계가 언어 요소들 간의 관계를 인식하는 것은 선험적으로 주어지는 데이터베이스와 통계 정보(빈도수, 언어 요소 간의 거리)에 의존하기 때문에 인간 연구자가 가지는 맥락에 대한 정의와는 분명한 차이를 보일 수밖에 없다고 할 수 있다.

고전적인 번역 연구와 마찬가지로 코퍼스 기반 번역 연구에서도 맥락은 매우 중요한 개념이다. 앞서 언급한 것처럼 텍스트를 직접 관찰하는 전통적인 연구 방법과 달리 코퍼스에 의한 연구는 분석 도구를 사용해서 특정한 기준에 의해 선별된 어휘 자료를 가지고 번역 판단을 내리게 된다. 그러나 독립된 단어만을 가지고 번역 연관성을 찾아낸다는 것은 매우 위험한 일이기 때문에, 정확한 번역 등가를 찾기 위해 맥락을 다시 참조할 수밖에 없다.

컴퓨터의 관점에서 단어는 공백에 의해 구분되는 일정한 문자의 집합이며, 맥락은 이러한 문자 집합의 집합이라고 할 수 있다. 그런데 맥락 검색을 통해 나온 결과물에 대한 판단은 인간 연구자에 의해 이루어지기 때문에 가능하다면 연구자가 번역 결과물에 대한 판단을 충분히 할 수 있는 분명한 의미 단위, 다시 말해서 '번역 단위'(unité de traduction)[13]를 기준으로 맥락을 보여주는 것이

12) 언어자동처리의 관점에서 보면, 코퍼스에 포함된 '어휘' 혹은 '단어'는 의미를 가진 언어 요소가 아니라 문자의 집합인 토큰(token 혹은 occurrence)이다(Lebart & Salem 1994: 36). 따라서 맥락은 일정한 수의 토큰의 집합이라고 할 수 있다. 컴퓨터의 관점에서 맥락을 탐색한다는 것은 바로 토큰의 집합을 탐색하는 것이고, 거기서 일정한 유형의 토큰, 즉 번역 등가어를 추출한다는 것을 의미한다.

13) 비네와 다르벨네(1977: 36-43)에 있어 '번역 단위'는 '사고의 단위'(unité de pensée)이며, 통사적 차원이든 어휘적 차원이든 번역 과정에서 분리해서 생각할 수 없는 하나의 의미적 단위라고 할 수 있다. 물론 원텍스트와 번역 텍스트 사이에서 번역 단위의 형식이 동일할 수 없으며, 서로 다른 형태의 단위로 변형될 수 있으며, '번역

타당하다. 물론 번역 단위에 대한 정의도 인간과 기계 사이에서 차별화 될 수밖에 없다. 형식적인 차원에서 번역 단위는 컴퓨터가 텍스트 내의 번역 요소를 식별할 수 있는 기준이 되는 범주를 가리킨다. 특히 병렬코퍼스의 전처리 과정인 코퍼스의 자동 정렬에서 번역 단위는 원문과 번역문 사이에서 컴퓨터가 번역 정렬을 수행할 때 필요한 식별 기준을 의미한다. 일반적으로 '구', '문장', '문단'이 번역 정렬의 기준으로 사용되었기 때문에, 형식적인 차원에서의 번역 단위는 곧 이들을 가리킨다.[14]

코퍼스 연구에서는 병렬코퍼스의 기계적인 처리 문제를 고려하지 않을 수 없다. 또한 연구자가 한 단어의 맥락 환경을 살필 때, 가능한 한 완전한 맥락을 제공할 필요가 있다. 이러한 다양한 환경적인 변수를 고려해야 하기 때문에, 코퍼스 분석 도구는 문장 혹은 문단을 맥락 기준으로 설정하는 것이 일반적이다. 그런데 번역 단위의 범위가 넓을수록 어떤 특별한 정보가 없다면 컴퓨터가 인식하는 어휘들의 연관성은 약화될 수 있기 때문에, '문단'은 맥락 기준으로 너무 큰 단위일 수 있다. 보다 섬세한 분석을 위해서 '구'를 번역 단위로 지정할 수도 있으나, 언어 차원뿐만 아니라 언어자동처리 관점에서도 이를 처리하는 것이 매우 난해하다. 따라서 문장 수준의 번역 단위가 분석 도구를 활용해서 맥락을 탐색할 때 고려할 수 있는 가장 일반적인 맥락 기준이라고 할 수 있다.

맥락 차원에서 마지막으로 고려해야 할 것은 병렬코퍼스는 번역 텍스트로 구성되었다는 점이다. 따라서 단순히 '문장'을 번역 맥락의 기준으로 고려한다면 여러 가지 문제가 야기될 수 있다. 원문과 번역문에서 문장 수준의 번역 대응 방식이 다양한 형태로 실현될 수 있다는 점을 고려해야만 한다. 원문에서 하나의 문장이 번역문에서는 두 개 혹은 그 이상의 문장으로 번역되거나 혹은 그 반대일 수 있기 때문이다. 본 논문에서 이러한 병렬코퍼스의 문장 정렬의 결과

단위'에 관한 비네와 다르벨네의 정의도 이론 번역학적인 관점에서 많은 논쟁을 야기하는 주제 중의 하나이다. '번역 단위'에 관한 비판적 논의에 대해서는 발라르(1992, 1993a, 1993b) 참조.

14) 현재는 '문장'이 가장 보편적인 정렬 기준이 되고 있다. 문단은 범위가 너무 크고, 구는 언어 간의 차이에 의해서 컴퓨터로 처리를 하기에는 해결하기 힘든 문제들을 내포하고 있기 때문이다. 병렬코퍼스의 자동정렬에 관해서는 Brown, Lai & Mercer(1991), Gale & Church(1993), Kay & in Röscheisen(1993) 참조.

를 '문장정렬맥락'이라고 부를 것이다. '문장정렬맥락'은 원문과 번역문 사이에서 1:1, 1:2 혹은 2:1, 다중대응과 같은 여러 가지 유형으로 실현되는 문장 수준의 형식적인 번역 맥락 단위를 가리킨다. 병렬코퍼스 기반 연구에서 '문장' 혹은 '문장정렬맥락' 중에서 어떤 것을 맥락의 기준으로 설정하는지에 따라서 그 결과가 달라질 수 있고, 혹은 새로운 형식적인 문제를 야기할 수 있다. 이에 대한 문제는 이후에 소개할 맥락 탐색 기법을 통해 논의할 것이다.

4. 번역 코퍼스에서 맥락 탐색 기법

4.1. 키워드 검색(KWIC)

4.1.1. 키워드 검색의 개념

키워드 검색은 코퍼스 기반 연구에서 가장 기초적인 어휘 검색 방법이다. 현재도 널리 사용되는 키워드 검색은 특정 어휘를 중심으로 앞과 뒤로 배열되는 언어 요소들을 제시하여 연구자가 해당 어휘의 출현 맥락을 관찰할 수 있는 '어휘색인'(Concordance)을 산출한다. 어휘색인은 키워드를 중심으로 전후 어휘들의 철자 배열을 기준으로 한 정렬된 일정한 범위의 맥락 목록을 보여준다. 키워드 검색 방법은 키워드와 연관된 어휘들의 횡적 배열을 볼 수 있기 때문에, 키워드를 중심으로 앞뒤로 확장되는 어휘 요소들의 관계를 직접적으로 파악할 수 있다는 이점이 있다. 코퍼스 기반 연구에서 가장 기초적인 키워드 검색 방법은 병렬코퍼스에서도 가장 기본적인 맥락 탐색 기법이라고 할 수 있다. 연구자는 원문과 번역문 사이에서 번역 대응 가능한 어휘들을 키워드로 검색하여 추출된 각 어휘들의 맥락을 비교하면서 대상이 된 어휘들의 맥락의 특징을 관찰하고 번역 가능성을 판단할 수 있다.[15]

15) 병렬코퍼스에서 키워드 검색의 결과는 원문과 번역문 각각의 어휘색인이 병치가 되기 때문에 '병렬어휘색인'(Bi-concordance)이라고 부른다. 키워드 검색을 활용한 번역 탐색의 한 예로 베이커(1995)를 참조할 수 있다.

〈표 2〉 병렬어휘색인의 예(*enfant*/아이)

맥락	원문		
4	to cantabile, dit l'	**enfant**	.
5			
6	#l'	**enfant**	resta immobile, la
11	§#l'	**enfant**	ne répondit pas.
13	#pas un cil de l'	**enfant**	ne bougea.
18	§#l'	**enfant**	, immobile, les yeux
19			
21	§#l'	**enfant**	ne jugea pas bon de
22			
26	ssire considéra cet	**enfant**	de ses pieds jusqu'
맥락	번역문		
4	§#모데라토 칸타빌레하고	**아이**	가 대답하였다.
5	§#피아노 선생은	**아이**	의 말에 구두점을 찍기라도
6	#	**아이**	는 얼굴을 악보로 향한 채
11	§#	**아이**	는 대답하지 않았다.
13	#	**아이**	는 눈썹 하나 까딱하지 않았다
18	내리깐채 꼼짝도 않고 있는	**아이**	만 이제 막 어스름 저녁이 되
19	#거기에 생각이 미치자	**아이**	는 오싹한 기분이 들었다
21	§#	**아이**	는 대답하지 않기로 작정하
22	앞에 목석처럼 앉아 있는 이	**아이**	를 한 번 더 쳐다보았다.
26	§#안데바레드도	**아이**	를 머리부터 발끝까지 훑어보

그런데 <표 2>에서 알 수 있듯이 어휘색인은 키워드를 중심으로 전후 어휘들의 철자 순서 이외에는 특별한 배열 기준 없이 너무 많은 정보를 제시하기 때문에, 연구자가 해당 어휘의 전체적인 맥락 환경을 살피는 데에는 많은 어려움이 있을 수 있다. 또한 서로 다른 언어를 비교하는 번역 연구에서 각 언어의 어휘 배열 및 구조가 다르기 때문에, 번역 대응어의 어휘색인을 직접적으로 비교하여 유의미적인 맥락 특징을 찾아내는 것도 쉽지 않다. 이를 위해서 어휘색인의 표준화 작업이 필요하다.

4.1.2. 어휘색인의 표준화 작업

키워드 검색의 결과물인 병렬어휘색인에서 키워드를 중심으로 전후로 나열되는 어휘의 배열을 보다 분명하게 확인하기 위해서 어휘색인 표준화 작업이

요구된다. 이 방법은 어휘색인에서 키워드를 중심으로 한 어휘의 배열 중에서 동일한 통사 구조를 가지는 형태를 빈도수와 함께 재분류하여 연구자가 함께 출현하는 어휘들의 구조를 명시적으로 확인할 수 있도록 한다. 이를 바탕으로 연구자는 번역어들의 맥락 구조를 비교하여 이 어휘들의 실제 번역 가능성을 조금 더 명확하게 판단할 수 있다.[16]

〈표 3〉 병렬어휘색인의 표준화 작업의 한 예

conditions prévues par le **droit**	---- ---- 4
dans les conditions prévues par le **droit**	---- 2
protégés par le **droit**	---- ---- ---- 2
sont régis par le **droit**	---- ---- 2
pour le **droit**	---- ---- ---- ---- 3
prévoit le **droit**	---- ---- ---- ---- 2
proclame le **droit**	---- ---- ---- ---- 2
ou complétant celle que procure le **droit**	---- 2

75 ---- ---- ---- ---- ----	**권리** 가
2 ---- ---- ---- ----	**권리** 가 결정
2 ---- ---- ---- ----	**권리** 가 실행
4 ---- ---- ----	**권리** 가 인정 된다
17 ---- ---- ---- ----	**권리** 가 있다
2 ---- ---- ---- ----	**권리** 가 있다는
2 ---- ---- ----	**권리** 가 있다고 선언
6 ---- ---- ---- ----	**권리** 가 있으며
7 ---- ---- ---- ----	**권리** 가 있음을

<표 3>에서 확인할 수 있듯이 어휘색인 표준화의 결과는 키워드를 중심으로 확장되어가는 어휘들의 모습을 명시적으로 관찰할 수 있을 뿐만 아니라, 확장 수준에 따라서 코퍼스에서의 출현 빈도를 확인할 수 있기 때문에, <표 2>에서 보이는 다소 무질서한 맥락의 배열보다는 코퍼스 내에서 키워드의 어휘적 양태를 보다 명시적으로 이해할 수 있다. 따라서 '어휘색인 표준화'는 기존의 키워드 검색에서 보다 발전된 형태의 기법이라고 할 수 있다.

16) 어휘색인의 표준화 작업에 관해서는 르바르와 살렘(Lebart & Salem 1994: 63-70), 조준형(2010: 256-269) 참조.

4.1.3. 키워드 검색에서 맥락의 문제

기존의 키워드 검색은 맥락에 관한 특별한 기준을 정하지 않고 원시 코퍼스를 그대로 활용했기 때문에, 어휘색인으로 제시된 맥락은 완전한 텍스트 구성단위가 아니라 불완전한 형태로 제시되었다(<표 2> 참조). 전통적인 키워드 검색 도구는 맥락의 범위를 지정할 때, 앞서 우리가 언급했던 것처럼 '문장' 혹은 '문단'과 같은 텍스트 구성단위를 기준으로 한 것이 아니라 키워드를 중심으로 한 어휘의 수를 기준으로 했기 때문이다. 따라서 연구자가 키워드를 중심으로 완전한 맥락을 확인하는 것은 불가능했다.

반면에 최근의 병렬코퍼스 분석 도구들은 코퍼스의 전처리 과정에서 이미 문장 수준의 번역 정렬 과정을 거치기 때문에 완벽한 문장을 맥락으로 제시하는 것이 가능하게 되었다. 예를 들어 엑셀을 이용해서 다음과 같이 문장 수준의 맥락으로 구성된 병렬어휘색인을 작성할 수 있다.

〈그림 2〉 엑셀을 활용한 병렬어휘색인

	A	B	C	D
1		français		coréen
5	$	§#- Moderato cantabile, dit l'enfant.	$	#"모데라토 칸타빌레"하고 소년이 대답했다.
7	$	#L'enfant resta immobile, la tête tournée vers sa partition.	$	#소년은 얼굴을 악보로 향한 채 움직이지 않았다.
12	$	§#L'enfant ne répondit pas.	$	§#소년은 대답하지 않았다.
14	$	#Pas un cil de l'enfant ne bougea.	$	#소년은 눈썹 하나 까딱하지 않았다.
19	$	§#L'enfant, immobile, les yeux baissés, fut seul à se souvenir que le soir venait d'éclater.	$	§#꼼짝 않고 눈을 아래로 향한 채 어둠이 벌써 활짝 번졌구나 하고 생각한 것은 그 소년뿐이었다.
22	$	§#L'enfant ne jugea pas bon de répondre.	$	§#소년은 대답하지 않는 것이 좋으리라고 생각했다.
27	$	§#Anne Desbaresdes aussi reconsidéra cet enfant de ses pieds jusqu'à sa tête mais d'une autre façon que la dame.	$	§#안느 데바레드도 발끝에서부터 머리끝까지 이 소년을 훑어 보았다. #그러나 피아노 선생과는 다른 표정으로.
29	$	§#L'enfant ne témoigna aucune surprise.	$	§#소년은 아무런 놀라움도 나타내지 않았다.
32	$	#Tout à côté des mains de l'enfant.	$	#바로 소년의 손 옆에서.

<그림 2>는 프랑스 소설인 마르그리트 뒤라스(Marguerite Duras)의 『모데라토 칸타빌레(*Moderato Cantabile*)』에서 *enfant*이 포함된 프랑스어 맥락과 그것의 한국어 번역 맥락을 나란히 병치한 병렬어휘색인이다. 검색의 매개변수를 조절하면 프랑스어 원문과 한국어 번역본 사이에서 다양한 번역 등가 검색이 가능하다. 각 맥락은 완전한 문장들을 포함하고 있기 때문에, 키워드의 의미를

쉽게 파악할 수 있다.

병렬코퍼스에서 맥락 범위는 연구 목적에 따라서 '문장' 혹은 '번역정렬맥락', 두 가지 기준으로 변화를 줄 수가 있다고 하였다. 먼저 가장 일반적인 기준인 '문장'을 맥락 범주 기준으로 설정할 수 있다. 문장을 기준으로 한 맥락 검색은 기존의 키워드 검색과 동일한 결과를 보여준다. 단지 기존의 키워드 검색과 달리 완벽한 문장을 제시하기 때문에 연구자는 해당 어휘의 언어적 환경을 기존의 키워드 검색보다는 명시적으로 확인할 수 있다. 그런데 앞서 언급한 것처럼, 원문과 번역본 사이에서 문장 수준의 번역 대응관계는 다양한 유형으로 나타난다. 이러한 점 때문에 단순히 문장을 맥락 기준으로 설정하게 되면 정확한 번역 맥락을 비교한다고 단언하기는 어렵다. 이를 위해서 문장 수준의 번역 대응 유형을 충분히 고려하여 '문장정렬맥락'을 맥락 기준으로 설정할 수 있다. 문장 수준 번역정렬을 의미하는 '문장정렬맥락'을 맥락 기준으로 설정하면 원문과 번역본 사이에서 완벽한 번역 대응을 이루는 맥락 검색이 가능해진다.

한편 병렬코퍼스에서 문장을 맥락 기준으로 설정한 키워드 검색이 완전히 무의미하다고 말할 수는 없는데, 번역 차원을 넘어서 코퍼스 내에서 번역 후보 어휘들의 어휘적·통사적 양태를 대조언어학적인 관점에서 관찰할 때는 유용한 방법이 될 수 있다. 마치 비교코퍼스를 기반으로 특정 어휘들의 다양한 언어적 양태를 비교하는 것에 해당하기 때문이다. 따라서 연구 목적에 맞게 맥락 기준으로 '문장정렬맥락'과 '문장'을 적절하게 사용하는 것이 필요하다.

4.2. 공기어 검색

4.2.1. 공기어 검색의 개념

'공기어'(Cooccurrence)란 일정한 맥락 범주에서 키워드를 중심으로 함께 출현하는 어휘들을 가리킨다. 하나의 키워드를 중심으로 출현하는 공기어를 살펴보고자 할 때 가장 기본적인 방식이 앞서 언급했던 키워드 검색이라고 할 수 있다. 키워드 검색은 키워드가 속한 맥락을 목록화하여 보여준다는 점에서 유용한 방법일 수 있지만, 분석을 위한 특별한 표지 없이 맥락 자체를 그대로 보여준다는 점에서 공기어들 자체의 출현 양상을 이해하는 데에는 충분하지 못하

다. 따라서 키워드가 포함된 번역 맥락에 나타나는 공기어의 분포 특징을 분석하기 위한 '공기어 검색' 기법이 필요하다.

'공기어 검색'의 주목적은 특정 어휘를 중심으로 나타나는 어휘들의 분포 특징과 맥락에 따른 출현 공기어들 자체의 특징을 관찰하는 데에 있다. 일반적으로 공기어 검색은 보다 전문적인 분석 도구를 필요로 한다. 왜냐하면 공기어 검색의 결과물은 키워드가 되는 어휘와 공기어 사이의 연관성을 설명하기 위한 추가적인 정보들이 필요하기 때문이다. 이러한 정보는 일반적으로 통계 자료의 형태로 제공된다. 통계 자료는 각 어휘들의 빈도수뿐만 아니라 어휘들의 공기 연관성을 의미하는 확률값 등으로 제시된다.[17]

<표 4>는 『모데라토 칸타빌레』의 프랑스어 원문에서는 *moderato* 그리고 한국어 번역본에서는 '모데라토'를 키워드로 하여 검색한 공기어 목록이다.[18] 각 목록은 공기어를 중심으로 각 공기어의 빈도수(F), 그리고 키워드와 공기어의 공기빈도수(co), 마지막으로 공기어 빈도변화를 의미하는 '공기빈도특이성'(sp)이 동시에 표시되어 있다.

〈표 4〉 공기어 검색의 한 예(*moderato*/모데라토)

프랑스어 원문				한국어 번역본			
어휘	F	co	sp	어휘	F	co	sp
leçon	7	1	2.5	노래	22	1	2.1
sonatine	23	1	2	깐따빌레	6	6	15.5
dame	35	1	1.8	소나띠네	23	1	2.1
enfant	150	1	1.3	선생	48	2	2.8
rythme	1	1	3.4	대답	17	1	2.2
fin	14	1	2.2	소년	173	2	1.8
cantabile	8	8	20	솜씨	3	1	2.9
				의미	3	1	2.9
				뜻	10	3	6.2
				레슨	15	1	2.2

17) 본 논문의 주제는 공기어에 관한 것이 아니기 때문에 공기어 검색의 원리와 통계적 원리에 대해서는 구체적으로 언급하지 않을 것이다. 이에 대해서는 라퐁(Lafon 1981, 1984), 마르티네즈(Martinez 2000, 2003) 참조.

18) 자세한 분석 내용은 조준형(2012) 참조.

‘공기빈도특이성’이라고 부르는 확률값은 각 어휘들의 빈도수와 함께 코퍼스 내에서 어휘들의 공기 관계, 조금 더 넓은 의미에서 연어 관계를 탐색하고 판단할 수 있는 주요 표지가 될 수 있다. 다시 말해서, 공기빈도특이성의 확률값이 높을수록, 해당 공기어는 다른 어휘들에 비해 키워드와의 공기 가능성이 높아지며, 따라서 연구자는 분석 코퍼스 내에서 이들 간의 분포적 연관성 및 연어의 가능성도 예상해 볼 수 있다.

번역 맥락 탐색의 관점에서는 문장과 같은 완벽한 맥락 텍스트를 참조하지 않더라도, 연구자는 원문과 번역문의 공기어들을 비교하면서 키워드가 되는 어휘의 번역 가능성을 일차적으로 판단할 수 있다. 공기어들이 곧 맥락을 구성하는 요소들이기 때문이다. 번역 코퍼스에서 비교하고자 하는 키워드를 중심으로 출현하는 공기어들의 분포 양상이 비슷하다면 키워드 간의 맥락적인 유사성이 높다고 할 수 있으며, 따라서 인간 연구자 혹은 컴퓨터는 번역 후보 어휘의 번역 대응성이 확률적으로 높다고 판단할 수 있다.

공기어 검색의 이점은 공기어를 다시 키워드로 해서 또 다른 공기어들을 검색하는 등, 다중적인 검색이 가능하다는 것이다. 이러한 복합적 다중검색을 통해서 병렬코퍼스 내에서 나타나는 어휘들 간의 번역 비교가 가능해진다.

4.2.2. 공기어 검색에서 맥락의 문제

공기어 검색에서도 검색 기준이 되는 맥락의 범주 설정이 중요한 문제가 될 수 있다. 앞서 키워드 검색에서 기준이 되는 맥락은 분석 목적에 따라서 ‘문장’ 혹은 ‘문장정렬맥락’, 두 가지가 될 수 있다고 하였는데, 공기어 검색에서도 동일한 태도를 취할 수 있다.

먼저, 맥락 기준으로 ‘문장’을 선택하면 한 문장 내에서 함께 나타나는 어휘들의 목록을 보여주기 때문에 ‘공기어’의 정의에 가장 부합하는 목록을 생성할 수 있다. 반면에 맥락 기준을 ‘문장정렬맥락’으로 설정하면 번역 차원에서 가장 적절한 공기어들을 살펴볼 수 있다. 그런데 후자의 경우는 어휘 연계성과 관련된 문제가 제기될 수 있다.

§#Un **client** entra, désoeuvré, seul, seul, et commanda également du **vin**.
§#한 **손님**이 한가한 듯 혼자서 들어섰다. #그러고는 **술**을 주문했다.

위에 제시된 예를 보면 원문에서 한 문장으로 구현된 맥락이 한국어 번역문에서는 두 개의 문장이 분리되어 있다. 이 경우에 원문에서 *client*과 *vin*은 하나의 문장 내에서 공기어 관계에 있다고 말할 수 있지만, 한국어 번역문에서 '손님'과 '술'은 과연 공기어 관계에 있다고 할 수가 있을 것인가? 맥락의 기준이 '문장정렬맥락'이라고 하면 두 한국어 어휘가 공기어 관계에 있다고 할 수 있지만, 문장이 맥락의 기준이라고 한다면 두 어휘는 공기어 관계에 있지 않다고 판단할 수밖에 없다.

이러한 문제에 해결하기 위해서는 어휘들 간의 거리, 다시 말해서 두 어휘가 출현하는 위치를 표상하는 통계 수치를 고려할 수 있다. 두 어휘의 거리는 일반적으로 두 어휘 사이에 나타나는 다른 어휘들의 수로 평가되는데, 이 수치가 높을수록, 다시 말해서 두 어휘 사이에 나타나는 어휘들의 수가 증가할수록 컴퓨터는 문제가 되는 어휘들의 의미적 연관성이 약화된다고 판단하게 된다. 그러나 해당 어휘가 멀리 떨어져 있다고 하더라도 의미적으로 완전히 무관한 어휘라고 단언하기는 어렵다. 어휘들 간의 조응 관계는 형식적인 텍스트 구성 단위를 넘어서 이루어진다. 앞서 언급한 것처럼 인간 연구자는 맥락을 넓은 범위에서 총합적으로 고려하기 때문에 서로 거리가 먼 어휘들이라도 의미적인 연관성이 있는지 없는지를 쉽게 판단할 수 있다. 반면에 컴퓨터는 주어진 맥락 기준에서만 어휘들 간의 관계를 평가하기 때문에, 위의 예에서와 같은 경우는 형식적으로도 의미적으로도 두 어휘가 공기 관계에 놓인다고 말하기는 어려울 것이다.

5. 결론

이미 존재하는 번역 텍스트를 기반으로 구축된 병렬코퍼스는 번역 연구에서 언어 자료를 관찰하기 위한 최적의 도구라고 할 수 있다. 더욱이 대용량 코퍼스를 통해서 많은 언어 자료를 효율적으로 관찰할 수 있다는 점에서 코퍼스

기반 연구는 매우 실용적인 방법 중의 하나라고 할 수 있다.

코퍼스 분석 도구와 분석 기법을 이용해서 연구자는 코퍼스로부터 다양한 번역 등가어 후보들을 찾아낼 수 있다. 그런데 코퍼스로부터 어휘 형태로 추출된 분석 결과물들은 빈도수 이외에는 번역에 대한 결정적인 증거를 제공하지 않으며, 번역 등가성을 빈도수만으로 평가하는 것은 의미가 배제된 형식적인 차원의 등가만을 고려하는 결과를 야기할 수 있다. 이러한 문제를 해결하기 위해서 코퍼스 기반 번역 연구에서도 맥락은 어휘들의 번역적 판단을 위한 가장 중요한 개념이라고 할 수 있다. 이를 위해서 언어자동처리 연구자들은 병렬코퍼스에서 맥락 탐색을 위한 다양한 기법을 제시하였다. 병렬코퍼스에서 맥락 탐색은 여러 가지 방식으로 이루어지지만 궁극적으로 코퍼스 내에서 해당 번역어의 분포와 그것과 함께 출현하는 어휘들의 분포를 살펴보는 것이 기본 개념이다. '키워드 검색'과 '공기어 검색'은 병렬코퍼스에서 번역 등가어를 검증하기 위해서 맥락을 탐색할 때 사용하는 가장 대표적인 방법이다. 이 기법들은 컴퓨터에 대한 전문적인 지식이 없더라도 기초적인 형태의 병렬코퍼스를 가지고 있다면, 엑셀과 같이 일반적인 프로그램 등을 이용해서 번역 연구자 누구나 충분히 활용 가능한 기법들이다. 이 기법들을 이용해서 인간 연구자는 원문과 번역문 사이에서 번역 등가어와 공기어들을 관찰하면서 번역어가 속한 번역 맥락을 확인할 수 있다.

한 가지 염두에 두어야 할 것은 이러한 기법들에 의해서 추출된 결과물은 통계 수치와 함께 번역 판단을 위한 최소한 근거들이지 그 자체가 번역 문제를 해결할 수 있는 결정적 증거는 되지 못한다는 점이다. 왜냐하면 앞서 살펴본 것처럼 기계에 의해 병렬코퍼스에서 추출된 맥락들은 하나의 완전한 단위이지만, 전체적인 관점에서는 그 자체가 다른 맥락 단위와의 연계성을 보여주지 못하는 단편적인 맥락들이며, 맥락의 기준도 매우 형식적인 수단에 의존하기 때문이다. 따라서 분석의 최종적인 판단은 인간 연구자의 몫이다. 코퍼스 기반 연구에 대한 비판적인 태도는 대부분 컴퓨터라는 도구를 사용함으로써 필연적으로 나타날 수밖에 없는 분석 방법의 지극히 형식적인 특징 때문에 비롯되는 것 같다. 그러나 대규모 코퍼스를 기반으로 실제 번역 용례를 빠른 시간 내에 탐색할 수 있다는 점은 분명히 이 연구의 장점이라고 할 수 있다.

현대 인문학 연구에서 컴퓨터 도구의 활용은 더 이상 무시하지 못할 새로

운 흐름이며, 이 도구에 의해서 인간 연구자는 양적으로도 질적으로도 기존의 자료 처리의 한계를 극복할 수 있게 되었다. 번역학 연구도 마찬가지라고 할 수 있다. 번역 연구에서 인간의 심층적이고 분석적인 사유는 매우 중요하다. 그러나 컴퓨터에 의한 코퍼스 연구의 장점을 충분히 활용할 수 있다면 이전보다 훨씬 효율적이고 체계적인 번역 연구가 가능해질 것이다. 또한 코퍼스 기반 번역 연구에서 나타나는 '형식'적인 결과물들도 인문학 연구자들과의 협력 작업을 통해서 '실질'적인 자료로 재탄생할 수 있을 것이다.

참고문헌

전성기 (2009) 「탐구번역론－하나의 인문학 번역론」, 『번역학연구』 10(2): 133-57.

정호정 (2003) 「코퍼스 중심의 번역학 연구」, 『번역학연구』 4(2): 71-88.

조준형 & 이영훈 (2012) 「『모데라토 칸타빌레(*Moderato Cantabile*)』의 번역코퍼스에 나타난 저빈도 명사 연구」, 『통역과 번역』 14(1): 237-71.

Baker, Mona (1995) 'Corpora in Translation Studies: An Overview and Some Suggestions for Future Research', *Target* 7(2): 223-43.

Ballard, Michel (1992) 'Concepts méthodologiques pour la mesure de l'équivalence', *Turjuman* 1(2): 17-30.

--- (1993a) 'Concepts méthodologiques pour la mesure de l'équivalence', *Turjuman* 2(2): 7-22.

--- (1993b) 'Unité de traduction. Essai de redéfinition d'un concept', *La traduction à l'université: Recherches et propositions didactiques*, Études réunies par Michel Ballard, Presses universitaires de Lille: 223-62.

--- (2006) 'La traductologie, science d'observation', *Qu'est-ce que la traductologie?* Études réunies par M. Ballard, Artois Presses Université: 179-94.

--- (2007) 'Pour un rééquilibrage épistémologique en traductologie', *Quo vadis Translatologie?*, Frank & Timme: 17-34.

Brown, Peter. F., Jennifer. C. Lai & Robert L. Mercer (1991) 'Aligning Sentences in Parallel Corpora', *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*: 169-76.

Cho, Joon-Hyung (2010) *Analyse textométrique des corpus parallèles français-coréens*, Thèse de doctorat, Université Paris 3.

Dubois, Jean, Mattée Giacomo, Louis Guespin, Christiane Marcellesi, Jean-Baptiste Marcellesi & Jean-Pierre Mével (2001) *Dictionnaire de linguistique*, Paris: Larousse.

Fuchs, Catherine (1993) *Linguistique et traitements automatiques des langues*, Paris: Hachette, « HU. Linguistique ».

Gale, William A. & Kenneth W. Church (1993) 'A program for aligning sentences in bilingual corpora', *Computational Linguistics* 19(1): 75-102.

Holmes, James S. (1972) 'The Name and Nature of Translation Studies', *Translated!: Papers on Literay Translation and Translation Studies*, Amsterdam: Rodopi: 67-80.

Isabelle, Pierre & Susan Warwick-Armstrong (1993) 'Les corpus bilingues: une nouvelle ressource pour le traducteur', Pierrette Bouillon et André Clas (Dirs.), *La Traductique: études et recherches de traduction par ordinateur*, Les Presses de l'Université de Montréal: 288-306.

Kay, Martin & Martin Röscheisen (1993) 'Text-Translation Alignment', *Computational Linguistics* 19(1): 121-42.

Lafon, Pierre (1981) 'Analyse lexicométrique et recherché des cooccurrences', *Mots* 3: 95-148.

--- (1984) *Dépouillements et statistiques en lexicométrie*, Paris-Gnève: Slatkine-Champion.

Lebart, Ludovic & André Salem (1994) *Statistique Textuelle*, Paris: Dunod.

Loffler-Laurian, Anne-Marie (1996) *La traduction automatique*, Paris : Presses Universitaires du Septentrion.

Martinez, William (2000) 'Mise en évidence de rapports synonymiques par la méthode des cooccurrences', *Actes des 5es Journées Internationales*

d'Analyse statistique des Données Textuelles: 197-203.

Martinez, William (2003) *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*, Thèse de doctorat, Université Paris III.

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

Vinay, Jean-Paul & Jean Darbelnet (1977) *Stylistique comparée du français et de l'anglais*, Paris: Didier.

[Abstract]

The Significance and Limitations of Context Exploration in Parallel Corpus

Cho, Joon-Hyung
(Korea University)

A parallel corpus contains numerous practical examples concerning the translation; it consists of the translation texts produced by a linguistic community. The main goal of corpus-based translation studies is to extract translation equivalences from a parallel corpus. Since the 1980s, computer-assisted tools have been utilized to accomplish this goal. However, the results obtained by these tools fail to validate translation relationships because they ignore the context in which the original text was used. To overcome this limitation, researchers have to return to the context. The *Keyword in Context* (KWIC) and the *co-occurrence analysis* are the most efficient methods for determining meaning in the domain of the corpus-based studies. However, to define the context for exploration, these two methods rely mainly on formal criteria such as the sentence or the translation alignment, which corresponds to a sentence-to-sentence correspondence between the original text and its translation. Such a formal definition of context can sometimes obscure the translation results because computers are unable to recognize the human cognitive processing. Therefore, a formal study of corpus-based translation studies must consider these inevitable problems and consider more various conditions of inquiry.

▸ Key Words: translation studies, parallel corpus, context exploration, KWIC, co-occurrence

조준형
고려대학교 불어불문학과 강사
chojh4net@naver.com
관심분야: 코퍼스 번역학, 번역평가

논문투고일: 2012년 10월 31일
심사완료일: 2012년 11월 26일
게재확정일: 2012년 12월 14일