

불한 번역 코퍼스에서 반복구문 분석에 의한 번역 대응 연구

조 준 형

(고려대)

1. 서론

코퍼스 기반 번역 연구에서 번역 등가어를 추출하거나 번역 텍스트의 특징을 개괄하기 위해 코퍼스 내의 어휘들을 관찰할 때 가장 많이 사용하는 기법이 키워드 검색(Keyword in Context)이다. 키워드 검색은 코퍼스 연구에서 가장 기본이 되는 어휘 탐색 방법이며, 그 결과물인 어휘맥락색인(concordance)은 키워드를 중심으로 어휘들의 다양한 관계망을 표상하는 목록을 제공하기 때문에 번역 연구에서 번역어의 맥락 상황을 이해하는 데에 필요한 기초자료가 될 수 있다.¹⁾ 어휘맥락색인은 키워드가 속한 맥락을 직접적으로 관찰할 수 있다는 점에

1) 어휘맥락색인은 코퍼스의 초기 연구부터 사용된 매우 보편적인 텍스트 분석 방식으로 싱클레어(Sinclair 1991, 2003), 유석훈 역(1999: 86-145) 등은 어휘, 구문, 연어 등 다양한 차원에서 어휘맥락색인을 활용한 어휘 관찰 예를 보여준다.

서 언어 연구뿐만 아니라 언어 교육에서도 효과적으로 사용할 수 있다.)²⁾

분석 주제에 따라서는 키워드로 단일어뿐만 아니라 두 개 이상의 단어로 구성된 어휘 그룹을 고려할 수도 있는데, 만일 연구 주제가 후자의 경우라면 해당 구문을 직접적으로 관찰하는 것이 분석에 보다 유리하다고 할 수 있다. 그럼에도 불구하고 코퍼스에서 번역 현상 관찰은 주로 단일어 검색에 의존하고 있는데, 이것은 한국어를 중심으로 한 번역 연구에서 한국어가 다른 서양어들과 언어적 특징이 다르기 때문에 전문 분석 도구에 의한 구문 탐색이 쉽지 않기 때문이다.

그런데 비교 차원의 연구에서 한국어와 타 언어 사이에서 번역적으로 대응 가능성이 있는 어휘들이 실제 텍스트에서 보이는 어휘적·문법적 양태의 차이 라든지 언어, 관용구, 복합어 등과 같은 어휘 그룹의 비교 연구에서는 구문 탐색이 매우 효과적인 방법 중의 하나라고 할 수 있다. 코퍼스에서 구문 탐색은 전혀 새로운 방법이 아니라 어휘맥락색인을 바탕으로 하는 전통적인 코퍼스 연구 기법이라고 할 수 있다(싱클레어 2003: 81-98). 단지 기존의 어휘맥락색인은 단순 목록에 지나지 않기 때문에 키워드를 중심으로 이루어지는 어휘들의 결합 양상을 체계적으로 보여주지 못하기 때문에 연구자가 이를 일목요연하게 관찰하는 것이 쉽지가 않다. 어휘맥락색인의 이러한 문제를 해결하기 위해 맥락 목록을 명시적으로 재구성할 필요가 있는데, 이를 위해서 가장 많이 사용되는 방법이 본 연구의 주제가 되는 반복구문(*segments répétés*) 분석이다. 반복구문 분석은 어휘맥락색인에 의해 제시된 목록에서 규칙적인 어휘 조합을 찾아내고 이를 일정한 기준에 따라서 재분류하고 빈도수와 함께 반복구문 목록을 생성한다. 이 목록을 가지고 코퍼스에서 규칙성을 가지는 어휘 그룹을 관찰하고 언어 간의 대조 연구가 가능해진다. 그렇지만 다른 코퍼스 분석 기법과 마찬가지로 반복구문 분석도 한국어와 프랑스어처럼 언어 유형이 전혀 다른 언어를 대상으로 했을 때는 여러 가지 한계가 있을 수 있음을 간과해서는 안 된다.)³⁾

2) 코퍼스 기반 외국어 교육에서 어휘맥락색인이 많이 활용되는 이유는 대부분의 코퍼스 분석 도구들에 기본적으로 포함되어 교습자가 특별한 사전 조작 없이 이 방법을 통해서 필요한 어휘들의 특징을 쉽게 학습자에게 제공할 수 있기 때문이다. 현재 언어 교육에서 어휘맥락색인은 워렌(Warren 2012)과 찰스(Charles 2012)가 보여주듯이 어휘 차원을 넘어서 구문 탐색과 담화분석 등 매우 다양한 범주에서 활용되고 있다.

본 연구는 전문 분석 도구를 활용해서 번역 코퍼스로부터 규칙적으로 반복되는 일정한 길이의 구문들을 추출하여 이를 비교 분석할 수 있는 엔그램(N-gram) 분석과 반복구문확장(expansion réccurente) 분석, 두 가지 기법을 살펴보고 이 기법들의 효용성과 한계를 점검하고자 한다. 이를 위해서 본 연구는 한편으로는 영어·프랑스어·한국어 세 가지 언어로 작성된 인권 관련 협약서로 구성된 코퍼스를 다른 한편으로는 프랑스어·한국어로 된 문학 소설 코퍼스를 대상으로 반복구문을 비교 분석하면서 언어 및 텍스트 장르에 따른 반복구문 분석의 결과를 평가하면서 반복구문 분석의 유용성을 살펴볼 것이다.

현재 국내에서 코퍼스 기반 연구는 대부분 영어-한국어를 대상으로 이루어지고 있으며, 국내 프랑스어 관련 연구에서는 컴퓨터를 활용한 코퍼스 기반 연구가 소수 연구자에 국한되어 아직은 저변이 탄탄하지 않다고 할 수 있다. 이러한 관점에서 본 연구가 프랑스어 관련 분야에서 정체되어 있는 코퍼스 연구의 흐름을 가속시킬 수 있는 계기가 되었으면 한다.

2. 코퍼스와 언어 자료의 탐색

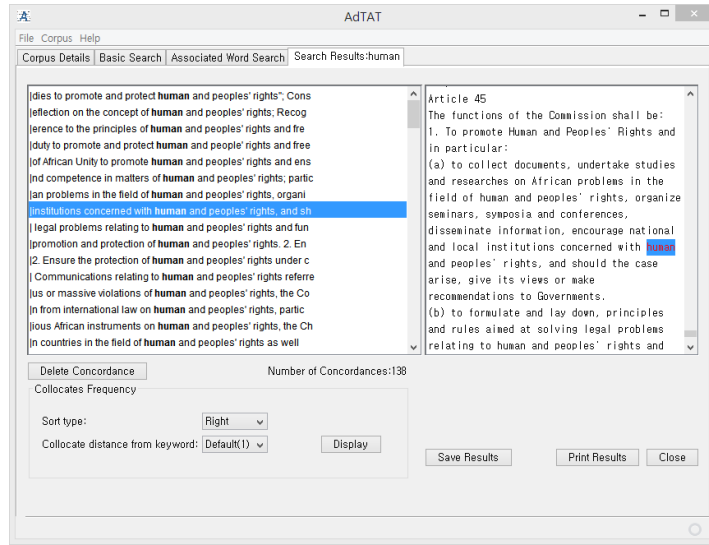
2.1. 키워드 검색과 어휘맥락색인

코퍼스 기반 연구의 목적은 코퍼스 내에 포함된 언어 자료를 다양한 기법들을 활용해서 심층적이고 체계적으로 관찰하고, 이를 통해서 언어 현상의 보편적인 개념을 설명하는 데에 있다. 코퍼스 내에서 언어 요소들을 체계적으로 관찰하기 위해서 특정 단어의 분포, 특정 단어가 속한 맥락 탐색, 공기어 검색,

-
- 3) 각 언어마다 어휘의 결합 방식, 문장의 구성 그리고 글쓰기 방식이 차별화되기 때문에 컴퓨터를 활용한 구문 연구에서는 이를 충분히 고려해야만 한다. 예를 들어 글쓰기 방식에서 중국어는 띄어쓰기가 없지만, 한국어와 프랑스어는 띄어쓰기가 존재하기 때문에 문장에서 단어들을 쉽게 구별할 수 있다. 그런데 한국어는 교착어의 특성 때문에 명사는 일반적으로 문장에서 조사와 같은 기능어들과 결합되어 존재하지만 프랑스어는 그렇지 않다. 현재 많이 사용되는 대부분의 분석도구들은 서양어를 대상으로 개발되었기 때문에 아시아언어들의 구문특징을 충분히 고려하지 못한다. 따라서 컴퓨터를 활용한 이개어 구문 비교는 어느 정도의 한계를 가질 수밖에 없다.

구문 탐색 등 여러 가지 기법들이 제안되었는데, 궁극적으로 이 모든 방법들의 출발점이 되는 것은 키워드 검색에 의한 어휘맥락색인이라고 할 수 있다.

〈그림 1〉 어휘맥락색인의 한 예



공기어 검색, 어휘 분포, 구문 검색도 결국엔 일정한 길이의 맥락을 기준으로 시행되는데, 어휘맥락색인이 바로 키워드를 포함하는 맥락을 일정한 기준으로 분류해놓은 목록이기 때문이다. 코퍼스 분야에서 주요 학자 중 한 사람인 싱클레어(Sinclair 1991)가 코퍼스에서 언어 자료를 관찰하기 위해 제시한 방법 중의 하나가 바로 어휘맥락색인이다. 싱클레어는 어휘맥락색인을 토대로 어휘, 문법, 연어 등 다양한 언어 현상을 계량적으로 분석하는 예를 보여주고 있다. 언어 교육에서 코퍼스의 활용을 기술하고 있는 베니트(Bennett 2010)에서도 어휘 맥락색인은 주요 분석 방법 중의 하나이다. 이처럼 어휘맥락색인이 코퍼스 연구에서 기본이 된 것은 이 목록에 의해 우리가 코퍼스에서 텍스트 단위들을 재구성하고 단어의 양태를 통합적으로 관찰할 수 있기 때문이다(Salem 1987: 41).

어휘맥락색인의 장점은 단일어 연구에서만 아니라 번역을 비롯해서 이질 언어 간의 비교 연구에서도 확인할 수 있다. 베이커(Baker 1995)가 번역학에서

코퍼스의 활용 예를 설명하면서 사용한 기본 방법도 키워드 검색에 의한 어휘 맥락색인의 비교였다. 이 경우는 단일어 어휘맥락색인이 아닌 병렬어휘맥락색인(Bi-concordance)에 해당한다. 병렬어휘맥락색인은 원문과 번역문을 병치하여 번역 등가어를 직접적으로 관찰할 수 있다는 점에서 번역 연구에서 매우 유용한 개념이다. 비교코퍼스 기반 연구에서도 분석 대상이 되는 언어 간의 차이를 비교하거나 번역 등가 표현이라고 할 수 있는 후보들을 관찰할 때도 (병렬)어휘 맥락색인은 흔히 사용되는 개념이다.⁴⁾

키워드 검색에 의한 번역 등가어 관찰에서 키워드로 선택되는 것은 일반적으로 단어인 경우가 대부분이다. 그러나 번역 연구에서 단어 자체만의 번역 대응 연구는 여러 가지 문제점이 있을 수 있다. 왜냐하면 언어에 따라서 사용되는 표현 방식이 다르고 문장에서 단어의 의미는 함께 출현하는 다른 단어들과의 상호 관계망 속에서 결정되기 때문이다.⁵⁾ 또한 한 언어에서 하나의 단어로 표상되는 개념이 다른 언어에서도 하나의 단어로 표현된다고 할 수는 없기 때문이다. 비네와 다르벨네(Vinay & Darbelent 1977)의 영어와 프랑스어의 비교 연구를 보면 *tout de suite/immediately, avoir lieu/to happen, simple soldier/private* 와 같이 동일한 개념이 영어와 프랑스어에서 서로 다른 어휘 구조로 실현되고 있음을 확인할 수 있다. 이영옥(2002)과 이승아(2012)를 비롯해서 실제 많은 번역 연구는 하나의 단어가 아닌 보다 상위의 언어 단위를 대상으로 하는 것이 일반적이다. 따라서 코퍼스 기반 연구에서도 단어 차원의 탐색이 아닌, 보다 상위의 단위인 구문 차원의 검색도 필요하다고 할 수 있다.

어휘맥락색인에서 다양한 어휘 요소들의 분석이 가능하다는 것은 *free hand* 와 같은 언어(collocation)의 관찰(싱클레어 2003: 23-25) 그리고 *as...as*와 같은 문법 구조 관찰(싱클레어 2003: 105-115)처럼 기존의 여러 연구 결과물을 통해

4) 번역 연구에서 병렬어휘맥락색인의 활용에 관해서는 조준형(2010: 107-110), 조준형(2011), 지미나(Zimina 2004: 27-33) 참조.

5) 이러한 문제는 언어 특징이 완전히 다른 언어들 간의 비교에서 두드러지게 나타난다. 예를 들어, 불한 코퍼스 분석에서 빈도수가 유사한 한국어와 프랑스어 단어들이 실제 맥락에서는 번역 등가를 이루지 못하는 경우가 많다. 반면 빈도수 차이가 크지만 실제 병렬어휘색인맥락에서는 번역 등가를 이루는 경우도 있다. 이에 대한 것은 조준형(2011) 참조.

확인할 수 있다. 엄밀한 의미에서 어휘맥락색인 혹은 병렬어휘맥락색인을 사용하는 목적은 키워드 자체의 의미를 이해하기 위한 것이라기보다는 이 키워드가 속한 맥락을 통해 키워드의 의미가 결정되는 어휘·통사적 환경을 파악하는 데에 있다고 할 수 있다. 이러한 점에서 어휘맥락색인을 참조하면서 우리는 이미 단어가 아닌 구문을 관찰하는 것이라고 할 수 있다. 더욱이 보다 분석적인 연구를 위해서 기존의 기법들이 지속적으로 발전하고 있기 때문에 이를 잘 활용한다면 컴퓨터의 도움을 받아 코퍼스로부터 번역 연구에 필요한 다양한 언어 요소들을 포함하는 자료들을 쉽고 빠르게 추출할 수 있을 것이다.

2.2. 구문 검색

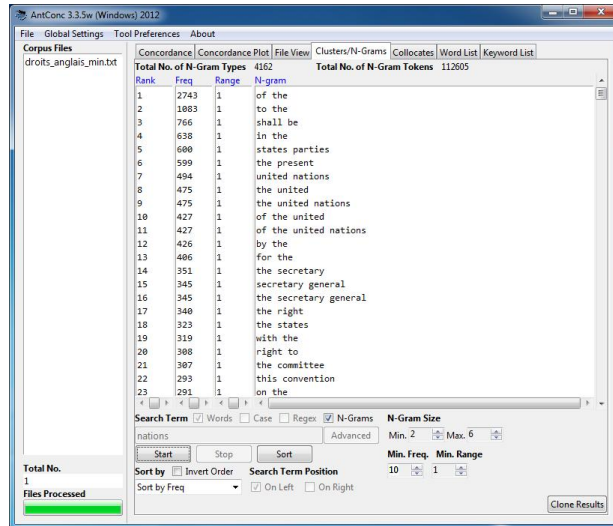
앞서 언급한 것처럼 코퍼스에서 언어 관계 혹은 복합어, 관용구와 같은 둘 이상의 어휘로 이루어진 어휘 그룹은 규칙적으로 반복해서 출현하는 구문을 일정한 기준에 따라서 선별하면서 관찰가능하다. 공기어 검색도 어휘 결합에 관한 연구를 위한 유용한 방법 중의 하나이지만, 공기어는 어휘들 간의 직접적인 연관성을 가리키는 것은 아니기 때문에 구문 분석을 위한 최적의 기법이라고는 할 수 없다.

코퍼스에서 반복되는 구문을 검색할 때 가장 많이 활용하는 방식이 엔그램(N-gram) 분석과 반복구문확장(expansion réccurente) 분석이라고 할 수 있다. 엔그램 분석은 n의 크기, 다시 말해서 구성 어휘 수에 따라서 코퍼스에서 통계적으로 빈번하게 발생하는 연속 단어 배열(이창수 2011: 321)을 빈도수와 함께 제시하면서, 규칙적으로 출현하는 반복구문을 관찰할 수 있도록 한다.⁶⁾ 다음은 코퍼스에서 n의 크기를 최소 2에서 최대 6으로 지정하여 추출한 반복구문의 한 예이다.⁷⁾

6) 이창수(2011: 321-322)에 따르면, 반복구문은 어휘묶음(lexical bundles), 다발(clusters), 어휘구(lexical phrase), 정형화된 표현(formulas), 규칙적 표현(routines), 미리 만들어진 표현(pre-abricated expressions) 등 다양한 이름으로 불린다. 본 논문에서는 코퍼스에서 규칙적으로 반복해서 나타나는 표현이라는 의미에서 반복구문이라는 용어를 사용할 것이다.

7) 본 논문에서 반복구문 탐색을 위해서 *AntConc*를 기본 분석 프로그램으로 사용할 것이다. 이 프로그램은 엔그램 분석뿐만 아니라 반복구문확장 분석을 위한 기능도 갖추고

<그림 2> 반복구문 추출의 한 예



<그림 2>에서 볼 수 있듯이 엔그램 분석은 특정 키워드를 대상으로 하지 않고 코퍼스에서 엔그램 목록, 다시 말해서 코퍼스에서 일정한 기준에 따라 규칙적으로 반복되는 모든 구문을 추출한다. 따라서 위 목록에서처럼 *states parties*, *united nations*, *secretary general*처럼 유의미적인 표현도 있지만, *of the*, *to the*, *shall be*처럼 문법 구조만을 확인할 수 있는 구문도 나타난다. 어쨌든 이러한 엔그램 분석의 결과물을 비교함으로써 번역 코퍼스 전체에서 번역 대응어들을 개괄적으로 탐색할 수 있다. 만일 두 목록 사이에서 구성 어휘 요소 중 하나가 의미적으로 대응된다면 이를 기반으로 어휘 및 문법 차원에서 두 언어의 대조 연구도 가능하다.

엔그램 분석이 코퍼스 전체에서 규칙적으로 반복해서 출현하는 어휘 결합을 추출하는 것이라면, 반복구문확장 분석은 특정 키워드를 중심으로 확장되는 어휘들의 양상을 살피는 것이다. 어휘 결합 양태를 관찰하기 위한 가장 쉬운 방

있다. 로렌스 앤토니(Laurence Anthony)에 의해 개발된 *AntConc*는 원시코퍼스뿐만 아니라 다국어 텍스트도 언어코드 충돌 없이 처리할 수 있기 때문에 유럽어로 작성된 텍스트를 분석할 때도 매우 유용한 도구라고 할 수 있다. 본 프로그램은 개발자의 홈페이지(<http://www.antlab.sci.waseda.ac.jp/software.html>)에서 무료로 받을 수 있다.

법이 바로 어휘맥락색인이지만, 어휘맥락색인은 <그림 1>이 보여주듯이 맥락의 개요만을 보여주며, 목록의 크기가 커질수록 이를 명확하게 해석하기가 난해해지기 때문에(살렘 1987: 135), 특징적인 어휘 분포 양상을 체계적으로 기술하기 위해서는 목록의 추가적인 조작이 필요하다. 이와 같은 관점에서 반복구문확장은 앵그램 분석의 원리를 이용해서 어휘맥락색인을 재구조화하는 방식이라고 할 수 있다. 반복구문확장은 어휘맥락색인에서 어휘 수를 매개변수로 해서 키워드를 중심으로 좌측 혹은 우측으로 전개되는 반복구문확장 분포목록(*Inventaire distributionnel des expansions récurrentes après/avant un pôle*, IDER)을 빈도수와 함께 제시한다.

<표 1> '권리'를 중심으로 좌측 반복구문확장 분포목록의 예

	가	권리	11
	가	권리	10
국적	가	권리	3
은	가	권리	2
모든 사람	은	가	2
은	국적	가	2
간섭 받지	은	가	2
아니하고	의견	가	2
	의견	가	2
	다음	과	11
	다음	과	7

<표 2> '권리'를 중심으로 우측 반복구문확장 분포목록의 예

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

구문확장목록은 앵그램 분석과 달리 특정 키워드를 중심으로 규칙적으로 반복되는 구문을 추출하기 때문에 특정 어휘들을 중심으로 직접적으로 번역 대응 관계를 관찰할 수 있고, 이와 더불어 어휘·통사 구조까지 비교가 가능하다. 예를 들어 '권리'의 영어역인 *right* 혹은 프랑스어역인 *droit*를 키워드로 해서 목록을 작성하면 코퍼스에서 '권리'/*right/droit* 자체의 번역 등가성뿐만 아니라 이 키워드를 포함한 번역 등가 표현들까지도 검증할 수 있다. 또한 반복구문을

단순히 철자순으로 나열하는 방식은 길이가 긴 구문이나 빈도수가 높은 구문을 명확하게 표시하기가 어렵기 때문에(살렘 187: 100), <표 1>과 <표 2>처럼 구문을 길이와 빈도수에 따라서 계층적으로 배열함으로써 구문의 구조를 보다 명확하게 파악할 수 있도록 한다.⁸⁾

엔그램 분석과 반복구문확장 분석 두 기법 중 어떤 방법이 더 효과적인지 단정할 수는 없다. 두 기법의 목적이 다르기 때문이다. 전자가 코퍼스 전체에서 규칙성을 가지는 모든 구문을 추출해내는 것이라면, 후자는 코퍼스에서 특정 키워드를 포함하는 반복구문을 선별하는 것을 목적으로 한다. 이창수(2011)와 이영옥(2002)처럼 분석이 총체적인 관점에서 이루어질 때는 엔그램 분석이 적절할 것이고, 김정우(2009)와 이승아(2012)처럼 분석 대상이 상당히 구체적인 경우에는 반복구문확장 분석이 조금 더 효과적이라고 할 수 있다.

물론 이러한 기법들이 모든 언어에서 만족할만한 결과를 도출한다고는 할 수 없다. 프랑스어와 영어처럼 유사한 언어적 특징을 가지는 언어들을 분석 대상으로 할 때는 구조적으로도 의미적으로도 상당히 유사한 구문 대응을 기대할 수 있지만, 한국어와 영어 혹은 프랑스어처럼 완전히 다른 언어 특징을 가지는 언어들 간의 비교에서는 상대적으로 그 결과가 기대에 미치지 못할 수도 있다. 현재 번역 코퍼스를 대상으로 한 엔그램 혹은 반복구문확장 분석의 예는 찾아볼 수 없기 때문에 이에 대한 답을 당장 제시할 수는 없다.⁹⁾ 더욱이 국내의 경우는 이창수(2011: 232)가 언급하듯이 코퍼스 기반 번역 연구에서조차도 반복구문에 대한 문헌은 찾기가 어렵기 때문에, 번역 연구에서 엔그램 분석과 반복구문확장 분석이 어느 정도 타당성이 있는지는 입증되지 않았다.

프랑스어가 포함된 병렬코퍼스를 대상으로 엔그램과 반복구문확장 분석을 활용한 구문 분석은 지미나(2004: 92-109)와 조준형(2010: 247-291)이 거의 유일한 예라고 할 수 있다. 지미나(2004)는 프랑스어와 영어로 구성된 병렬코퍼스를 대상으로 반복구문확장 방법의 유용성을 검증한 이 연구는 구문 구성 어휘들을 비롯해서 빈도수 차원에서도 두 언어가 매우 유사한 번역적 특징을 보이고 있음을 예증하고 있다. 조준형(2010)은 한국어와 프랑스어 병렬코퍼스¹⁰⁾를

8) 반복구문확장 목록의 배열 방식에 대해서는 살렘(1987: 93-105) 참조.

9) 살렘(1987)의 연구도 분석 대상은 프랑스어 단일어 코퍼스이지 다국어 코퍼스가 아니다.

10) 이 연구에서 사용된 불한병렬코퍼스는 본 논문에서 사용된 인권코퍼스와 동일하지

대상으로 하였으며 비록 지미나의 결과에는 미치지 못하지만 어느 정도 유의미한 결과를 이끌어 낼 수 있었다. 그러나 이러한 범주의 연구가 매우 희소하기 때문에 구문 분석 방법이 병렬코퍼스 기반 번역 연구에서 유용한 방법이 될 수 있다고 단언하기는 어렵다. 국내에서 찾을 수 있는 유일한 연구인 이창수(2011)도 엄밀한 의미에서 비교코퍼스(한국어 번역문과 한국어 비번역문)를 대상으로 한 것이기 때문이다.

구문 분석을 비롯해서 문법 연구에도 응용할 수 있는 이러한 연구의 일반화를 위해서 다양한 코퍼스를 대상으로 한 검증이 필요할 것이다. 이를 위해서 본 논문에서는 프랑스어와 한국어를 중심으로 한 두 가지 종류의 코퍼스를 통해서 앵그램과 구문확장분석을 차례로 적용하면서 유의미한 결과를 이끌어낼 수 있는지 실험해 볼 것이다. 영어를 포함시킨 것은 앞서 언급한 것처럼 언어의 유형론적인 특징에 따라서 결과가 어떤 식으로 달라지는지 확인하고, 현재 국내 코퍼스 기반 연구에서 예를 찾기가 어려운 세 언어 이상의 다국어 연구의 가능성을 확인하기 위함이다.

3. 반복구문에 의한 번역 표현 탐색

3.1. 분석 코퍼스

본 연구에서 반복구문 분석을 검증하기 위해서 두 가지 종류의 코퍼스를 사용할 것이다. 하나는 유럽 의회의 인권 관련 협약서를 바탕으로 하는 코퍼스로 영어와 프랑스어 그리고 한국어 세 가지 언어로 작성된 번역 코퍼스이다. 다른 하나는 문학 코퍼스로 프랑스어와 한국어로 작성된 문학 텍스트로 구성된다. 인권 코퍼스에서 영어 텍스트를 포함시킨 것은 영어/프랑스어를 대상으로 한 분석과 프랑스어/한국어를 대상으로 한 분석에서 어떤 차이가 있는지 확인하기 위해서이다. 문학 코퍼스는 총 58개의 문학작품을 포함하고 있는데, 프랑스어 소설과 한국어 번역문 그리고 한국어 소설과 프랑스 번역문이 혼합되어 있다.

만, 영어를 제외한 프랑스어와 한국어만을 분석 대상으로 하였으며, 반복구문확장을 주요 분석 방법으로 하였다.

이는 한국어와 프랑스어에서 번역 방향에 따른 번역의 차이를 간접적으로 확인하기 위해서이다. 마지막으로 유럽 의회에서 발간되는 모든 문서는 회원국의 언어와 영어가 공식 언어로 지정되어 있기 때문에, 프랑스어와 영어 텍스트는 각각이 하나의 정본이자 번역 텍스트라고 할 수 있다. 두 코퍼스의 계량적 특징은 다음과 같다.

〈표 3〉 인권 코퍼스의 통계 정보

	어절수	어형수	단발어수	최대빈도수	최대빈도어 ¹¹⁾
영어	112,199	4,073	1,227	10,399	the
프랑스어	121,758	5,197	1,651	7,004	de
한국어 형태소	104,968	4,316	1,537	4,937	의
한국어 원문 ¹²⁾	64,422	12,206	6,575	1,243	이

〈표 4〉 문학 코퍼스 통계 정보

	어절수	어형수	단발어수	최대빈도수	최대빈도어
프랑스어	418,995	24,613	10,975	16,280	de
한국어 원문	287,533	70,427	46,179	3,739	이
한국어 형태소	524,386	33,084	17,424	18,277	그

- 11) 형태소 분석된 한국어 텍스트와 그렇지 않은 한국어 텍스트에서 최대빈도어가 다른 것은 형태소 분석에 의해 명사와 결합된 (보)조사들이 분리되면서 교빈도어를 구성하기 때문이다. 예를 들어 형태소 분석 이전에는 ‘인권의’가 하나의 어형으로 분류되지만, 형태소 분석 이후에는 ‘인권’+‘의’로 분리되면서 ‘의’가 독립된 어형으로 고려된다.
- 12) 한국어 텍스트는 비교를 위해서 명사와 기능어들이 분리된 형태소 분석 텍스트도 활용하였다. 명사와 같은 실사의 정확한 통계 정보를 확인하기 위해서는 형태소 분석이 필요할 수도 있으며, 구문 분석에서도 ‘명사+조사’ 형태의 구문을 위해서는 원문보다는 형태소 분석된 텍스트를 관찰하는 것이 이점이 있다고 할 수 있다. 원문 혹은 형태소 분석 텍스트의 선택은 연구 목적에 따라서 달라질 수 있으며, 본고에서는 다양한 관점에서의 비교를 위해서 형태소 분석 텍스트도 분석 대상에 포함하였다.

3.2. 인권 코퍼스 분석

3.2.1. 엔그램에 의한 반복구문의 전체적인 특징

〈표 5〉 반복구문 상위 20개

영어		프랑스어		한국어	
빈도	구문	빈도	구문	빈도	구문
2,743	of the	1,379	de la	929	할 수
1,082	to the	839	à la	886	에 관한
766	shall be	596	les états	760	에 대한
638	in the	584	états parties	605	그 는
600	states parties	518	la présente	571	것 을
599	the present	482	nations unies	532	국제 연합
493	united nations	479	des nations	519	수 있다
474	the united	475	des nations unies	515	이 협약
474	the united nations	459	présente convention	477	권리 를
426	of the united nations	459	la présente convention	476	나 는
426	of the united	379	et de	455	에 의하여
424	by the	377	les états parties	445	당사국 은
405	for the	343	secrétaire général	436	사무 총장
340	the right	284	du présent	429	에 따라
323	the states	279	des droits	426	그 의
319	with the	274	la convention	403	협약 의
308	right to	259	ou de	400	것 은
303	the committee	253	tous les	358	하는 것
293	this convention	246	de leur	338	할 수 있다

위 표는 인권 코퍼스에서 빈도수가 가장 높은 상위 20개의 반복구문을 정리한 것이다. <표 5>에서 구문이 완전히 다른 한국어와 나머지 두 언어의 대비는 여러 가지 측면에서 무리가 있지만 구문에 포함된 명사형의 유형들을 보면 어느 정도 현재의 코퍼스에서 빈번하게 쓰이는 어휘 결합 형태를 유추해서 비교할 수는 있다.

먼저, 위 표에서 영어와 프랑스어 간의 구문의 유사성은 분명하게 확인할 수 있다. 비네와 다르벨네(1977)가 다양한 번역 층위에서 두 언어 간의 차이를 보여주듯이 영어와 프랑스어의 구문 구성이 완전히 동일하다고는 할 수 없다. 그럼에도 불구하고 <표 5>에서는 대부분의 구문이 구조적으로도 의미적으로도

유사한 양태를 보이고 있다. 물론 위 표에서 유사한 구문의 출현은 사용된 코퍼스가 원천 텍스트로 하는 문헌의 특징에서 그 원인을 찾을 수도 있다. 현재의 코퍼스는 실용 텍스트를 바탕으로 하며, 원천 텍스트 자체가 헌장, 협약서 등과 같이 매우 명시적인 텍스트이기 때문에 번역에서 어휘 변화의 가능성이 상당히 낮다고 할 수 있다. 더욱이 고유 명사, 용어들의 대응 관계는 확실적이기 때문에 문학 작품과 비교해서 좀 더 많은 유사 구문을 발견할 수 있다.

명사형을 포함하고 있지 않은 구문을 비교하더라도 두 언어 간의 유형적 유사성은 쉽게 확인된다. 예를 들어, 영어에서 가장 빈도가 높은 구문은 *of the*(F=2,739)로 ‘전치사+정관사’ 구조이다. 프랑스어에서 가장 빈도가 높은 구문은 *de la*(F=1,379)로 이 역시도 ‘전치사+정관사’ 구조이다. 두 번째로 빈도가 높은 구문도 각각 *to the*(F=1,082)와 *à la*(F=839)로 모두 ‘전치사+정관사’ 구조이다.¹³⁾ 더욱이 ‘전치사+정관사’ 구조라고 하더라도 사용된 전치사의 의미가 다르다면 대응이 잘못되었다고 할 수 있는데, 여기서는 *of/de* 모두 ‘소유(所有)’를 의미하는 수식 관계를 표현하는 전치사이며, *to/à* 역시도 맥락에 따라서 ‘방향(方向)’ 혹은 ‘소속(所屬)’을 의미하는 전치사로 동일하다고 할 수 있다.

<표 5>의 구문 목록에서 가장 많이 보이는 명사형은 *nations*이다. 이를 중심으로 한 어형의 결합 관계를 고려하면 더욱 흥미롭다.

13) 빈도수의 차이는 프랑스어가 가지는 언어적 특징 때문에 비롯된다고 할 수 있다. 프랑스어에서 관사, 형용사 등과 같이 명사를 수식하는 한정사는 모두 명사의 성(性)과 수(數)에 일치해야만 한다. 따라서 프랑스어 정관사는 성수에 따라서 *le, la, les* 세 가지 형태를 가지는데, 여성형 단수인 *la*를 제외하고, *le*와 *les*는 전치사 *de*와 *à* 바로 다음에 위치하면, 각각 *du, des/au, aux*로 축약이 된다. 이 축약 형태는 코퍼스 분석에서 하나의 어절로 인식되기 때문에 또 다른 어형과 함께 반복해서 출현하지 않는 이상 반복구문에서는 배제된다. <표 5>에서 *des nations unies*는 비록 세 가지 어형이 결합된 반복구문이지만 실질적으로는 *de+les+nations+unies*의 결합형이라고 할 수 있다.

〈표 6〉 Nations 중심으로 한 반복구문 비교

영어		프랑스어	
빈도	구문	빈도	구문
493	united nations	482	nations unies
474	the united	479	des nations
474	the united nations	475	des nations unies
426	of the united nations		
426	of the united		

<표 6>을 보면, 영어 *united nations*와 프랑스어 *nations unies*는 빈도수에서도 거의 일치하는 구문이며, 명사와 형용사의 위치만 다를 뿐 모두 한국어 ‘국제연합’에 해당하는 명칭이다. 구문 길이, 다시 말해서 어형이 확장되는 양태를 보면 두 번째와 세 번째 영어 구문을 제외하면 나머지 구문은 모두 일치하고 있다.¹⁴⁾

of+the+united : des(de+les)+nations¹⁵⁾

of+the+united+nations : des(de+les)+nations+unies

이러한 점을 고려한다면 영어와 프랑스어처럼 유사한 언어적 특징을 가진 언어들 간의 번역 비교에서 ‘반복구문’은 매우 중요한 분석 방법이 될 수 있다.

동일한 관점에서 한국어를 고려하면, <표 5>에서 영어 *united nations* 구문에 대응되는 구문으로 ‘국제 연합’(F=532)을 고려할 수 있다. 언어 차이를 고려한다면 빈도수도 매우 유사한 경우라고 할 수 있다. 이것은 영어와 프랑스어의 표현이 한국어에서 다른 수단에 의존하지 않고 직역이 되었음을 추측할 수 있다. 그런데 ‘국제 연합’이라는 구문을 제외하고는 <표 5>에서 이 구문을 구성하는 어형(‘국제’ 혹은 ‘연합’)이 포함된 다른 한국어 구문은 나타나지 않고 있다.

14) 프랑스어 텍스트에서 *the+united+nations* 조합에 해당할 수 있는 *les+nations+unies*는 14번 출현하는 저빈도 구문에 속한다. 이를 통해서 프랑스 텍스트에서 *nations unies* 구문은 14번의 경우를 제외하고는 모두 *de+les*의 축약형인 *des*와 공기한다고 추측할 수 있다.

15) 영어에서는 *united*가 *nations* 앞에 위치하기 때문에 궁극적으로 동일한 구문이라고 할 수 있다.

그런데 빈도수를 고려하면 영어/프랑스어와 한국어의 구문 구성 어형의 양태는 차이가 있음을 확인할 수 있다. 다시 말해서, 영어에서 *nations/united*, 프랑스어에서 *nations/unies* 그리고 한국어에서 ‘국제’/‘연합’의 관계는 다를 수 있다. 어형들 자체만을 고려한다면, 번역 차원에서 한국어 ‘연합’은 영어 *united*와 프랑스어 *unies*에 대응된다고 생각할 수 있지만, ‘국제’는 *nations*에 직접적으로 대응된다고 말할 수는 없다. 영어든 프랑스어이든 *nations*는 한국어에서 ‘국가’ 혹은 ‘국민’으로 번역될 수 있기 때문이다. 만일 *United Nations*를 사전적인 의미로 직역을 한다면 ‘국가 연합’ 혹은 ‘연합 국가’가 될 것이다. 일반적인 상황이라면 한국어 ‘국제’는 *nations*보다는 *international*이라는 형용사를 먼저 상기시킨다. 세 언어에서 각 구문의 구성 어형들의 빈도수를 비교하면 다음과 같다.

영어	: <i>nations</i> (F=506), <i>united</i> (F=494)
프랑스어	: <i>nations</i> (F=497), <i>unies</i> (F=483)
한국어	: 국제(F=770), 연합(F=534)

위의 결과에서 보듯이 영어와 프랑스어에 비해서 한국어는 ‘국제’와 ‘연합’의 빈도수가 상대적으로 많은 차이를 보이고 있다. 이것은 두 어형 간의 관계에서 영어와 프랑스어는 현재의 코퍼스에서 거의 하나의 의미 단위를 이루는 반면, 한국어에서는 ‘국제’가 좀 더 자유롭게 사용되고 있음을 알 수 있다. 영어 형용사 *international*(F=258)과 프랑스어 형용사 *international*+(F=259)¹⁶⁾의 빈도수를 고려하면, 한국어 ‘국제’는 결국 *nations*과 *international*을 포괄하는 번역어로 사용되고 있음을 알 수 있다. 만일 반복구문 분석에 의존하지 않고 어형들을 개별적으로 관찰한다면 이와 같은 차이를 확인하기 어려울 수도 있다.

<표 5>에서 기능어들로만 구성된 반복구문을 제외하면, 영어 텍스트에서 방금 살펴본 *united nations* 이외에도, *states parties*(F=600), *secretary general*(F=345), *the committee*(F=307), *this convention*(F=293)을 고려할 수 있다. 현재의 표에서 프랑스어 대응어들은 각각 *nations (unies)*를 비롯해서 *états*

16) 프랑스어 형용사 *international*은 수식하는 명사의 성과 수에 따라서 *international* (F=86), *internationale*(F=91), *internationaux*(F=41), *internationales*(F=41), 네 가지 형태로 존재한다.

parties(F=596), *secrétaire général*(F=343) 그리고 목록에는 보이지 않지만 *la commission*(F=216), *la convention*(F=274)을 고려할 수 있다. 반면 한국어에서는 ‘국제 연합’과 더불어, ‘사무 총장’(F=394), ‘이 협약’(F=452)과 같이 번역 차원에서 대응되는 형태를 찾을 수 있지만, 나머지 *states parties*, *the committee*에 대응되는 한국어 번역어는 해당 개념의 표현 방식의 차이로 인해서 한국어 반복구문 목록에서는 나타나지 않는다. 예를 들어 *states parties*는 영어 혹은 프랑스어에서 두 개 어형의 결합으로 나타나지만, 한국어에서는 ‘당사국’이라는 하나의 어형으로 출현하기 때문이다. 두 개 이상 어형의 결합을 의미하는 반복구문에서는 단일어인 ‘당사국’ 자체는 배제가 되며, ‘당사국 은’(F=402)처럼 다른 기능어들과의 결합 형태로만 나타난다(<표 5> 참조). 이것은 *committee*의 번역어인 ‘위원회’도 마찬가지이다.¹⁷⁾

3.2.2. 반복구문확장 목록 비교

인권 코퍼스에서 반복구문확장 분석의 검증을 위해서 *convention* / ‘협약’을 중심으로 한 구문들의 번역 양상을 살펴보기로 하자.¹⁸⁾ 추출 자료의 신뢰성을 위해서 최소 빈도수는 12로 지정하였다.

<표 7> 영어 *convention*의 반복구문확장

좌확장		우확장	
구문	빈도	빈도	구문
this convention	293	150	convention shall
the convention	248	57	convention on
present convention	230	52	convention shall be
the present convention	230	38	convention on the
of the convention	112	35	convention and
of this convention	112	30	convention may
of the present convention	80	25	convention to

17) 프랑스어 혹은 영어의 복합어가 한국어에서는 한자의 존재 때문에 단일어로 표현되는 경우가 많다. 이러한 이유로 엔그램 분석에서는 서양어와 한국어 사이의 번역 대응 구문이 일치하지 않는 경우가 일반적이다.

18) *nations* 혹은 *unies*의 반복구문확장은 <표 5>에서도 간접적으로 확인할 수 있다.

to the convention	60	24	convention for
to the present convention	53	24	convention or
to this convention	48	24	convention , the
parties to the present convention	39	22	convention shall enter
in the present convention	37	22	convention shall enter into
parties to the convention	27	22	convention shall enter into force
parties to this convention	23	21	convention in
states parties to the present convention	23	20	convention shall be open
states parties to the convention	22	19	convention for the
in this convention	21	18	convention shall enter into force on
states parties to this convention	18	16	convention is
provisions of this convention	17	16	convention may be
american convention	16	16	convention on human
provisions of the convention	16	16	convention on human rights
the provisions of this convention	16	15	convention shall not
application of this convention	14	14	convention shall apply
party to this convention	14	14	convention undertake
the american convention	14	12	convention of
entry into force of the convention	13	12	convention on the rights
entry into force of this convention	13	12	convention on the rights of
force of the convention	13	12	convention on the rights of the
force of this convention	13	12	convention shall be open for
into force of the convention	13		
into force of this convention	13		
provisions of the present convention	13		
the parties to the present convention	13		
the states parties to this convention	13		
in the convention	12		
international convention	12		
the provisions of the convention	12		

〈표 8〉 프랑스어 *convention*의 반복구문확장

좌확장		우확장	
구문	빈도	빈도	구문
la présente convention	459	38	convention et
présente convention	459	38	convention s
la convention	274	34	convention entrera
de la présente convention	185	34	convention entrera en
de la convention	126	34	convention entrera en vigueur

à la présente convention	97	34	convention ne
cette convention	66	29	convention ou
parties à la présente convention	65	25	convention sera
à la convention	64	21	convention à
états parties à la présente convention	45	20	convention entrera en vigueur le
de cette convention	39	18	convention pour
dans la présente convention	34	18	convention qui
dispositions de la présente convention	33	17	convention de
parties à la convention	32	17	convention est
états parties à la convention	28	15	convention concernant
par la présente convention	25	15	convention relative
en vigueur de la convention	19	15	convention , les
entrée en vigueur de la convention	19	14	convention américaine
partie à la présente convention	19	14	convention américaine relative
vigueur de la convention	19	14	convention américaine relative aux
article 2 de la présente convention	18	14	convention américaine relative aux droits
dispositions de la convention	18	14	convention américaine relative aux droits de
l'article 2 de la présente convention	18	14	convention s'engagent
les dispositions de la présente convention	17	13	convention ou y
application de la présente convention	15	13	convention sera ouverte
ladite convention	15	12	convention relative aux
l'application de la présente convention	14	12	convention relative aux droits
application de la convention	13	12	convention relative aux droits de
aux fins de la présente convention	13	12	convention relative aux droits de l
fins de la présente convention	13	12	convention sera ouverte à
des dispositions de la présente convention	12	12	convention , le
l'application de la convention	12		
partie à la convention	12		
reconnus dans la présente convention	12		
état partie à la présente convention	12		

〈표 9〉 한국어 ‘협약’의 반복구문확장

좌확장		우확장	
구문	빈도	빈도	구문
이 협약	452	304	협약 의
본 협약	141	109	협약 은
은 이 협약	50	105	협약 에
관한 협약	41	93	협약 제
에 관한 협약	41	45	협약 이
국제 협약	24	40	협약 의 당사국
는 협약	24	38	협약 을
은 협약	22	30	협약 의 규정
는 이 협약	16	28	협약 당사국
이 이 협약	16	27	협약 에 규정
인권 협약	16	26	협약 의 적용
미주 인권 협약	15	22	협약 에 규정 된
당사국 은 이 협약	14	22	협약 의 당사국 은
은 본 협약	13	21	협약 의 제
권리 에 관한 협약	11	20	협약 의 발효
의정서 는 협약	11	19	협약 에서
이 의정서 는 협약	11	17	협약 제2조
아동 의 권리 에 관한 협약	10	16	협약 에 의하여
에 이 협약	10	16	협약 의 체결국
의 권리 에 관한 협약	10	14	협약 상 의
		14	협약 에 대한
		14	협약 은 국제
		14	협약 은 국제 연합
		14	협약 은 비준
		14	협약 의 해석
		13	협약 의 적용 을
		12	협약 규정
		12	협약 이 발효

*Convention*을 중심으로 어휘가 확장되는 양상을 보면, 프랑스어와 영어에서는 유사한 반복구문확장 구조를 확인할 수 있다. 문장 구조 및 어휘의 일치성은 곧 두 언어의 번역이 일대일 대응 관계를 보이고 있음을 보여준다. 예를 들어, 영어 *convention shall enter* 등으로 전개되는 확장은 프랑스어에서 *convention entrera*라는 동일한 어휘 및 시제 구조에 대응된다. 그 외에도 영어 *be* 동사와 프랑스어 *être* 동사의 대응(*is/est*)을 비롯해서 좌측 확장에서는 관사와 지시사

(*this/cette*) 등의 대응 관계를 확인할 수 있다. 흥미로운 점은 형용사와 연관된 것인데, 프랑스어에서 *convention américaine*는 우측 확장에 나타나지만, 여기에 대응되는 영어의 *american convention*은 좌측 확장에 위치한다. 이것은 두 언어 사이의 번역 대응 관계뿐만 아니라 언어적 차이도 예증한다.¹⁹⁾ 이러한 정보는 번역 교육에서 교수자가 학습자에게 번역어뿐만 아니라 비교 대상이 되는 언어의 통사적 차이도 동시에 이해시킬 수 있는 매우 유용한 자료가 된다.

한국어 ‘협약’의 반복구문확장 목록을 프랑스어와 영어의 목록과 비교를 하면, 비록 언어 특성이 다르긴 하지만 여러 가지 측면에서 다양한 관찰이 가능하다. 번역적인 차원에서는 앞서 엔그램 분석에 비해 상당히 많은 부분에서 번역 등가 표현을 선별해 낼 수 있다. 특히 한정사 혹은 형용사의 수식 구조에서 다음과 같이 번역적으로 대응될 수 있는 표현들이 발견된다.

this convention/cette convention/이 협약
american convention/convention américaine/미주 인권 협약
provisions of the convention/dispositions de la convention/협약의 규정

또한 언어 유형적인 차이도 관찰가능하다. 현재의 목록을 보면 지시사 ‘이’를 비롯한 한정사는 한국어에서도 명사 앞에 위치하기 때문에 좌측 확장으로 고려되면서 프랑스어/영어와 동일한 구조를 보인다. 그렇지만 통사적 구조(수식어 및 보어의 위치)의 차이로 인해서 한국어는 나머지 두 언어와는 다른 구성적 특징을 보일 수 있다. 몇 가지 예를 들면 다음과 같다.²⁰⁾

이 협약 : *this convention; cette convention*
 본 협약 : *the present convention; la présente convention*
 의 권리에 관한 협약 : *convention on the rights of; convention relative aux droits*

19) 프랑스어 형용사는 일반적으로 길이가 짧은 것을 제외하고는 명사 뒤에 위치하며, 특히 국적을 나타내는 형용사는 반드시 명사 뒤에 있어야 한다.

20) 만일 한국어와 상당히 많은 유형적인 특징을 공유하는 일본어 텍스트를 같은 방식으로 비교한다면, 프랑스어와 영어의 대응만큼이나 많은 구문 상의 번역 대응 관계를 확인할 수 있을 것이다.

이러한 예들이 축적되면 국내 번역가들뿐만 아니라, 외국인 번역가들이 모국어를 한국어로 번역하거나 혹은 그 반대의 경우 그리고 자동 번역에서 문장 구조 차이에 의해 나타날 수 있는 번역 오류를 보정할 수 있는 자료들이 될 수 있다.

이러한 관찰은 코퍼스에서 단순히 하나의 어휘를 비교하는 것과는 많은 차이가 있다. 번역 코퍼스에서 하나의 어형만을 고려할 경우, 번역적 판단은 이 어형의 빈도수가 일차적 기준이 될 수 있지만, 언어적 특징이 전혀 다른 언어들의 비교에서는 신뢰할 수 있는 증거가 될 수 없다. 더욱이 하나의 단어는 맥락에 따라서 여러 가지 의미로 실현될 수 있기 때문에 정량 자료만으로 번역적 판단은 내리는 것은 위험하다. 공기어 분석은 해당 키워드와 포함된 맥락 정보를 알려준다는 점에서 유용한 방법이 될 수 있지만, 이 역시도 특정 맥락에 함께 출현하는 어휘들을 가리키지 공기어와 키워드 간의 실질적인 관계를 직접적으로 설명하는 것은 아니다. 반면 구문 검색은 최소한 일정한 크기의 표현을 대상으로 하기 때문에 단어 검색의 다의적인 모호성에서는 벗어날 수 있다. 물론 <표 5>에서 보는 것처럼 특정한 키워드가 없는 앵그램 방식의 구문 검색은 기능어들만으로 구성되어 의미 차원에서는 무용한 구문들이 나올 수도 있지만, *convention* '협약'의 예처럼 특정 키워드를 중심으로 한 구문 탐색은 번역 차원에서도, 통사 차원에서도, 유용한 자료를 이끌어낼 수 있다.

3.3. 문학 코퍼스에서 반복구문 탐색

인권 관련 협약서로 구성된 인권 코퍼스는 번역이 비교적 정형화되어 있으며, 이 때문에 영어와 프랑스어 간의 비교뿐만 아니라 한국어와의 비교에서도 의미 및 통사 차원에서도 매우 유사한 표현들을 확인할 수 있었다. 보다 신뢰성 있는 검증을 위해서 문학 텍스트를 대상으로 동일한 검증을 해보기로 하자. 인권 코퍼스는 전체가 인권 관련 문서로 주제가 동일하기 때문에 동일한 표현의 반복성이 두드러지고 이것은 앵그램 분석을 통해서 확인할 수 있었다. 반면 문학 코퍼스는 다양한 작품들로 구성이 되어 있기 때문에 관용구를 제외하고는 표현의 반복성은 상당히 낮다고 할 수 있다. 문학 코퍼스는 인권 코퍼스와 달리 영어로 된 텍스트를 구하기가 어렵기 때문에 프랑스어와 한국어 텍스트만을 분

석 대상으로 한다.

3.3.1. 앵그램 분석

아래 표는 문학코퍼스에서 프랑스어 텍스트, 한국어 원문 그리고 형태소 분석된 한국어 텍스트 각각에서 빈도수가 높은 상위 20개의 반복구문을 분류한 것이다.

〈표 10〉 문학 코퍼스에서 반복구문

프랑스어		한국어 원문		한국어 형태소	
빈도	구문	빈도	구문	빈도	구문
225	il n y	47	알 수 없는	2249	하 았 다
218	n était pas	34	하고 내가 말했다	1331	있 었 다
214	est ce que	28	있는 것 같았다	1194	고 있 었
197	n est pas	27	다시 한 번	957	이 었 다
196	il y avait	21	잘 알고 있었다	808	고 있 었 다
195	qu il avait	20	아무 말도 하지	752	말 하 았
187	il y a	19	안 데바레드가 말했다	681	말 하 았 다
185	qu est ce	19	하고 그녀가 말했다	611	것 이 다
179	ce qu il	19	하고 안 데바레드가	587	지 았 았
162	ce n est	17	그 여자가 말했다	536	는 것 이
149	se mit à	17	하고 그 여자가	432	하 지 았
137	il n avait	17	한 번 더	427	았 았 다
134	tout à coup	16	어느 날 저녁	405	고 있 는
132	je ne sais	14	견딜 수 없는	397	하 고 있
126	ce n était	14	았을 수 없었다	373	지 았 았 다
122	n avait pas	14	알 수 없었다	363	어 있 었
121	qu est ce que	14	하고 그가 말했다	342	다 그 는
117	ce n est pas	13	것을 알고 있었다	315	었 다 그
116	comme s il	13	나는 더 이상	307	하 았 습니다
114	n y avait	13	말도 하지 았 았 다	304	르 것 이
111	d une voix	13	볼 수 있었다	297	없 었 다

앵그램에서 n, 다시 말해서 구문에 포함된 최소 어형 수는 3으로 하였다. 선별 기준을 2로 했을 때는 빈도수를 기준으로 상위 그룹을 차지하는 구문들은

기능어들로만 구성된 구문들만 추출되어서 최소 어형 수를 3으로 하였다. <표 10>에서 프랑스어의 경우는 기능어로 구성된 반복구문 목록으로도 볼 수 있지만, *il y a*(영어 *there is*), *se mit à*(영어 *start, begin*), *tout à coup*(영어 *suddenly*) 등처럼 그 자체로 유의미적인 그룹들을 발견할 수 있다. 한국어 원문에서는 ‘여자’, ‘테바레드’, ‘저녁’ 등 명사형이 포함된 구문들을 확인할 수 있다. 반면, 형태소 분석된 한국어 텍스트에서는 대부분이 서술어로 동사 혹은 형용사의 시제 구조가 표시된 구문들에 해당한다. 따라서 유의미적인 구문을 찾기가 어렵다.²¹⁾

그런데 <표 10>의 목록은 <표 5>와는 분명한 차이가 있다. 인권 코퍼스에서는 텍스트 전체가 인권이라는 하나의 주제와 연관되었고, 정형화된 표현이 많았기 때문에 엔그램 분석에서도 ‘국제 연합’, ‘사무 총장’과 같은 번역 대응 표현들을 발견할 수 있었지만, <표 10>에서는 이러한 유형의 표현을 전혀 발견할 수 없다. 비록 텍스트 전체가 문학에 속하지만 주제 및 작가 그리고 번역가가 서로 다르기 때문에 인권 코퍼스처럼 코퍼스 전체에 걸쳐 동일하게 반복되는 표현을 기대하기는 어렵기 때문이다. 이러한 경우 키워드를 지정하지 않은 엔그램 방식의 구문 검색은 적절하지 않다고 할 수 있다. 따라서 문학 코퍼스처럼 표현의 다양성을 가지는 경우, 그래서 번역에서도 다양한 가능성을 내포하는 경우는 엔그램보다는 특정 키워드를 기준으로 하는 반복구문확장 분석이 보다 유용하다고 할 수 있다.

3.3.2. 반복구문확장 분석

문학 코퍼스에서 특정 키워드를 중심으로 프랑스어와 한국어 사이의 반복구문확장을 비교하기 위해서 고빈도 명사와 복합어 두 가지를 선택하고자 한다. 물론 이 경우도 두 가지 사항을 염두해 두어야 한다. 먼저, 프랑스어와 한국어 간의 번역 대응 가능성이 있는 어형이라고 하더라도 현재의 반복구문확장은 해당 어형 하나만을 비교하는 것이 아니라 일정한 어형들의 집합인 구문을 대비

21) <표 10>은 고빈도 구문 일부만을 선별했지만, 만일 *n*의 크기를 조절하고 저빈도 구문까지 모두 고려한다면 유의미적인 구문들을 찾아낼 수 있을 것이다.

시키는 것이기 때문에 각 언어의 통사 구조 및 어휘 결합 방식을 고려하지 않을 수 없다. 이러한 관점에서 상대적으로 고빈도 구문을 찾기는 쉽지 않을 것이다. 다른 하나는 만일 코퍼스에 사용된 문학 작품이 장편 소설이라고 한다면 반복되는 유사한 표현이 출현할 가능성이 높다고 예상할 수도 있지만, 단편 소설 및 중편 소설로 이루어진 현재의 코퍼스에서는 부사구나 서술어 등을 제외하면 명사형이 포함된 구문이 규칙적으로 반복되기는 어렵다고 예측할 수 있다. 이러한 관점에서 구문의 빈도 기준을 3으로 하였으며, 이를 통해서 고빈도 구문이 아니라 오히려 저빈도 구문의 양태를 확인할 수 있는 이점도 동시에 가질 수 있다.²²⁾

<표 10>에서 프랑스어 구문은 대부분 인칭대명사, 대명사, 동사 등으로 구성된 구문들이며, 명사형은 *voix*가 있다. 프랑스어 *voix*는 코퍼스에서 433번 출현하는 고빈도어이다. 번역적인 차원에서 이 어형에 대응될 수 있는 한국어 어형은 ‘목소리’ 혹은 ‘소리’라고 할 수 있다. 그런데 이 두 어형은 각각 한국어 텍스트에서 각각 302번과 1008번 출현하는 고빈도어에 해당한다. 빈도수만을 고려하면 *voix*의 한국어 번역어는 ‘목소리’라고 간주할 수 있지만, ‘소리’도 맥락에 따라서 *voix*의 번역 대응어가 될 수 있다. 그러나 일차적인 의미에서 ‘목소리’는 프랑스어로 *voix*로 번역이 되지만, ‘소리’는 *son*, *bruit*, *voix* 등 맥락에 따라서 다양한 프랑스어에 대응될 수 있기 때문에 ‘소리’의 빈도수가 훨씬 높은 것은 당연하다고 할 수 있다. 여기서는 빈도수를 고려해서 한국어 텍스트에서 ‘목소리’만을 키워드로 반복구문확장을 조사하고 실제 맥락에서 대응이 어떻게 되는지 살펴보기로 한다.

22) 코퍼스 기반 연구에서 어휘 혹은 구문을 관찰할 때 주로 고빈도어들을 분석 대상으로 하는 경향이 있다. 이것은 관찰 결과의 일반화 및 대표성을 고려하기 때문이라고 할 수 있다. 그럼에도 불구하고 저빈도 요소들도 매우 중요할 수 있는데, 저빈도 요소들은 어떤 의미에서는 텍스트를 특징짓는 하나의 표지가 될 수도 있고, 번역의 관점에서도 최소한 분석 코퍼스에서 일반적이지 않은 언어 요소의 번역 대응어들을 관찰할 수 있다는 점에서 중요한 요소라고 할 수 있다. 저빈도 요소를 대상으로 한 분석에 대해서는 조준형과 이영훈(2012) 및 지미나(2002) 참조.

〈표 11〉 프랑스어 *voix*의 반복구문확장

좌확장		우확장	
구문	빈도	빈도	구문
d'une voix	110	37	voix de
la voix	77	14	voix basse
sa voix	52	12	voix était
une voix	40	9	voix basse,
à voix	29	9	voix du
de sa voix	15	9	voix plus
des voix	8	9	voix qui
cette voix	6	8	voix basse.
dit d'une voix	6	4	voix douce
et sa voix	6	4	voix si
les voix	6	4	voix tremblait
même voix	6	4	voix très
de voix	5	3	voix basse :
la même voix	5	3	voix basse et
avec une voix	4	3	voix basse...
de la voix	4	3	voix claire
parler à voix	4	3	voix d'anne
répondit d'une voix	4	3	voix d'un
avec la voix	3	3	voix de la
c'était une voix	3	3	voix de son
dit à voix	3	3	voix des
entendit la voix	3	3	voix forte
et la voix	3	3	voix haute,
et, d'une voix	3	3	voix triste
haute voix	3	3	voix à
le ton de sa voix	3		
lui dit d'une voix	3		
lui dit à voix	3		
parler d'une voix	3		
petite voix	3		
puis, d'une voix	3		
ton de sa voix	3		
à haute voix	3		
à la voix	3		
à parler d'une voix	3		

〈표 12〉 한국어 ‘목소리’ 반복구문확장

좌확장		우확장	
구문	빈도	빈도	구문
낮은 목소리로	11	18	목소리로 말했다.
그의 목소리는	8	7	목소리로 이렇게
떨리는 목소리로	5	6	목소리로 대답했다.
안느의 목소리가	5	5	목소리가 들렸다.
그 목소리는	4	5	목소리로 그는
분명한 목소리로	4	5	목소리로 물었다.
부드러운 목소리가	3	3	목소리가 갑자기
부드러운 목소리로	3	3	목소리가 나를
없는 목소리로	3	3	목소리가 다시
있는 목소리로	3	3	목소리가 대답했다.
		3	목소리로 나를
		3	목소리로 말했습니다.

〈표 13〉 형태소 분석 한국어 텍스트에서 ‘목소리’ 반복구문확장

좌확장		우확장	
구문	빈도	빈도	구문
의 목소리	55	121	목소리 로
ㄴ 목소리	51	70	목소리 가
는 목소리	33	32	목소리 를
하 ㄴ 목소리	32	26	목소리 로 말
은 목소리	27	24	목소리 는
그 목소리	14	18	목소리 로 말 하
그 의 목소리	5	17	목소리 로 말 하 았
그의 목소리	5	10	목소리 로 그
떨리 는 목소리	5	9	목소리 가 들리
부드럽 은 목소리	5	7	목소리 에
안느 의 목소리	5	6	목소리 로 대답
다 . 그의 목소리	4	6	목소리 로 이렇
분명 하 ㄴ 목소리	4	6	목소리 로 이렇 게
쉼 목소리	4	5	목소리 가 나
여자 의 목소리	4	5	목소리 도
그 여자 의 목소리	3	5	목소리 로 그 는
나지막 하 ㄴ 목소리	3	5	목소리 로 말 했다
다 . 그 목소리	3	5	목소리 이
롭 ㄴ 목소리	3	5	목소리 처럼
아버지 의 목소리	3	4	목소리 로 나

아주 낮은 목소리	3	4	목소리 로 대답 하
왔다 . 그의 목소리	3	4	목소리 로 대답 하 았
은은 하 나 목소리	3	4	목소리 로 이야기
		4	목소리 를 낮추
		4	목소리 를 듣
		4	목소리 와
		3	목소리 가 갑자기
		3	목소리 가 나 를
		3	목소리 가 다시
		3	목소리 가 대답
		3	목소리 가 대답 하
		3	목소리 가 대답 하 았
		3	목소리 가 들렸다
		3	목소리 가 들리 어
		3	목소리 가 들리 어 오
		3	목소리 로 나 를
		3	목소리 로 묻
		3	목소리 로 묻 었
		3	목소리 로 묻 었 다
		3	목소리 로 이야기 하
		3	목소리 였다
		3	목소리 의
		3	목소리 이 었

<표 12>는 형태소 분석이 되지 않은 한국어 텍스트를 대상으로 했기 때문에 만일 ‘목소리’를 키워드로 지정하면 어떤 결과도 나타나지 않는다. 한국어에서 명사형은 문장에서 조사와 결합하여 문장 성분을 구성하기 때문이다. 이러한 이유로 ‘목소리’가 포함된 어형을 키워드로 하였으며, 그 결과는 ‘목소리가’, ‘목소리를’ 등과 같이 명사+(보)조사 형태의 어형들이 포함된 구문이 추출되었다.

이 목록에서 번역적으로 대응될 수 있는 표현을 보면, 의미적으로도 빈도적으로도 *entendit la voix*(F=3)와 ‘목소리가 들렸다’(F=5)(<표 12>), ‘목소리 가 들리’(F=9)/‘목소리를 듣’(F=4)(<표 13>)를 우선적으로 고려할 수 있다.²³⁾ ‘목소리가 들렸다’에서 ‘들렸다’는 수동의 의미이기 때문에, ‘목소리를 듣’과 의미상

23) ‘목소리가 들렸다’와 ‘목소리 가 들리’는 형태소 분석의 유무에 따른 차이기 때문에 궁극적으로는 같은 표현이라고 할 수 있다.

동일하다고 할 수 있다. 실제 맥락을 통해 프랑스어와 한국어의 대응 관계를 살펴보면 다음과 같다.

- a) chactas cessa de parler, et l'on **entendit la voix** du flamant qui, retiré dans les roseaux du meschacebé, annonçait un orage pour le milieu du jour.
 삭타스가 말을 마치자 미시시피강의 갈대밭 숲 속 깊은 곳에서 한 낮에 비바람 치는 뇌우를 알리는 홍학 **소리가 들렸다**.
- b) loin derrière lui, il **entendit la voix** de l'autre petite fille qui reprenait de la vigueur, et celle de sa soeur qui au contraire faiblit tout d'un coup.
 그리고 등 뒤로 도로 빨라 가는 뒷집 계집애의 말소리와 급작스레 떠가는 누이의 **말소리를 들으면서도** [...].
- c) ce fut à ce moment-là qu'il **entendit la voix** de l'oiseau-mère.
 그때 어미 새의 **목소리가 날아왔다**.

위 맥락을 보면 프랑스어 *entendit la voix*는 각각 한국어에서 ‘소리가 들렸다’, ‘말소리를 듣다’, ‘목소리가 날아왔다’에 대응된다. 첫 번째 예문에서 ‘소리가 들렸다’는 소리의 주체가 인간이 아닌 새이기 때문에 ‘목소리’가 아닌 ‘소리’로 번역되었다는 점을 고려하면 ‘목소리가 들렸다’에 해당한다고 할 수 있다.²⁴⁾ 예문 b)와 c)는 원문이 한국어 소설이고 프랑스어가 번역문이라고 할 수 있다. 예문 b)는 ‘말소리를 듣다’가 *il entendit la voix*로 번역이 되어 있기 때문에, 앞서 예측한 대응 관계 범주에 포함된다. 반면 예문 c)에서 프랑스어는 *entendit la voix*라고 표현되어 있지만 한국어 원문이 ‘날아왔다’라는 문학적 표현으로 되어 있어서 매우 특수한 경우라고 할 수 있다. 어쨌든 이 경우에 프랑스어 표현만을 보면 전혀 의미상의 문제가 없으나 문학적인 차원에서는 충실한 번역은 아닌 것 같다.²⁵⁾

24) 보다 상세한 분석을 위해서는 한국어 텍스트에서 ‘소리’를 키워드로 해서 반복구문 확장을 조사해야 할 것이다.

25) 예문 c)는 매우 특수한 경우이기 때문에, 이를 자동번역을 위한 자료로 축적하는 것이 타당한 것인가에 대해서는 의문점이 든다. 어떤 의미에서는 번역 오류라고 볼 수도 있기 때문이다.

보다 명확한 분석을 위해 한국어 ‘목소리가 들렸다’/‘목소리 를 듣’를 기준으로 맥락에서 프랑스어 대응 표현을 추출하면 다음과 같다.

- d) 샘솟는 내 눈물로 이내 뿌옇게 흐려진 그 얼굴 쪽에서 다시 그런 부드러운 목소리가 들렸다.
de la direction de son visage, brouillé par mes larmes, me parvint le son d une voix aimable.
- e) 다시 사람의 목소리가 들렸다.
j’ai entendu la voix pour la deuxième fois.
- f) 뒤이어 누구의 입에선가, 누가 빈틈을 냈어? 하는 흥분에 찬 목소리가 들렸다.
quelqu’un gueula : “qui c’est qui l’a laissée filer?”.
- g) 휘몰아가는 북풍과 처량한 가로수의 울음 속에서도 네 목소리를 들었다.
je t’ai entendue dans les hurlements du vent du nord et dans les pitoyables sanglots des arbres de la route.
- h) 그 여자의 목소리를 듣는 건 처음이다.
pour la première fois je l’entends, [...].

예문 e), g), h)를 제외하면²⁶⁾, *parvint le son*, *gueula* 등 다양한 동사들이 사용되고 있다. 비록 형태적인 차원에서는 서로 다르지만, 이러한 구문 검색을 통해서 한국어 ‘듣다’가 프랑스어에서는 기본 동사인 *entendre*뿐만 아니라 다른 표현 및 동사에 의해서도 실현될 수 있음을 실제 맥락을 통해서 확인할 수 있다. 더욱이 프랑스어 *voix*는 *entendre*와의 결합에서 ‘동사+목적어’ 구조로 실현되어 있는 반면, 한국어에서는 ‘목소리’가 ‘듣다’와의 결합에서 목적어뿐만 아니라 주어로도 구현되고 있다. 이를 통해서 이 어휘들의 결합에서 프랑스어와 한국어의 차이를 확인할 수 있다. 그럼에도 불구하고 술부 구조에서는 앞서 살펴본 프랑스어와 영어의 대응만큼 완벽하게 일치하는 표현은 없다고 할 수 있다. 만일 코퍼스를 병렬코퍼스로서 사용하였다면, 최소한 위에 제시된 예문만을 기준으로 한다면 두 언어 사이에서 번역 차원에서 완전히 일치하는 것은 없다

26) 예문 e), g), h)에서 프랑스어 표현은 *entendre*로 되어 있지만, 형태는 *entendu*, *entends*로 되어 있기 때문에 형태 차원에서는 *entendit la voix*와는 다르다고 할 수 있다.

고 할 수 있다. 그렇지만 비교코퍼스 관점에서는 번역적으로 대응될 수 있는 표현들을 관찰하면서 두 언어 사이에서 다양한 번역적 실현을 확인할 수 있었다.

술부 구조에서는 분석의 난점이 있었지만, 형용사+명사 수식 구조에서는 좀 더 분명한 번역 대응을 확인할 수 있다. <표 11>-<표 13>을 비교해 보면, 다양한 형용사들이 *voix* ‘목소리’를 직접적으로 수식하는 구문들을 찾을 수 있는데, 이를 통해서 각 형용사의 번역어들을 분류할 수 있다. 각 언어 별로 대응되는 형용사들을 정리하면 다음과 같다.

voix : basse, douce, claire, haute, forte, triste
목소리 : 낮은, 나지막한, 떨리는, 분명한, 부드러운

명사를 중심으로 한 이러한 수식 구조는 특히 연어와 관련해서 언어학적으로도 번역의 측면에서도 매우 어려운 문제이다. 연어의 한 예로 가장 대표적인 예라고 할 수 있는 ‘새빨간 거짓말’도 반복구문 분석을 통해서 영어 혹은 프랑스어에서 이에 해당하는 다양한 가능성으로 번역 텍스트에서 찾아낼 수 있다. 또한 위 예에서 프랑스어 형용사 *basse*가 한국어에서 ‘낮은’ 혹은 ‘나지막한’으로 다양하게 표현될 수 있음을 확인할 수 있다.²⁷⁾

이처럼 구문에 의한 번역 대응 관찰은 하나의 어휘 자체뿐만 아니라, 통사 구조, 어휘 결합, 어휘 그룹 등 다양한 층위에서 언어 간의 차이 등을 번역이라는 범주를 통해서 확인할 수 있도록 한다는 장점이 있다. 분명 프랑스어와 한국어는 서로 다른 언어적 양태를 보이기 때문에 완벽한 대응관계를 추적하는 것은 불가능하다고 할 수 있다. 그럼에도 불구하고 번역 텍스트에서 실질적으로 연관을 가지는 어휘들의 결집 양상을 문법적인 차원에서 혹은 번역의 차원에서 관찰할 수 있으며, 이를 번역 교육에서도 활용할 수 있다는 점에서 구문 검색은 유용한 방법론이 될 수 있다.

27) 이러한 정보가 축적되면 자동번역에서도 *voix basse*를 무조건적으로 ‘낮은 목소리’로 번역하는 것이 아니라, ‘나지막한 목소리’라는 또 다른 번역을 제시할 수 있다.

5. 결론

지금까지 우리는 번역 코퍼스에서 언어 요소를 관찰하기 위한 방법 중의 하나인 반복문 분석을 살펴보았다. 코퍼스 내에서 언어 요소들의 분포 양상 및 맥락을 살피기 위해서 다양한 방식이 사용된다. 키워드 검색과 그 결과인 어휘맥락색인은 키워드를 중심으로 한 일정한 길이의 어휘 그룹을 보여준다. 반면, 특정 맥락에서 함께 출현하고 있는 어휘들을 관찰하기 위해서는 주로 공기어 분석에 의존한다. 그리고 코퍼스에서 규칙적으로 반복되는 구문을 관찰하는 것이 곧 본 연구의 주제인 반복문 분석이라고 할 수 있다. 대부분의 코퍼스 분석 도구들은 이러한 역할을 담당하는 기능들을 갖추고 있으며, 코퍼스 언어학 및 기술의 발달과 함께 각 기법들은 이러한 언어 요소들을 다양한 관점에서 보다 세밀하게 관찰할 수 있도록 진화하였다.

그럼에도 불구하고 국내 코퍼스 기반 번역 연구에서는 단어 수준의 검색이 일반적이었고, 구문 탐색은 거의 이루어지지 않았다고 할 수 있다. 물론 이것이 번역 연구에서 단언어 이상의 언어 요소에 대한 번역 연구가 없었다는 것을 의미하는 것은 아니다. 오히려 전통적인 연구에서는 구문 수준의 연구가 보편적이라고 할 수 있다. 그럼에도 불구하고 코퍼스 기반 연구에서는 본 연구에서 살펴본 것과 같은 기법들을 활용한 구문 연구는 거의 없었다고 할 수 있다. 이것은 여러 가지 요인이 있지만 아무래도 국내에서는 공개된 대규모 번역 코퍼스가 존재하지 않을뿐더러, 분석 도구 및 분석 기법도 실용적인 관점에서 제대로 소개되어 있지 않기 때문인 것 같다. 연구자의 입장에서는 힘들게 코퍼스를 구축해서 분석 도구에 맞게 코퍼스를 조작하기보다는 텍스트를 직접적으로 관찰하는 기존의 방식의 훨씬 편리하기 때문이다.

본 연구는 어떠한 선행 작업 없이 원시 번역 코퍼스를 그대로 사용하여 반복문을 추출하고 이를 통해서 번역 대응 표현을 비교하는 방법을 보이고자 하였다. 분석 기법의 유용성을 검증하기 위해서 두 가지 종류의 코퍼스를 사용하였다. 영어와 프랑스어 그리고 한국어로 작성된 인권 협약서로 구성된 코퍼스와 한국어와 프랑스어 쓰인 문학 코퍼스를 분석 자료로 하여 언어에 따라서 텍스트 장르에 따라서 반복문 분석이 어느 정도의 타당한 결과를 이끌어 낼 수 있는지 여러 가지 예를 통해서 살펴보았다.

실용 텍스트에서는 언어적 특징이 다르더라도 번역 대응이라고 추정할 수 있는 구문들을 선별할 수 있었지만, 문학 텍스트에서는 그렇지 못했다. 앤그램 분석에 의한 반복구문 검색에서는 번역적으로 대응 가능한 구문을 발견하기가 어려웠다. 이것은 한국어와 프랑스어가 완전히 다른 유형의 언어이고, 실용 텍스트와 달리 서로 다른 이야기 구성된 문학 코퍼스라는 구성적 특성 때문이기도 하다. 그럼에도 불구하고 특정 키워드를 기준으로 한 반복구문확장 분석에서는 번역 대응어로 고려할 수 있는 구문들을 찾을 수 있었다. 물론 영어와 프랑스어처럼 대응 구문이 완전히 같을 수는 없지만, 오히려 그러한 구문들을 통해서 한국어와 프랑스어의 어휘 결합의 차이를 확인할 수 있는 계기가 되었다.

이러한 관점에서 반복구문 분석은 번역 대응어 추출뿐만 아니라, 번역을 통해서 분석 대상이 된 두 언어 사이에서 구문 구조의 차이를 예증하는 자료를 수집할 수 있다는 점에서 매우 유용한 방법이라고 할 수 있다. 또한 이러한 자료들은 구문의 번역에서 여전히 많은 문제점을 지닌 자동번역에서도 그리고 학교 현장에서 연어 및 관용어 같은 어휘 그룹의 교육에서도 효과적으로 재활용할 수 있는 일종의 참조목록이 될 수 있다. 구문 분석에 의해서 코퍼스로부터 추출된 구문 번역 대응 자료, 다시 말해서 구문 데이터베이스는 어휘가 아닌 문장 구조 차이에 의해서 나타날 수 있는 자동번역의 오류를 수정할 수 있는 참조 사전이 된다. 외국어 교육 및 번역 교육에서도 번역 차원에서 분류된 이러한 목록은 하나의 단어가 아닌 구문 자체를 비교차원에서 예증할 수 있는 지침서가 된다.

물론 본 연구에서 살펴본 것처럼 반복구문 분석 자체가 항상 만족할만한 결과에 이르는 것은 아니기 때문에, 신뢰할 수 있는 자료를 만들어내기 위해서는 다른 분석 기법과의 병행이 필수적이며, 분석 결과에 대한 연구자의 면밀한 검증이 필요할 것이다.

참고문헌

- 김정우 (2009) 「한국어 번역문의 구조적 명시화: 특정 보어 구문의 열세적 분포화 관련하여」, 『번역학연구』 10(3): 77-97.

- 유석훈 역 (1999) 『언어와 컴퓨터』, 서울: 고려대학교출판부.
- 이승아 (2012) 「‘Be being + 서술 형용사’ 구문의 영한번역전략: 영한 병렬 코퍼스를 활용한 분석」, 『번역학연구』 13(4): 159-83.
- 이영옥 (2002) 「한국어와 영어간 언어구조의 차이에 따른 번역의 문제: 인용문의 번역을 중심으로」, 『번역학연구』 3(1): 59-81.
- 이창수 (2011) 「엔그램(n-gram) 분석을 통한 번역한국어와 비번역한국어간의 어휘묶음 유형 차이 사례 연구」, 『통번역학연구』 15(1): 317-40.
- 조준형 & 이영훈 (2012) 「모데라토 칸타빌레 *Moderato Cantabile*』의 번역코퍼스에 나타난 저빈도 명사 연구」, 『통역과 번역』 14(1): 237-71.
- 조준형 (2010) *Analyse textométrique des corpus parallèles français-coréens*, Thèse de doctorat: Université Paris 3.
- 조준형 (2011) 「병렬코퍼스와 텍스트어휘추정법」, 『언어정보』 12: 85-109.
- Baker, Mona (1995) ‘Corpora in Translation Studies: An Overview and Some Suggestions for Future Research’, *Target* 7(2): 223-43.
- Bennett, Gena R. (2010) *Using CORPORA in the Language Learning Classroom - corpus linguistics for teachers*, The University of Michigan Press.
- Charles, Maggie (2012) ‘Using hands-on concordancing to teach rhetorical functions: evaluation and implications for EAP writing classes’ in Ana Frankenberg-Garcia, Lynne Flowerdew & Guy Aston (eds), *New Trends in Corpora and Language Learning* 26-43.
- Salem, André (1987) *Pratique des segments répétés: Essai de statistique textuelle*, Paris: Klincksieck.
- Sinclair, John (1991) *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.
- Sinclair, John (2003) *Reading Concordances*, London: Pearson Education.
- Vinay, Jean-Paul & Jean Darbelent (1977) *Stylistique comparée du français et de l’anglais*, Paris: Didier.
- Warren, Martin (2012) ‘Using corpora in the learning and teaching of phraseological variation’ in Ana Frankenberg-Garcia, Lynne Flowerdew

& Guy Aston (eds), *New Trends in Corpora and Language Learning*,
153-66.

Zimina, Maria (2002) 'Repérages lexicométriques des équivalences à basse
fréquence dans les corpus bilingues', *Lexicométrica* 1-27.

Zimina, Maria (2004) *Approches quantitatives de l'extraction de ressources
traductionnelles à partir de corpus parallèles*, Thèse de doctorat:
Université Paris III.

[Abstract]

**A Study on Translation Correspondences by Repeated Segments in
French-Korean Translation Corpora**

Cho, Joon-Hyung
(Korea University)

When extracting translation units from a translation corpus and making comparisons, it is common to use the Keyword in Context (KWIC) function. The results create a *concordance* which shows the contexts of a keyword, or a group of words which surround a keyword. However, a concordance list cannot provide a systematic explanation for the contextual and lexical patterns of this keyword if the list contains a wide variety of contexts. Therefore, context should be reorganized according to the lexical units of which it is composed. In this respect, it is also valuable to look directly at short groups of words (phrases), known as *clusters* or *repeated segments* in the field of textual statistics.

The aim of this study was to compare repeated segments in two types of translation corpora: an English-French-Korean corpus and a French-Korean corpus. To examine these units, we used two methods: *N-gram analysis* and *Repeated Segments Expansion*. The first method facilitated the observation of frequent clusters in the corpora, while the second one revealed lexical patterns to the left and right of a keyword. English and French showed very similar lexical patterns, whereas French and Korean did not. Nevertheless, the results of the comparison through lexical expansion verified a degree of similar lexical distribution between the French clusters and the Korean ones in terms of translation. These methods allow us to examine not only words but also translated expressions in the corpus. Moreover, they enable us to demonstrate contrasting characteristics between French and Korean in terms of lexical association.

▶ Key Words: translation studies, translation corpus, concordance, n-gram, repeated segments, segment expansion, KWIC

조준형

고려대학교 불어불문학과 강사

chojh4net@naver.com

관심분야: 코퍼스 번역학, 번역평가

논문투고일: 2013년 7월 31일

심사완료일: 2013년 9월 6일

게재확정일: 2013년 9월 16일