번역학연구
2014년 봄 제15권1호

번역교육을 위한 코퍼스기반 용어추출 방법*

박 명 수
(상명대)

I. 서론

외국어 지식이 있으면 누구나 번역을 할 수 있을까? 외국어에 대한 단순 지식만으로 할 수 있는 번역의 영역도 있지만, 지식의 다양화, 체계화, 세계화 등으로 인해 번역을 필요로 하는 영역은 일반적인 수준을 벗어나 '해당 분야를 모르면, 번역을 할 수 없다'라는 말이 통용될 정도로 전문번역을 필요로 하는 사례가 점점 늘고 있다. 전문번역(specialized translation)의 경우, 용어와 일반 어휘가 그리 어렵지 않게 구분이 되는 경우가 많지만, 관련 분야에 대한 사전 지식, 번역 경험 등이 부재할 경우에는 전문용어와 일반 어휘와 구분이 쉽지 않은 경우도 허다하다. 전문번역사들은 회계, 기술, 의학, 과학 등과 같은 특정 전문영역 번역을 위해 특정분야에서 사용되는 언어(Language for Specific Purposes)를 체계적으로 관리할 수 있도록 독립형 또는 웹기반으로 구동되는

* 이 논문은 2012학년도 상명대학교 교내선발과제 연구비의 지원을 받았음.

TM 소프트웨어, 용어추출 소프트웨어 등과 같은 여러 다양한 컴퓨터 및 코퍼스 기반 기술을 활용하고 있다. 전문번역사와 달리 특정 전문영역에 이제 막 진입하여 번역을 시작하거나, 번역사가 되기 위해 교육을 받는 학습자들에게 해당 분야에서 자주 사용되는 전문용어, 빈출어휘, 빈출어구 등을 분류하고 이를 체계적으로 숙지하고 관리하는 것은 전문가가 되기 위한 필수과정이 아닐 수 없다. 모나 베이커(Mona Baker 1992)는 번역에서 직면하게 되는 여러 문제의 원인이 어휘, 연어, 문법, 응집성, 일관성 등과 같이 다양한 수준에서 나타나는 등가어휘 및 표현으로 인해 빚어진다고 하였다. 이 중에서 가장 원초적이고 중요한 단초가 되는 요소가 어휘 및 어구라 할 수 있다. 본 논문은 특정분야에서 자주 사용되는 빈출어휘, 빈출어구 등을 코퍼스(corpus) 및 콘코댄서(concordancer)를 활용하여 추출해내는 효과적인 방법을 제안하는데 있다.

II. 선행연구 분석

우리말로 말뭉치라고 부르는 코퍼스는 텍스트의 모음이다. 코퍼스에 대한 학문적 정의를 살펴보자면, 싱클레어(Sinclair 1995: 17)는 '한 언어의 샘플로 사용하기 위해 언어적 기준에 따라 선별하여 배열한 특정언어의 모음'이라고 하였다. 컴퓨터 기술의 발달과 대중화 덕분에 방대한 양의 언어 텍스트를 한데 모으는 일이 가능해졌는데, 방대한 양의 데이터를 처리하기 위해 수집된 텍스트를 컴퓨터를 이용해 체계적으로 검색, 분류, 활용이 가능해져야 할 필요가 대두되었다. 이에 토니니-보넬리(Tognini-Bonelli 2001: 55)는 코퍼스를 '자동 또는 반자동으로 처리 또는 분석이 가능하도록 컴퓨터로 처리된 텍스트의 모음'이라고 정의하였다.

코퍼스 언어학(Corpus linguistics)이 번역학에 활용되기 시작한 것은 역사가 그리 길지 않다. 개인용 컴퓨터가 일반적으로 보급되면서부터 코퍼스를 기반으로 한 번역학 연구가 가능해졌기 때문이다. 학문적으로 코퍼스 언어학이 번역학에 등장한 것은, 베이커(Baker 1993)가 번역 텍스트 연구를 위한 코퍼스 활용법을 처음 논하면서 본격화되었다고 볼 수 있다. 아울러 영국 맨체스터 대학교의 통번역학 센터에서 번역영어코퍼스(Translational English Corpus) 구축을 위

해 수행한 프로젝트에서 코퍼스 언어학의 다양한 분석 기법이 번역학에 도입되었다(Stubbs 1996). 당시 번역학에 코퍼스 분석기법이 도입된 주요 목적은 다양한 번역체 문장들의 언어적 특성을 파악하기 위해서였다.

컴퓨터 기술의 발달과 코퍼스 분석 도구의 발달, 분석 기법의 다양화 덕분에 코퍼스 기반 연구가 늘어나고 있으며 코퍼스 연구가 번역에 미치는 영향에 대한 다양한 연구도 크게 늘어났다(Manca 2011; Partington 2011; Olohan 2001). 국내에서도 코퍼스를 활용한 번역학 연구가 늘고 있는 추세인데, 코퍼스 언어학과 코퍼스 중심 번역학 접근법의 개괄(정호정 2008), 병렬 말뭉치에 기반한 연구(조의연 2009; 최승권 & 김영일 2010; 최정아 2003; 황은하 2013, 2009), 전문분야 번역을 위한 소규모 코퍼스 구축 및 활용(유정주 2013) 등 연구영역이 점차 확대되고 있다. 코퍼스를 기반으로 한 번역학 연구는 번역 텍스트에 대한 어휘 또는 문체 분석뿐 아니라 기계번역(machine translation), 더 나아가 번역사 교육 등의 목적으로도 활용되고 있다. 번역사 양성 및 교육과 관련한 코퍼스의 활용에 대해 최초로 언급한 것으로 알려진 버나르디니(Bernardini 1997)는 번역 학습자들이 전문 번역가가 되기 위해 필요한 기술을 익힐 수 있도록 기존 번역교육 방식에 대규모 코퍼스 콘코댄싱 기법이 보완적으로 활용되어야 한다고 주장하였다. 자네틴(Zanettin 1998)도 번역교육에 코퍼스를 기반으로 한 번역 실습 활동이 포함되는 'Translator Trainee Workstations(번역교육용 워크스테이션)'를 제안하기도 했다.

번역사들의 경우 특정 전문영역 텍스트를 반복적으로 번역할 경우 이전에 번역한 부분을 재활용하기 위해 번역메모리(Translation Memories/ TM)를 구축하고 활용하는데, 이러한 TM이나 예제 기반의 기계번역(example-based MT programs) 그리고 혼합형 기계번역(hybrid MT programs)들도 컴퓨터 기술을 활용한 것으로 구축한 코퍼스를 통계적으로 어휘, 문법적으로 분석하는 기법이 적용되어 번역에서 컴퓨터 및 코퍼스 활용이 점점 많아질 수밖에 없다. 마하디(Mahadi 2010)는 코퍼스 활용 덕분에 원천 텍스트를 보다 잘 이해할 수 있게 되고, 다양한 텍스트 유형뿐 아니라 전문용어도 쉽게 파악할 수 있고, 이들 용어의 연어, 개념, 용어 등에 대한 정보도 보다 손쉽게 얻어낼 수 있다고 주장하였다. 여러 학자들이 번역에서 코퍼스 활용의 중요성과 장점에 대해 언급하였는데(Maia 2005; Kubler 2003; Yao 2008), 마이아(Maia 2005)는 번역사 교육에

서 코퍼스의 일반적 사용을 접목하는 것이 아주 중요하며, 이를 통해 어휘에 대한 사전적 정의를 뛰어 넘어, 해당 어휘의 문맥을 파악하고 연어적 관점에서 올바른 어휘를 선택하는 능력을 기를 수 있다고 하였다. 야오(Yao 2008)는 중국어에서 영어로 번역하는 학생들을 사전기반 번역 그룹과 코퍼스기반 그룹으로 나눠 비교한 실험연구에서 역시 동일한 주장을 하였다. 코퍼스를 활용한 번역 교육을 받은 학생들은 번역을 할 때 더 풍부한 자료를 기반으로 실제 번역에서 폭넓은 어휘와 표현의 선택권을 갖게 된다고 밝혔다.

이상에서 언급된 바와 같이 번역과 코퍼스는 이제 불가분의 관계라고 해도 무방할 정도이고, 다양한 연구자들의 번역학 연구에서 코퍼스를 기반으로 연구의 영역을 넓혀가고 있다. 다만 번역교육 현장에서 코퍼스를 활용한 장점에 대한 연구는 국내에서 그리 많지 않은 실정이다. 이에 본 논문은 특정분야에서 자주 사용되는 빈출어휘, 빈출어구 등을 코퍼스 및 콘코댄서를 활용하여 추출해 내는 효과적인 방법을 제안하고, 코퍼스 기반의 번역교육 모델을 제시하고자 한다.

III. 코퍼스 기반 용어 추출 방법

3.1 코퍼스 구축

영어 통번역 전문용어를 수강하고 있는 학부재학생 75명이 우리나라 대표 영자신문이라 할 수 있는 The Korea Herald와 The Korea Times에서 2000년~2013년까지 실린 날씨에 관련된 신문기사를 텍스트 파일로 변환하여 미니 코퍼스인 Korean Weather Corpus (KWC)를 구축하였다. KWC는 총 291개의 파일에 어휘 크기는 총 88,042개이다.

3.2 코퍼스 도구

KWC 분석을 위해 활용한 콘코댄서는 무료로 쉽게 누구나 사용할 수 있는 것을 선택하였다. 코퍼스 연구에서 가장 많이 활용되고 있는 콘코댄서로 Word Smith와 AntConc가 있는데, Word Smith의 경우 유료이며, AntConc는 무료로

컴퓨터에 설치해서 사용할 수 있다. 아울러, 웹에서 무료로 사용할 수 있는 콘코댄서 및 코퍼스 분석 도구로 Compleat Lexical Tutor(그림 2. 참조)가 있다. 본 연구에서는 코퍼스를 활용해서 번역텍스트에 자주 등장하는 어휘를 쉽게 추출할 수 있도록 하기 위해 학생들에게 AntConc(그림 1. 참조)를 사용하였다.

그림 1 AntConc 캡쳐 화면

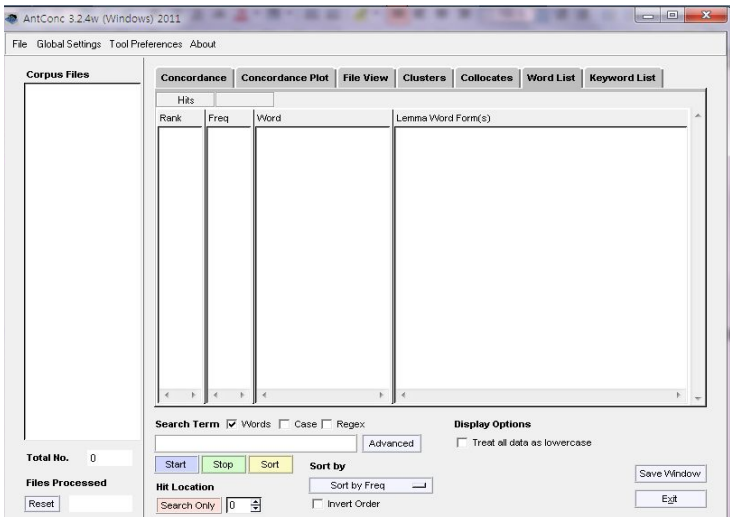
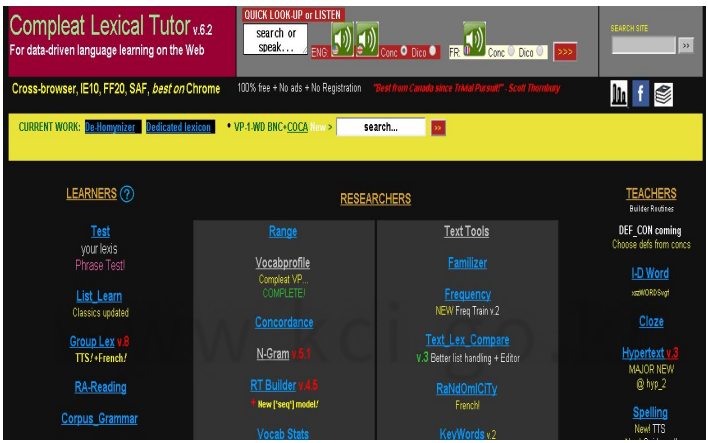


그림 2 Compleat Lexical Tutor 웹페이지



3.3 용어 추출방법

콘코댄서를 활용하면 분석대상 코퍼스의 어휘 분석을 손쉽게 할 수 있다. 어휘 분석에서 가장 일반적으로 하게 되는 것이 빈도수에 따른 어휘 분류이다. 빈도수에 따라 어휘를 분석하면 일반적으로 의미상 역할은 적거나 문법적인 기능면에서 필수적인 어휘들인 기능어(function words)와 의미상 역할이 크게 부여되는 내용어(content words)들이 모두 함께 분석되어, 흔히 코퍼스에서 최빈도 어휘를 상위 30개 추출하면 아래 그림과 같이 내용어보다 기능어들이 상위를 차지하게 된다.

〈표 1〉 BNC 상위 30개 최빈도 어휘목록

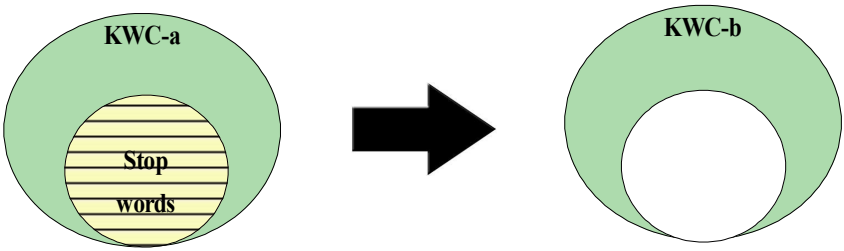
순위	어휘	출현빈도	순위	어휘	출현빈도
1	the	6041234	16	with	658584
2	of	3042376	17	as	653612
3	and	2616708	18	be	650082
4	to	2593729	19	he	639449
5	a	2164238	20	at	521694
6	in	1937819	21	by	512214
7	that	1118985	22	are	464272
8	it	1054279	23	have	460626
9	is	990281	24	this	453528
10	was	881473	25	not	451291
11	for	878727	26	but	445735
12	i	868634	27	from	424972
13	's	783990	28	had	420247
14	on	729518	29	they	419562
15	you	667363	30	his	408970

<표 1>은 1억 단어 규모의 대형 코퍼스인 British National Corpus (BNC)에서 최빈도 어휘를 추출한 것이다. 위의 표에 포함된 대부분의 어휘들은 기능적 측면에서 문법적인 역할이 크지만 번역에 필요한 용어라고 보기는 어려운 것들

이다. 이러한 기능중심 어휘들을 제외하고 분석하면 원래 의도했던 어휘목록을 추출할 수 있는데, 위와 같이 관사, 전치사, 대명사, 접속사, 조동사 등의 기능적인 어휘들을 언어학에서 stopwords(불용어 또는 정지어)라고 부른다. 이러한 stopwords 목록을 만들어 콘코댄서를 이용한 분석에서 해당목록을 제외한 분석결과를 도출하면 손쉽게 내용어 위주의 어휘를 추출할 수 있다.

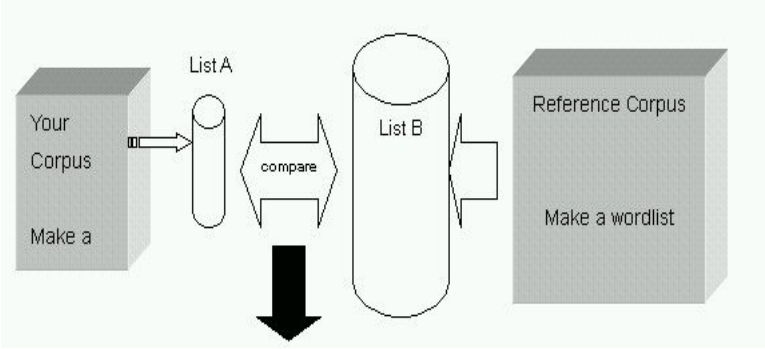
첫 번째, 용어추출 방법은 stopwords 개념을 활용한 코퍼스 분석 기법이다. 그림 3은 본 연구에서 용어추출 훈련을 위해 학생들에게 코퍼스를 기반으로 콘코댄서를 활용한 방법을 쉽게 설명하기 위해 제시한 용어추출과정을 그림으로 도식화한 것이다. 먼저 그림 3의 경우, 좌측의 KWC-a 전체 어휘에서 stopwords를 제외하면 우측과 같이 내용어 위주의 용어만이 남은 KWC-b가 남게 되고 이를 콘코댄서로 빈도분석을 실시하면 KWC에서 가장 많이 사용되는 어휘 및 표현목록 추출을 위한 틀이 마련된다.

그림 3 stopwords를 이용한 용어추출 방법 도식화



두 번째 방법은, 코퍼스 분석방법에서 키워드 분석(keyword analysis)을 이용하는 것이다. 키워드 분석을 위해서 두 개의 코퍼스가 필요하다. 키워드 분석은, 분석하고자 하는 코퍼스와 참조코퍼스(reference corpus)를 비교하여, 참조코퍼스보다 출현빈도가 많거나 적은 어휘 목록들을 추출하는 방식이다.

그림 4 키워드 분석을 통한 용어추출 방법 도식화[1)]



이상과 같은 두 가지 방법, 즉 stopwords 개념 활용과 keywords 개념 활용을 이용해 본 연구의 대상인 우리나라 날씨 미니 코퍼스인 KWC에서 날씨 관련 어휘 및 표현들을 추출하고 분석하였다.

IV. 코퍼스 기반 용어추출 사례

4.1 Stopwords를 활용한 용어 추출방법

앞서 설명한 방법을 활용한 KWC 용어추출 방법을 설명한다. 먼저 AntConc를 활용하여 어휘 분석을 실시하였다. 그림 5는 stopwords가 적용되기 이전의 어휘 분석 결과 화면이다.

1) 이 그림은 http://www.lancaster.ac.uk/fss/courses/ling/corpus/blue/l03_2.htm 에서 캡쳐한 것으로, 코퍼스연구에서 키워드 분석에 대한 개념을 설명하고 있다.

그림 5 KWC 1차 어휘분석

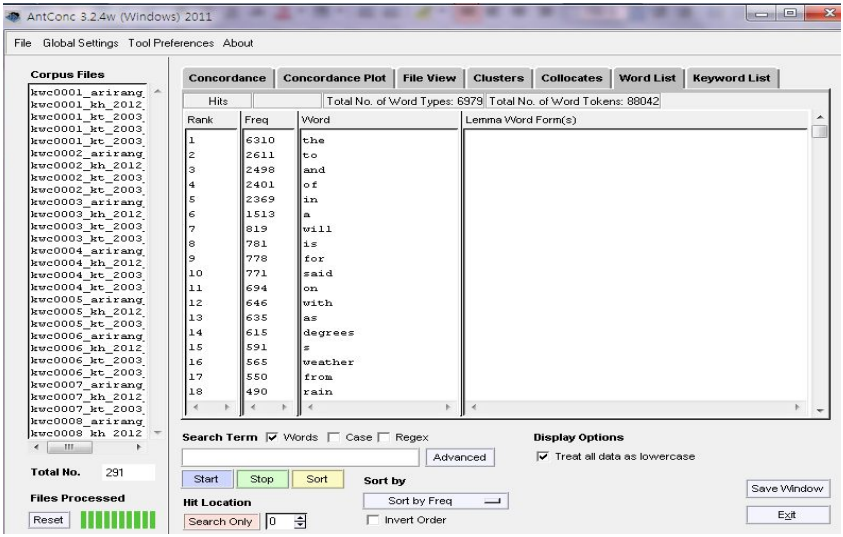


그림 5에서 분석한 어휘 목록을 빈도순서로 나열해 상위 50개를 추출하면 앞서 <표 1>에서 BNC 결과와 마찬가지로 기능어들이 많이 나타난다. 아래 [표 2]는 KWC에 stopwords를 적용하지 않고 분석한 결과를 정리한 것이다. 상위 50개 중에서 날씨와 관련된 용어로 분류할 수 있는 어휘는 degrees, weather, rain, KMA, expected, heavy, typhoon, temperatures, snow, celsius, morning, minus 정도로 12개에 불과하다. 나머지 어휘들은 앞서 언급한 바와 같이 내용적인 의미에서 중요도가 낮지만, 언어의 기능적 측면에서 문법적인 역할이 크다고 볼 수 있는 어휘들로, 실제 번역에서 큰 의미를 찾을 수 없거나 날씨와 같이 특정 전문영역에서 자주 출현하는 어휘, 용어, 표현의 기초가 된다고 보기 어려운 것들이다.

〈표 2〉 KWC 상위 50개 최빈도 어휘목록

1	the	26	were
2	to	27	was
3	and	28	by
4	of	29	province

5	in	30	KMA
6	a	31	expected
7	will	32	this
8	is	33	nation
9	for	34	heavy
10	said	35	have
11	on	36	typhoon
12	with	37	more
13	as	38	temperatures
14	degrees	39	snow
15	s	40	celsius
16	weather	41	morning
17	from	42	year
18	rain	43	than
19	seoul	44	up
20	that	45	but
21	it	46	country
22	be	47	south
23	korea	48	minus
24	at	49	also
25	are	50	some

전문영역에 대한 번역을 반복적으로 하는 경우, 자주 사용되는 용어, 표현 등을 TM 등을 이용해 체계적으로 관리할 수도 있으나, 이러한 TM 소프트웨어들은 비용이 만만치 않아, 번역을 공부하는 학생들의 입장에서 쉽게 구입해서 사용하기도 어렵다. 위와 같은 목록에서 기능어에 해당하는 어휘들을 stopwords로 만들어서 사용할 때 한 가지 염두에 두어야 하는 것이 있다. stopwords 목록은 인터넷 검색을 통해 쉽게 찾을 수 있는데, 수록된 어휘들이 서로 상이한 경우가 있다. 이는 stopwords가 단 하나가 아니라는 의미이며, 목적에 맞게 가감 과정을 통해 작성해서 사용할 수 있다는 것을 의미한다. 본 연구에서 KWC 코퍼스의 경우 우리나라 지명, 기관명 등과 같은 고유명사 또는 일반명사 중에서 코퍼스의 특성상 자주 등장하는 어휘들 예를 들어 위의 [표 2]에서 Seoul, Korea, province, country, south, year, also, some 등은 전문용어라고 보기 어려운 것들이다. 이런 어휘들을 모아 stopwords 목록에 추가해서 사용하면 목적에 맞는 어휘 목록들만을 추출해낼 수 있다.

다음 <표 3>은 수작업으로 추가한 어휘들이 포함된 stopwords 목록을 적용한 후 빈도별로 어휘검색을 하여 추출한 상위 50개 KWC 어휘 목록이다.

〈표 3〉 Stopwords를 적용한 KWC 상위 50개 최빈도 어휘목록

1	degrees	26	yellow
2	weather	27	wave
3	rain	28	low
4	province	29	rainfall
5	KMA	30	storm
6	expected (v)	31	winds
7	heavy	32	millimeters
8	typhoon	33	centimeters
9	temperatures	34	pressure
10	snow	35	damage
11	celsius	36	warning
12	morning	37	likely
13	cold	38	sea
14	administration	39	strong
15	forecast (n) (v)	40	water
16	meteorological	41	front
17	high	42	hot
18	heat	43	coast
19	dust	44	caused (v)
20	hit (v)	45	record
21	issued (v)	46	reported (v)
22	continue (v)	47	skies
23	season	48	snowfall
24	mm	49	mercury
25	average	50	kilometers

위의 표에 나타난 바와 같이 추출된 상위 50개 어휘들이 <표 2>에 비해 훨씬 날씨 영역에 훨씬 근접한 어휘 목록임을 알 수 있다. KWC에서 이들 어휘의 출현빈도는 63~615에 해당하는데, 이들 어휘들이 날씨를 보도하는 글에서 많이 사용된다는 것을 확인할 수 있다.

코퍼스를 활용하면 해당 어휘가 어떤 문맥 상황에서 사용되는지를 파악할 수 있는데, 바로 각 검색 어휘의 콘코던스(concordance)를 통해서 볼 수 있다. 이를 Key Word in Contexts (KWIC) 검색이라고 한다. 위의 <표 3>에서 가령

일기예보에서 가장 많이 사용되는 동사 중 하나인 20번째 최빈출 어휘인 'hit'에 대한 KWIC 검색을 살펴보면 다음 그림 6과 같다. 학생들은 hit의 사전적 의미에 국한되지 않고, 실제 일기예보 텍스트에서 hit가 어떤 모습으로 사용되는 지에 대한 실제적인 용례를 학습할 수 있는 장점이 있다. 인터넷의 생활화, 스마트폰의 보급으로 요즘 학생들은 어휘를 학습할 때 온라인에서 쉽게 접할 수 있는 네이버, 다음과 같은 포털 웹사이트에서 제공하는 사전을 사용하거나 사전 앱을 사용하면서 단지 사전적 정의만 확인하고 실제 예문을 통해 해당 어휘가 등장하는 다양한 용례는 간과하는 경우가 많다는 점을 생각해 볼 때, 이러한 코퍼스와 코퍼스 툴을 활용하는 것이 외국어 학습측면에서도 아주 긍정적이라 볼 수 있다.

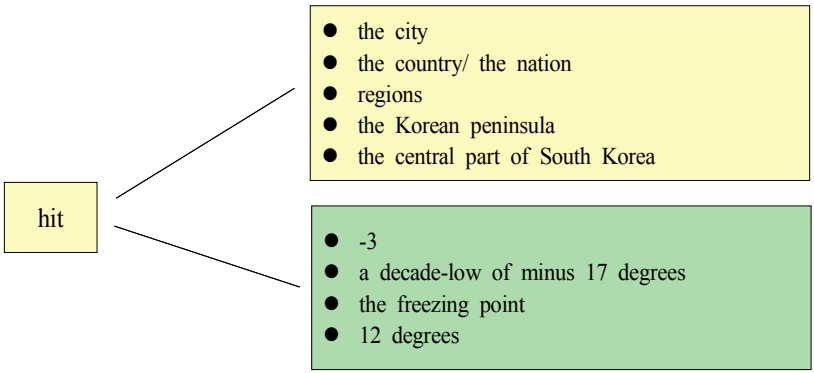
그림 6 "Hit"의 KWIC 검색 결과

Concordance | Concordance Plot | File View | Clusters | Collocates | Word List | Keyword List

Hit	KWIC	File
1	zau region, one of the worst hit, on Saturday. He called on t	kwc0001_
2	this season, with one or two hitting the Korean Peninsula. The	kwc0002_
3	ll from the heaviest rain to hit Beijing in over 60 years has	kwc0003_
4	n district, the most rain to hit the city in a 14-hour period	kwc0003_
5	ge from Typhoon Maemi, which hit the country during the Chusok	kwc0003_
6	ge from Typhoon Maemi, which hit the country during the Chusok	kwc0004_
7	travel to Alabama the worst-hit state where some one million	kwc0006_
8	Province, one of the worst hit regions by the typhoon. Some	kwc0006_
9	cture of the damage emerges. Hitting the nation at the end of	kwc0006_
10	e, the fastest winds ever to hit the nation. Typhoon Sara, whi	kwc0006_
11	use a strong cold front will hit the Korean peninsula once aga	kwc0007_
12	g Province, one of the worst hit regions by the typhoon. Some	kwc0007_
13	cture of the damage emerges. Hitting the nation at the end of	kwc0007_
14	e, the fastest winds ever to hit the nation. Typhoon Sara, whi	kwc0007_
15	be at 5… and Mt.Gumgang will hit -3.	kwc0008_
16	The temperature in Seoul hit a decade-low of minus 17 degre	kwc0009_
17	e up to -2… while Daegu will hit the freezing point. Both Gwan	kwc0009_
18	ernoon. Seoul is forecast to hit 12 degrees, Daegu will peak a	kwc0010_
19	Korea Heavy snowfall that hit the central part of South Kor	kwc0010

아래 그림은 위의 KWIC 검색 결과를 토대로 날씨와 관련해서 자주 사용되는 동사 'hit'의 용례를 20개의 콘코던스 라인을 바탕으로 추출하여 도식화 한 것이다.

그림 7 동사 "Hit"의 용례 도식화



동사 hit는 위의 예문들에서 완전 타동사로 사용되고 있음을 확인할 수 있다. 고빈도 목적어 결합 양상을 살펴보면 첫째, 지명, 장소가 이어져, hit the city/ the nation/ the country/ regions/ the Korean peninsula 등이 날씨 예보 글에서 자주 등장함을 알 수 있다. 둘째, 기온을 나타내는 숫자가 목적어로 이어져, hit -3/ a decade-low of minus 17 degrees/ the freezing point/ 12 degrees 등의 구문을 확인할 수 있다. 실제 이러한 구문이 사용된 아래 코퍼스의 예문을 통해 보다 정확하게 전체적인 맥락을 파악할 수 있다.

- The death toll from the heaviest rain to **hit Beijing** in over 60 years.
- Up to 46 centimeters of rain fell in Fangshan district, the most rain to **hit the city** in a 14-hour period since records began in 1951.
- Meanwhile, property damage from Typhoon Maemi, which **hit the country** during the Chusok holidays, surged to 4.3 trillion won
- A strong cold front will **hit the Korean peninsula** once again on Thursday
- The temperature in Seoul **hit a decade-low of minus 17 degrees** Celsius today.
- Temperatures for Friday…Seoul will rise up to -2… while Daegu will **hit the freezing point**.

이와 같이 해당 어휘와 자주 결합하여 사용되는 예제들을 직접 찾아 정리

를 해보는 분석적 접근을 통해, 목표언어 학습 효과(Boulton, 2010), 자기주도적 학습 효과(Huang, 2011), 의식적 주목 효과(Schmidt, 1990)를 통해 번역사로서 특정 영역에 자주 출현하는 어휘 및 표현 등을 정리해서 보다 정교하고 자연스러운 표현이 가능해질 수 있다.

4.2 Keyword 검색을 활용한 용어 추출방법

앞서 설명한 바와 같이 키워드 검색을 하려면 분석대상 코퍼스와 더불어 참조코퍼스(reference corpus)가 더 필요하다. 그림 4에 도식화된 대로, 분석대상 코퍼스에 비해, 참조코퍼스는 어휘 규모도 커야 하고, 분석대상 코퍼스의 특정 영역의 특성이 잘 드러나도록 하려면 코퍼스의 내용도 특정대상이기보다 일반적인 내용을 담은 코퍼스가 바람직하다. 무엇보다 참조코퍼스의 어휘 규모가 분석대상 코퍼스 보다 최소 5배 이상 커야 키워드분석의 통계적 의미면에서 유의미한 결과를 얻을 수 있다(Berber-Sardinha 2000).

AntConc를 사용하면 이러한 키워드 검색이 쉽게 이뤄질 수 있지만, 먼저 번역 교육을 받는 학생들이 이러한 코퍼스 간 비교 검색에서 필수적으로 알아야 하는 개념이 비교 분석에 적용되는 통계적인 개념이다. 키워드 검색을 위해 일반적으로 카이제곱 검정(Chi-square test)과 로그-가능도(Log-likelihood) 함수 개념이 적용된다. 본 연구에서는 통계적 유의미를 파악하기 위한 해당 어휘들의 출현빈도를 정밀 측정하는 것이 아니라, 키워드 검색의 개념을 활용한 특정 분야의 어휘, 용어, 표현 등을 파악하는 것이기에, 이러한 통계적인 개념에 대해서는 자세히 다루지 않는다.

그림 8 KWC의 키워드 검색 결과

Concordance | Concordance Plot | File View | Clusters | Collocates | Word List | Keyword List

Hits | Keyword Types Before Cut: 6673 | Keyword Types After Cut: 926

Rank	Freq	Keyness	Keyword
1	615	3889.676	degrees
2	566	3303.645	weather
3	490	3032.930	rain
4	384	2426.359	province
5	486	2419.419	seoul
6	375	2412.684	kma
7	372	2198.032	expected
8	294	1891.544	temperatures
9	296	1891.110	typhoon
10	281	1807.905	celsius
11	306	1687.788	heavy
12	284	1517.116	snow
13	230	1466.981	forecast
14	224	1441.177	meteorological
15	230	1403.833	administration
16	219	1377.469	southern
17	277	1351.612	morning
18	205	1306.365	agency

본 연구에서 키워드 검색에 활용한 참조 코퍼스는 YELC이다. YELC는 연세대학교에서 신입생들을 대상으로 쓰기 평가에서 수집한 대학 신입생들의 일반적인 주제에 대한 쓰기 내용을 기반으로 한 100만어 규모의 코퍼스이다. YELC의 내용이 일반적인 주제에 대한 쓰기라, 본 연구의 분석 대상 코퍼스인 날씨와 내용이 아주 다르기 때문에, 일반적인 주제의 글에 비해 상대적으로 출현 빈도가 높은 어휘들이 키워드 검색에서 추출된다는 전제가 가능해진다.

위의 그림 8은 이러한 참조코퍼스와 비교해서 추출한 KWC의 키워드 분석 결과화면이다. 앞서 <표 3>의 결과와 크게 다르지 않음을 알 수 있다. 그림 8에서 세 번째 칼럼의 'keyness'에 나타난 수치들은 로그우도 값을 바탕으로 해당 어휘가 참조코퍼스에 비해 KWC에서 나타나는 출현 가능성에 대한 통계값을 의미한다. 즉, 이 값이 클수록 참조코퍼스에 비해 KWC에서만 특이하게 자주 출현함을 의미하게 되고, 이러한 값이 높은 상위 최빈도 어휘를 추출하면, KWC에서 자주 사용되는 어휘목록을 만들어 낼 수 있다.

〈표 4〉 Keyword 분석을 통한 KWC 상위 50개 최빈도 어휘목록

1	degree	26	storm
2	weather	27	winds
3	rain	28	wave
4	province	29	yellow
5	KMA	30	centimeters
6	expected	31	continue
7	temperature	32	spell
8	typhoon	33	highs
9	celsius	34	lows
10	heavy	35	warning
11	snow	36	pressure
12	forecast	37	hit
13	meteorological	38	rainy
14	administration	39	snowfall
15	southern	40	skies
16	morning	41	reported
17	peninsula	42	tropical
18	heat	43	nationwide
19	dust	44	sea
20	cold	45	showers
21	mm	46	winter
22	season	47	percent
23	issued	48	range
24	average	49	downpours
25	rainfall	50	atmospheric

위의 <표 4>의 최빈도 어휘 50개는 참조코퍼스와 비교해 볼 때 KWC에서 출현빈도가 높은 어휘들 목록이다. 이 표의 어휘 목록만 보더라도 해당 코퍼스의 주된 내용이 날씨와 관련되었을 것이라는 짐작이 가능한데, 바로 이런 것을 코퍼스 언어학에서 해당 텍스트의 'aboutness'라고 부른다. 이는 언어학, 도서정보학 등의 학문에서 사용되는 용어로 해당 텍스트의 주제, 내용, 의도를 파악할 수 있는 개념으로 볼 수 있다. 즉 위의 최빈도 어휘 50개는 KWC의 글의 성향을 알려주는 일종의 지표 같은 역할을 한다는 것을 의미한다. 이러한 어휘가 KWC의 대표성을 지니며, 각각의 어휘들의 쓰임새를 파악하여 정리하면, 날씨와 관련한 전문 번역을 위한 번역사로서 전문영역의 빈출 어휘목록 작성과 같

은 기초 준비 작업을 값비싼 Translation Memory 소프트웨어를 구입하지 않고 손쉽게 해결할 수 있다.

V. 결론

5.1 결론

본 연구의 목적은 전문영역 번역을 위해 필수 사전작업 과정 중 하나인 특정 전문영역에서 자주 사용되는 어휘, 표현 등을 추출하는 방법을 코퍼스와 코퍼스 분석 도구를 활용하여 해결하는 방법을 모색하는 것이었다. 첫째 방법으로 제시된 “stopword list”를 기반으로 한 용어추출 방법은, 관사, 전치사 등의 문법적 기능어에 연구 및 교육 목적에 맞게 목록에 지명, 인명 및 검색 분석에서 불필요한 어휘를 추가하여, 사용 빈도가 높은 어휘 목록을 추출하는 방식이었다. <표 2>와 <표 3>의 비교를 통해 stopword 목록 기반 검색으로 날씨 보도만을 모은 코퍼스인 KWC에서 자주 사용되는 어휘를 추출이 가능하다는 것이 확인되었다. 컴퓨터만을 의존하여 분석 결과를 100% 신뢰하는 것은 어려우나, 번역이나 통역에서 특정 전문영역을 전문적으로 번역하여 관련 어휘, 용어 등의 목록을 체계적으로 관리할 필요성 측면에서 볼 때 이러한 방법이 전문가가 되기 위한 기초 작업으로서 충분히 가치가 있다고 판단된다. 둘째 방법은, 참조 코퍼스(reference corpus)와 분석 대상 코퍼스를 비교분석하여 통계값을 기반으로 한 “keyword” 분석이었다. 그림 8과 <표 4>를 통해 알 수 있듯이 이러한 키워드 검색 분석을 통해서도 날씨라는 특정 영역에서 자주 사용되는 어휘, 표현 등의 추출이 가능하다는 것을 확인하였다.

이러한 두 가지 용어 추출방법을 용어와 관련하여 번역사들이 겪는 어려움이나 문제에 대한 완벽한 해결책으로 제시한 것은 아니다. 위에서 제시한 두 가지 방법은 단지 어휘, 표현 목록만을 추출해내는 것이고, 전문 분야에 종사하는 전문가의 자문, 번역사의 추가 작업 등을 토대로 번역, 통역을 위한 특정 분야의 용어집을 만드는 과정으로 이어져야 할 필요가 있다. 실제 번역사들에게는 이들 용어를 엑셀 파일 등으로 병렬 구조로 만든 용어집이 더 직접적으로 도움

이 될 것이다. 비용이 많이 들고, 접근성이 상대적으로 떨어질 수 있는 전문가용 소프트웨어 등이 아니더라도, 무료로 활용할 수 있는 AntConc와 같은 코퍼스 분석 도구를 활용하여, 코퍼스 언어학의 분석 기법을 통해 번역교육, 번역학, 번역 실습 등의 과정에 필요한 용어 추출을 가능하게 만드는 방법을 제시하는 것이 본 연구의 의도였다.

5.2 교육적 함의

컴퓨터 사용이 일상화되고, 다양한 언어학 연구에서 대규모 데이터를 기반으로 한 연구가 활성화되고 있다. 특히 코퍼스를 기반으로 한 언어학적 연구는 단순한 어휘 분석에서부터 담화분석에 이르기까지 그 활용도가 나날이 높아져 많은 연구자들이 이에 대한 많은 관심과 연구물을 쏟아내고 있다. “classroom-based concordancing” “data-driven learning(DDL)”라는 용어로 Johns (1986)가 코퍼스를 언어교육에 활용할 수 있다는 주장을 내세운 지 30년이 다 되어가고 코퍼스가 번역학 연구에도 적극적으로 활용되고 있다. 본 연구는 이러한 코퍼스와 코퍼스 분석 도구를 이용해, 번역사들에게 필요한 목표분야 특히 날씨와 같은 특정 영역의 빈출어휘를 검색하여, 해당 분야의 용어와 그 용례를 파악하는 방법을 소개하고자 하였다.

전문 번역사들의 경우 오랜 경험과 개인의 노력을 통해 특정 전문영역에 대한 전문성을 확보하여, 해당 분야의 전문용어를 이미 잘 알고 있기도 하지만, 번역사 양성이라는 관점에서 본 연구가 의도한 것은 전문적이고 자연스러운 번역을 위해 필요한 과정 중 하나인 용어, 표현 정리라는 관점에서 출발한 것이다. 이는 번역사 개인의 노력에 의한 전문 분야 지식 습득이 아니라, 보다 체계적인 번역사 양성 교육의 관점에서 번역 교육현장에서 보다 체계적인 실습활동을 통한 실전 연습을 쌓고, 현장에 나가는 하나의 교과과정 모델을 구축하기 위한 전초작업이라고 할 수도 있을 것이다.

전문 번역사가 되기 위해 외국어 능력, 컴퓨터 기술, 폭넓은 상식, 번역 분야에 대한 전문적 지식 등이 요구된다. 이러한 지식적 측면에 더해 번역사가 되기 위해 대학교 이상의 기관에서 교육을 받는 학생들의 경우 능숙한 외국어 능력이상으로 실제 번역을 위해 필요한 것이 바로 실전 경험이다. 구아덱

(Gouadec 2007)은 번역사 양성교육에서 번역능력 배양에 대해 숙련된 기술공을 비유로 들며, 번역능력 향상을 위해서 다른 이들의 번역과정을 보고 함께 작업해보고, 실전 경험을 해야 한다고 주장하였다. 번역 교육현장에서 보다 체계적으로 전문적인 번역 교육을 위해 코퍼스를 활용하는 것이 장점이 있음에도, 아직까지 번역사 양성 교과과정에서 하나의 교과목으로 다뤄지지 않는 이유 중 하나가 기존 코퍼스와 분석 도구들이 언어학적 연구를 위해 만들어져 있기 때문이다. 아울러, 코퍼스 도구 활용이 전문 연구자들만의 전유물이라는 잘못된 인식으로 인해, 손쉽게 활용할 수 있는 코퍼스와 코퍼스 도구를 배우려는 동기가 부재하다. 이러한 현상은 코퍼스를 외국어 학습 교실에 적용하는 데서 빚어지는 여러 가지 어려움을 통해서도 알 수 있다. 쿡(Cook 1998)은 이와 관련해서 언어학적 관점에서 교육적 관점으로 전환하는 것이 결코 간단치 않다라고 주장했다.

코퍼스를 번역사 양성과정에 접목시키는 시도가 무엇보다 절실하다. 이를 위해 필요한 것이 첫째, 소규모 코퍼스 구축이다. 기존의 BNC와 같은 초대형 코퍼스가 아니라 실제 번역 대상인 번역텍스트를 기반으로 한 번역 교육용 소규모 코퍼스를 구축하는 것이 필요하다. 번역교육을 받고 있는 학생들과 더불어 협동으로 소규모 코퍼스를 구축해서 언어학 연구를 위해 필요한 분석이 아니라, 번역을 위해 필요한 어휘, 용어, 표현 등에 대한 분석을 시도하면서 Tim Johns가 언급한 것과 같은 컴퓨터 기반의 DDL이 다양하게 시도되어져야 한다. 소규모 코퍼스만으로도 교실에서의 번역교육 활동이 충분한 이유는 무엇보다 BNC, American National Corpus(ANC), Corpus of Contemporary American English (COCA) 등과 같은 대규모 코퍼스들을 활용하기에 버겁다는 점이다. 전체 큰 말뭉치를 사용하기보다는 적은 양이지만 교육 목적에 맞는 내용과 크기의 코퍼스만으로도 교육이 가능하기 때문이다. 실제 DDL 기반의 수업활동에서도 콘코던스 라인(concordance line)의 일부만을 보여주거나 KWIC 검색결과만으로 교육을 하는 경우가 빈번하다(Schmitt & Schmitt 2005).

둘째 바로 앞에서 언급한 DDL을 참고하여, 코퍼스의 교육적 활용시도가 필요하다. 외국어 교육 분야에서는 코퍼스를 기반으로 한 다양한 교실활동을 통해 학습자들의 목표언어 능력 향상을 시도하여 그 효과도 검증된 바 있다. 예를 들어, 외국어 교육 분야에서는 살아있는 교수 학습 자료의 활용과 학습과정

을 학습자가 조절하는 학습자 중심의 교육 그리고 연구자의 자세로 학습자가 적극적으로 수업활동에 참여(Braun 2008; Huang 2008; Hunston 2002; Johns 1994; O'Keefe, McCarthy, & Carter 2007)라는 장점을 적극 활용해 컴퓨터 기반의 DDL을 교실현장에서 활용하고 있다. 코퍼스 기반의 수업활동을 만드는 것이 번거롭고 어렵게 여겨질 수도 있지만, 앞서 언급한 학습자들의 적극적인 수업활동 참여, 특히 연구자적인 관점에서 주어진 데이터를 직접 주도면밀하게 관찰하여, 특정 영역에 자주 사용되는 어휘, 표현, 더 나아가 전문용어를 추출해내는 발견학습(discovery learning)을 통해 학습효과가 높아질 수 있기에 충분히 시도할만한 가치가 있다.

참고문헌

유정주 (2013)「법률 번역에서 DIY 코퍼스 활용사례」,『번역학연구』 14(2): 149-186.

조의연 (2009)「병렬 말뭉치에 기반한 번역학 연구」,『번역학연구』 10(2): 207-246.

정호정 (2003)「코퍼스 중심의 번역학 연구」,『번역학연구』 4(2): 89-115.

최승권· 김영길 (2010)「번역 말뭉치로부터 추출한 어휘 번역패턴의 의미 분류와 자동번역시스템에의 활용」,『번역학연구』 11(3): 277-301.

최정아 (2003)「병렬 말뭉치를 통한 한국어-영어의 번역 단어수 연구」,『번역학연구』 4(2): 89-115.

황은하 (2013)「말뭉치에 기반한 한중 뉴스표제의 문장부호 번역 연구」,『번역학연구』 14(2): 283-311.

황은하 (2009)「한중 인터넷 신문 기사 표제 병렬말뭉치 연구 －기계번역을 위한 시험적 연구」,『번역학연구』 10(3): 217-245.

Baker, Mona (1992) *In Other Words: a Coursebook on Translation*, Routledge, London.

_____ (1993) 'Corpus Linguistics and Translation Studies: Implications and Applications', In Baker, Mona; Francis, Gill; Tognini-Bonelli, Elena

(Eds), *Text and Technology: In Honor of John Sinclair*. Amsterdam/Philadelphia: John Benjamins, 233-250.

Berber-Sardinha, Tony (2000) 'Comparing Corpora with WordSmith Tools: How Large Must the Reference Corpus Be?', *Proceedings of the Workshop on Comparing Corpora* 9, 7-13.

Boulton, Alex (2010) 'Data-Driven Learning: Taking the Computer out of the Equation', *Language Learning* 60(3), 534-572.

Braun, Sabine (2008) 'From Pedagogically Relevant Corpora to Authentic Language Learning Contents', *ReCALL* 17(1): 47-64.

Cook, Guy (1998) 'The uses of reality: A reply to Ronald Carter', *ELT Journal* 52(1), 57-64.

Gouadec, Daniel (2007) *Translation as a Profession*, Amsterdam: John Benjamins.

Huang, Li-Shih (2008) 'Using Guided Corpus-Aided Discovery to Generate Active Learning', *English Teaching Forum*, 46(4): 20-27.

_____ (2011). Corpus-Aided Language Learning, *ELT Journal* 65(4): 481-484.

Hunston, Susan (2002) *Corpora in Applied Linguistics*, Cambridge, England: Cambridge University Press.

Johns, Tim (1994) 'From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-Driven Learning', In T. Odlin (Ed.), *Perspectives on Pedagogical Grammar* (pp. 293-313). Cambridge, England: Cambridge University Press.

Kubler, Natalie (2003) 'Corpora and LSP Translation', In Zanettin, Federico; Bernardini, Silvia; Stewart, Dominic (Eds), *Corpora in Translator Education*. Manchester: St. Jerome Publishing, 101-112.

Mahadi, Tengku, Tengku Sepora, Helia Vaezian and Mahmoud Akbari (2010) *Corpora in Translation. A Practical Guide*. New York & Bern: Peter Lang.

Maia, Berlinda (2005) 'Terminology and Translation – Bringing Research and Professional Training Together through Technology', *Meta* (50)4: 1079-98.

Manca, Elena (2011) Corpus Linguistics and Cultural Studies: a Combined Approach in the Translation Process' In M. Bondi (Ed.), RILA, Rassegna Italiana di Linguistica Applicata.

O'Keefe, Anne, Michael McCarthy and Ronald Carter (2007) *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge, England: Cambridge University Press.

Olohan, Maeve (2001) 'Spelling out the Optionals in Translation: A Corpus Study', *UCREL Technical Papers* 13: 423-432

Partington, Alan (2011) 'Corpus Linguistics: What It is and What It Can Do', In Cultus: *the Journal of Intercultural Mediation and Communication* 4, D. Katan, E. Manca and C. Spinzi (eds), Terni: Iconesoft, pp. 35-58.

Schmidt, Richard (1990) 'The Role of Consciousness in Second Language Learning', *Applied Linguistics* 11(2), 129-58.

Schmitt, Diane and Norbert Schmitt (2005) *Focus on Vocabulary: Mastering the Academic Word List*, London: Pearson.

Sinclair, John (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stubbs, Michael (1996) *Text and Corpus Analysis*, London: Blackwell.

Tognini-Bonelli, Elena (2001). *Corpus Linguistics at Work, Amsterdam and Philadelphia*: John Benjamins.

Yao, Xing (2008) 'A Corpus-Based Translation Class', In Xiao, Richard, Lianzhen He and Ming Yue (Eds), Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies 12 (UCCTS, 2008). Retrieved Sept 20, 2013, from http://www.lancs.ac.uk/fass/projects/corpus/UCCTS2008Proceedings.

Zanettin, Federico (1998) 'Bilingual comparable corpora and the training of translators', *Meta* 43(4), 616-630.

[Abstract]

Corpus-based Term Extraction Methods for Translator Training

Park, Myongsu
(Sangmyung University)

This paper reports on how to extract terms from a small specialized corpus of Korean Weather Corpus (KWC). The KWC was built from three different sets of data: the Korea Herald, the Korea Times, and Arirang News and its size was 88,042 tokens. It is more than true that the developments in computer technology have made tremendous contribution to the widespread use of corpus in various disciplines and its effects are also felt in the field of the translation studies as well. As part of efforts of encouraging the use of corpus and the corpus-based analytic approaches, the present research aimed at making use of two corpus-based approaches in extracting terms. The first method was using "a list of stopwords" which mainly consists of grammatical function words such as articles and prepositions. By filtering out these words prior to making a list of most frequent words in the KWC, it was made possible to create a list of words that were almost all term candidates. The second one was based on "a keyword analysis." Keywords are those whose frequency is unusually high in comparison with a reference corpus. These unusually high frequent words can represent the aboutness of a given text and reveal some salient features related to a genre. The method also provided a list of positive keywords, which can result in a good list of term candidates of KWC. The suggested methods, hopefully, can serve as alternative ways of extracting terms and contribute to the widespread us of corpus in the translation study.

▸ Key Words: corpus, concordancer, term, extraction, translator training

박명수
상명대학교 영어영문학과 교수
myongsu@smu.ac.kr
관심분야: 코퍼스와 통번역, 제2언어습득론, 영상번역

논문투고일: 2014년 1월 20일
심사완료일: 2014년 2월 15일
게재확정일: 2014년 3월 12일