번역학연구
2014년 가을 제15권3호

다차원통계분석법을 활용한 번역보편소 사례연구*

이 창 수
(한국외국어대)

1. 연구 목적 및 배경

번역연구에 코퍼스언어학 기법이 본격적으로 도입된 것은 1990년대 초로 당시 베이커(Baker 1993, 1995, 1996)는 명시화, 단순화, 규범화, 균일화 등 4가지 번역보편소를 제안하면서 연구방법으로 코퍼스 활용의 필요성을 역설하였다. 그 후 전 세계적으로 번역보편소에 관한 다양한 연구가 진행되면서 코퍼스 기반 연구가 번역연구의 주류를 이룰 정도로 크게 활성화되었다.

그러나 지금까지의 번역보편소 연구는 크게 두 가지 문제를 안고 있다. 첫째는 번역보편소의 실체에 대한 논란이다. 번역보편소에 반기를 든 대표적 학자로는 하우스(House 2008)를 들 수 있다. 하우스(2008: 11)는 번역문에 나타나는 현상은 해당 번역 언어 쌍에 국한되는 것이라며 언어보편소는 있을 수 있지만 번역보편소는 있을 수 없다고 단언하였다. 이에 덧붙여 일부 학자는 명시화

* 본 연구는 2014년도 한국외국어대학교 교내연구비 지원을 받아 작성되었음.

같은 번역보편소 가설에 대하여 실체가 없는, 입증되지 않은 가설이라며 폐기를 주장하기도 한다(Becher 2010). 또한 단순화의 예로 연구되어 온 평균 문장길이의 경우는 연구마다 결과가 상이하여 단순화의 언어 표지가 되지 못한다고 보는 학자도 있다(Xiao 2010: 20-22).

둘째는 연구방법상의 문제이다. 지금까지 번역보편소 연구의 대부분은 특정 언어현상을 개별적으로 분석하였을 뿐 여러 언어현상을 통합적으로 분석한 경우는 드물다. 자오(2010)의 경우처럼 하나의 연구에서 어휘밀집도, 평균문장길이, 접속사 발생건수, 수동태 발생건수처럼 여러 언어현상을 분석한 경우라고 해도 각 언어현상을 개별적으로 통계 분석하는 데 그쳤을 뿐 전체를 통합적으로 분석하지 못했다. 그러나 모든 언어현상이 그렇듯이 번역문과 비번역문의 관계도 다차원적이다. 따라서 연구방법도 여러 언어현상을 동시에 분석할 수 있는 다차원적인 데이터분석법을 활용할 필요가 있다(De Sutter et al. 2012: 327).

이와 같은 배경에서 본 연구에선 한영 문학번역 코퍼스와 비번역 영어 문학코퍼스를 대상으로 다차원 데이터분석법 중 하나인 주성분분석법을 활용한 사례연구를 실시하였다. 본 연구에서 사용한 번역보편소 분석변수는 문헌에서 단순화의 언어 표지로 연구되어 온 평균어휘유형비율(STTR)과 평균문장길이, 그리고 명시화의 예로 언급되어 온 접속사 사용비율과 say동사 뒤의 that 접속사 명시화 등 4가지이다. 단순화란 용어 그대로 번역 언어를 단순화 하는 경향을, 명시화란 번역에서 의미나 표현을 더 구체화하는 경향을 일컫는다(Baker 1996: 180-83). 사례 연구의 일차적 목표는 분석 코퍼스에서 상기 4가지 언어지표가 번역문과 비번역문를 구분하는 기준이 될 수 있는지를 검증하는 것이지만, 보다 궁극적 목표는 사례 연구를 통하여 코퍼스 기반 번역연구에서 다차원 데이터분석법의 필요성을 제기하는 것이다.

2. 이론적 배경

2.1. 번역보편소 정의와 문제점

번역보편소는 베이커(1993: 245)가 처음 제안한 개념으로 번역문과 비번역문을 구분 짓는 번역문의 "내재된" 특징을 일컫는다. 같은 맥락에서 체스터만(Chesterman 2004: 3)도 번역보편소를 모든 번역문에서 나타나는 특징이라고 설명하였다. 부연설명하면 번역문에서 발견되는 어떤 특징이 번역보편소로 인정받으려면 번역 언어 쌍, 번역 언어 방향, 텍스트유형, 번역사, 번역문의 생성 시기 등 외적 변수의 영향을 받지 않고 모든 번역문에서 항상 존재하는 특징이어야 한다.

이와 같은 관점에서 앞서 언급하였듯이 베이커(1993, 1995, 1996)는 명시화, 단순화, 규범화, 균일화 등 4가지 번역보편소를 제안하였으며, 체스터만(2004)은 원문과 번역문간의 관계에 적용되는 s-보편소 9가지와 번역문과 비번역문간이 관계에 적용되는 t-보편소 4가지를 제안하였다.

이와 같은 배경 하에서 그동안 연구 대상이 되었던 번역보편소를 들자면 명시화(Chen 2006; Laviosa-Braithwaite 1995; Olohan 2003; Olohan & Baker 2000; Øver s 1998), 단순화(Laviosa-Braithwaite 1996, 1997; Malmkjær 1997), 규범화(Kenny 2000, 2001; Teich 2001, 2003; Toury 1995: 102-12), 균일화(Laviosa-Braithwaite 1996), 과소표현(Eskola 2000), ST 간섭현상 또는 비침현상(Teich, 2003) 등이 있다.

그러나 번역보편소는 원래 그런 것이 있을 것이라는 가설에서 출발한 개념으로 그 동안 위에 언급한 것처럼 다양한 번역보편소 가설을 뒷받침하는 연구가 학계에 보고되었지만 여전히 그 존재 여부를 놓고 논란이 일고 있다. 번역보편소에 가장 회의적인 학자는 하우스(2008: 11)로 번역은 활동(performance), 언어 사용(parole) 및 언어사용능력의 문제로 근본적으로 언어 쌍에 따라 특징이 달라질 수밖에 없기 때문에 번역보편소는 존재하지도 않고 존재할 수도 없다고 단언하였다. 또한 하우스(2008: 12)는 번역보편소로 주장되는 현상이 번역 언어 방향이나 장르에 따라 편차가 크다는 점을 지적하면서 번역문의 보편적 특징은 아니라고 주장하였다. 또한 원문, 번역문, 비교문 등은 이들이 속한 장르가 공

시적으로 어떻게 발전하고 어떤 지위를 갖느냐에 따라 큰 영향을 받기 때문에 번역문의 특징도 시대 배경에 따라 달라진다고 주장하였다.

번역보편소에 대한 이와 같은 근본적 문제 제기는 차치하고라도 번역보편소로 주장되는 현상들에 대한 엇갈린 연구결과와 해석은 이들의 실재에 대한 의문을 불러일으킨다. 가령, 푸르티넨(Puurtinen 2003)의 핀란드 아동도서의 영어 번역 연구에선 비정형동사절(non-finite clause)보다 정형동사절의 비율이 높게 나타났다. 동 연구에선 정형동사절은 비정형동사절에 비하여 어휘밀집도가 높고 어휘 반복율이 낮다는 이유로 이런 결과를 단순화 번역보편소에 반하는 것으로 해석하였으나, 베이커(1996: 180)는 유사한 사례에서 정형동사절에서는 시제와 절 관계 등이 명시적으로 드러나기 때문에 독자 이해를 돕는다는 점에서 단순화와 명시화의 표상으로 해석하였다. 이렇듯 같은 현상을 놓고도 번역보편소와의 연관성에 대하여 상이한 해석이 나오는 것은 번역보편소의 개념이 아직 모호하고 불안정하다는 것을 반증한다.

자오(2010: 10)는 연어관계, 어휘사용, 구문구조와 관련된 최근 연구에서 기존 연구의 단순화 가설을 뒤집는 결과가 나옴에 따라 단순화 가설은 논란거리라고 지적하였다. 같은 맥락에서 자오(2010: 20)는 번역중국어와 비번역중국어 간의 평균문장 길이를 비교하였는데 번역문의 평균문장길이가 짧다는 기존 주장과 달리 큰 차이가 없는 것으로 나타나서 평균문장길이는 단순화와 관계가 없다고 평가하였다.

베커(2010)는 명시화가 번역보편소 연구에서 독선적인 지위를 누리고 있다면서 명시화로 거론되는 언어 현상은 다른 가설로도 충분히 설명될 수 있으며 명시화를 뒷받침할 객관적 증거가 부재하기 때문에 명시화 가설은 폐기되어야 한다고 주장하였다. 구체적으로 베커는 명시화의 고전적 연구로 받아들여지는 올로한과 베이커(Olohan & Baker 2000)의 ‘say와 ’tell’동사에 붙는 ‘that’ 접속사 연구와 외베라스(Øver s 1988)의 영어-노르웨이어 문학번역에서의 명시화 연구의 결과에 대해 원문간섭현상으로 설명이 가능하다며 날카로운 비판을 가했다.

2.2. 다차원 탐구적 데이터분석법

번역보편소의 연구 방법론적 측면에선 단일 언어현상에 대한 단순한 발생빈도나 통계유의성 검증에 의존하는 방식의 한계점이 지적되고 있다. 젠셋과 맥길리브레이(Jenset and McGillivray 2012: 302-303)는 베이커나 그 뒤를 이은 대부분의 연구들이 단순 발생빈도 비교와 단편적 예문 분석 등 매우 기초적인 분석방법에 의존하고 있다면서 다변수 통계분석법의 활용 필요성을 강조했다. 디 셔터르 외(2012: 327)도 번역어와 비번역어의 관계는 한 두 언어 현상으로 설명할 수 없는 다차원적 관계라며 여러 변수를 동시에 분석할 수 있는 다차원 분석의 활용을 예시하였다.

이와 같은 배경 하에 젠셋과 맥길리브레이(2012)는 번역영어코퍼스(TEC)에서 '-ly', '-ment'와 같은 접미사 사용과 번역원어, 번역사, 원저자, 텍스트유형과 같은 변수의 상관관계를 분석하였다. 동 저자들은(Jenset & McGillivray 2012: 307-8) 바이버(Biber 1988)가 장르 분석에 사용하여 유명해진 요인분석법(factor analysis)은 데이터 내의 분산을 충분히 설명하지 못하며, 분석결과 해석이 주관적이고, 데이터에 맞는 모델을 찾기 어렵다는 점에서 코퍼스 분석에 적합하지 않다고 평가하였다. 저자들은(Jenset & McGillivray 2012: 309) 그 대안으로 주성분분석법(principal component analysis)을 제안하였는데, 주성분분석법은 요인분석법에 비하여 데이터 내의 분산을 모두 반영하며, 연구자의 주관적 해석 여지를 줄여주고, 변수가 독립적이지 않아도 된다는 점을 장점으로 거론하였다.

디 셔터르 외(2012: 327)는 번역보편소 연구에 다차원 탐구분석법을 적용한 예로 번역문에서 격식체를 선호하는 표현의 보수주의(conservatism) 가설을 시험하는데 프로파일 기반 주성분분석법을 활용하였다. 해당 연구는 불어와 영어에서 네덜란드어로 번역된 번역문과 네덜란드어 비번역문을 분석코퍼스로 사용하였으며, 10쌍의 격식-비격식 어휘를 선정하여 두 대안 간의 선택에 있어 텍스트유형과 번역원어가 미치는 영향을 조사하였다.

상기 연구에서 소개된 요인분석법이나 주성분분석법은 피어슨 상관계수분석, t-테스트, ANOVA, 회귀분석 등과 같이 데이터 내 변수 간 차이나 상관관계의 통계 유의성을 테스트하는 확증적 데이터분석법(confirmatory data analysis)과 달리 데이터 내 여러 변수 간에 어떤 상관관계가 있는지를 탐구하는 탐구적 데

이터분석법(explanatory data analysis)이다(Jenset & McGillivray 2012: 304). 이에 대한 논의는 다음 장의 사례연구에서 분석기법으로 채택한 주성분분석법을 설명할 때 좀 더 언급하기로 한다.

3. 사례연구

3.1. 분석 목적과 데이터

본 사례연구에선 영어로 번역된 한국소설과 순수 영어소설을 비교코퍼스로 사용하여 평균어휘유형비율(STTR), 평균문장길이, 접속사 빈도와 say동사 뒤의 that 사용 여부 등 4가지 언어현상이 두 코퍼스를 구분하는 요인이 될 수 있는지를 조사하고자 한다.

어휘유형비율(TTR)은 텍스트 내 어휘유형수를 총 단어수로 나눈 비율인데, 이 비율은 텍스트나 코퍼스의 크기에 따라 크게 달라진다. 이와 같은 문제를 해결하기 위하여 텍스트를 1000단어마다 나눠 구한 TTR에 100을 곱한 퍼센트 값을 가지고 다시 전체 평균을 낸 것을 STTR이라고 한다(Scott 2014: 205-86). 자오(2010: 19)의 중국어 번역문 연구에선 비번역코퍼스와 비교했을 때 STTR에서 큰 차이가 나타나지 않았다. 따라서 STTR이 번역문과 비번역문을 구분해주는 언어지표인지는 여전히 논의의 대상이라고 할 수 있다. 평균문장길이는 문장의 평균 단어 수로 앞서 언급한대로 여러 연구에서 비번역문에서 더 높게 나온 경우가 있어서 단순화의 언어지표로서 논란거리가 되고 있다. 접속사는 ‘However’, ‘Instead’, ‘In addition’과 같이 문장 앞에 붙어 앞뒤문장을 논리적으로 연결하는 역할을 하는 접속부사를 의미하며, 각 텍스트 당 발생건수를 전체 텍스트 단어 수로 나눠 1000을 곱한 값을 사용하였다. 접속사의 경우 자오(2010: 23)의 연구에선 비번역문에 비하여 번역중국어에서 접속사 빈도가 높게 나타났다. 특히 다른 장르에 비하여 소설 장르에서 두드러지게 높았다. 따라서 과연 한영 소설번역문에서도 같은 결과를 발견할 수 있는지 확인해 볼 가치가 있다. 마지막으로 ‘say’ 동사에 붙는 ‘that’ 접속사 명시화는 전체 텍스트에서 ‘that’가 붙은 경우를 그렇지 않은 경우의 수로 나눠서 100을 곱한 수를 사용하

였다. 앞서 언급한 올로한과 베이커(2000)의 연구에서 ‘say’와 ‘tell’ 뒤에 ‘that’을 명시화하는 비율이 비번역영어보다 번역영어에서 더 높게 나타났다.

이와 같은 4가지 언어지표를 가지고 분석한 데이터는 영어로 번역된 한국소설 21종과 순수 영어 소설 21종이다. 영어로 번역된 한국소설은 주로 60, 70년대 한국문학 대표작품들로 2000년대에 영어로 번역이 되었다. 내용은 주로 사회상, 가정사, 연애, 인간관계를 다룬 일반 소설이다. 이와 비교할 순수영어 소설은 1990과 2000년대에 발표된 현대소설이다. 영한번역코퍼스와 어느 정도 내용을 일치시키기 위하여 역사, 공상 과학, 판타지 같은 특정한 장르의 작품은 배제하였다. 한영번역코퍼스는 총 단어수가 1,449,422이며 비교영어코퍼스는 1,878,156으로 비교코퍼스가 더 크다. 이는 21종이라는 작품 수를 맞추다 보니 나타난 결과이다. 그러나 본 연구에서 사용한 분석지표들이 앞서 설명하였듯이 모두 표준화(평균치나 퍼센트)된 수치이기 때문에 양 코퍼스 간의 크기 차이는 큰 문제가 되지 않는다.

3.2. 데이터 분석과 결과

앞서 언급하였듯이 본 연구의 통계분석은 주성분분석법을 사용하였으며 모든 통계작업은 R 통계프로그램을 통해 이뤄졌다. 본 사례연구에서 사용된 분석기법은 기본적으로 베이언(Baayen 2008)의 논의에 기초하였으며 앞서 언급한 젠셋과 맥길리브레이(2012)와 디 셔터르 외(2012)의 논문을 참고하였다.

먼저 R 통계프로그램에 탑재된 데이터프레임의 구조부터 살펴보도록 한다. R에서 19열에서부터 25열까지의 분석데이터를 불러내면 <표 1>과 같다. 맨 왼쪽 행의 일련번호는 작품번호로 일부만 나와 있지만, 실제 데이터프레임에는 번역작품 t1부터 t21까지, 번역작품 n_1부터 n_21까지 나열되어 있다. 그리고 횡으로는 STTR, mSen(평균문장길이), Cnn(접속사발생비율), That(that발생비율), Text(TR=번역, NT=비번역) 등 5개의 분석변수가 배열되어 있다.

〈표 1〉 분석 데이터프레임

	STTR	mSen	Cnn	That	Text
t19	45.45	14.42	9.14	43	TR
t20	43.19	12.38	17.00	29	TR
t21	44.05	11.87	12.78	0	TR
n_1	41.58	10.16	3.64	0	NT
n_2	47.78	10.18	5.23	0	NT
n_3	49.65	12.22	6.51	0	NT
n_4	36.07	15.93	3.23	12	NT

〈표 2〉 분석데이터 구조

```
data.frame': 42 obs. of  5 variables:
$ STTR: num  42.2 43.3 45.2 43.5 43.5 ...
$ mSen: num  13.1 10.6 11 10.2 11.1 ...
$ Cnn : num  11.92 6.82 5.55 13.67 10.42 ...
$ That: int  0 10 19 4 0 8 0 45 19 13 ...
$ Text: Factor w/ 2 levels "NT","TR": 2 2 2 2 2
```

<표 2>는 분석 데이터의 구조를 보여주는 또 다른 표이다. 이를 보면 분석 데이터는 총 42개의 텍스트(obs. observation의 준말)와 5개의 변수로 구성되어 있다는 것을 알 수 있다. 그리고 STTR, mSen, Cnn, That 변수는 숫자변수(num=number, int=integer)이고 Text는 2개의 레벨로 구성된 범주형 변수(Factor)이다.

베이커(1996: 181-3)가 처음 제시한 대로 STTR과 mSen이 단순화 보편소를 보여주는 언어지표가 되려면 비번역문에 비하여 번역문의 값이 더 적어야 한다. 또한 올로한과 베이커(2000) 그리고 자오(2010)의 연구결과에 따르면 명시화 지표로서의 Cnn와 That은 비번역문에 비하여 번역문에서 값이 더 커야 한다. 따라서 이와 같은 가설이 맞다면 STT와 mSen, 그리고 Cnn과 That 사이에는 각각 정의 상관관계가 성립되어야 하고 두 쌍 간에는 역의 상관관계가 성립되어야 한다. 이를 확인하기 위하여 데이터 중 번역문만 뽑아서 4 변수 간의 상관계수행렬을 구해보면 <표 3>과 같다. 이를 보면 STTR과 mSen는 피어슨 상관계수가 －0.05로 미약한 수치지만 예상과 반대로 역의 상관관계를 보여주고 있다. Cnn와 That도 －0.10으로 역의 상관관계를 갖고 있다. 이는 본 데이터에선 4개의 분석변수 중 일부는 번역보편소로서 문제가 있을 수 있음을 시사한다.

참고로 피어슨 상관계수행렬은 주성분분석의 기본 데이터가 된다.

〈표 3〉 번역코퍼스의 4 변수 간 상관행렬

	STTR	mSen	Cnn	That
STTR	1.00	-0.05	0.31	-0.20
mSen	-0.05	1.00	0.13	0.24
Cnn	0.31	0.13	1.00	-0.10
That	-0.20	0.24	-0.10	1.00

그러면 본격적인 주성분분석으로 들어가 보자. 주성분분석의 주목적은 다변수 데이터의 변수를 줄여 그래프로 보여줌으로써 변수와 개별 측정치 간의 상관관계를 보다 직관적으로 쉽게 파악할 수 있도록 하는 것이다. 본 연구의 분석데이터는 5개의 변수가 있으므로 5차원 데이터라고 할 수 있다. 이와 같이 5개의 변수가 서로 얽혀있는 상황에서 변수 간 그리고 변수와 개별 텍스트 간의 관계를 파악하는 것은 매우 난해한 일이다. 보통 변수가 10개, 20개씩 되는 설문조사의 경우는 더욱 그러하다. 변수가 많으면 변수 간에 상관관계가 존재하게 마련이다. 변수 간에 상관관계가 있다는 것은 해당 변수들이 결국 같은 관계를 반복적으로 반영한다는 것을 의미한다. 이런 반복적 변수를 통합하여 상관관계가 없는 소수의 새로운 설명변수를 찾아내는 것이 주성분분석이며, 그와 같은 새로운 설명 변수를 성분(component)이라고 부른다. 이론적으로 성분의 수는 변수의 수와 일치하지만 그 중 2, 3개의 주성분이 대부분의 데이터를 설명하기 때문에 나머지는 무시할 수 있게 된다.

실제 주성분분석을 통해 성분을 찾아내는 과정은 데이터를 표준화하고(scaling), 공분산이나 상관행렬을 구하고, 이를 활용해서 고유값(eigen value)과 고유변수(eigen vector)를 계산하고 새로운 고유변수에 대한 각 텍스트의 좌표를 구해서 그래프로 표시하는 등 다소 복잡한 계산과정을 거치지만 R에는 이와 같은 계산을 대신 수행해주는 기능이 있다. 가장 대표적인 것이 R에서 기본적으로 제공하는 prcomp과 FactoMineR 패키지에서 제공되는 PCA 명령어이다. 본 연구에선 아래 R 입력코드에 예시한 대로 PCA 기능을 사용하여 주성분분석을 실행하였다. 아래 코드에서 ‘PCA’는 주성분분석 명령어이고 ‘data’는 R에 탑재된 데이터 명이며, ‘quali.sup=5’는 데이터 중 5번째 범주변수 ‘Text’를 데이터를 설명

하는 보조변수로 사용한다는 것을 의미한다. 이와 같이 하여 data.pca라는 새로운 개체를 생성하는데 여기에 분석결과가 담겨있다. FactoMineR의 PCA기능은 기본적으로 데이터를 표준화하기 때문에 본 연구의 데이터와 같이 표시단위가 다른 변수를 분석하는데 용이하다. 여기서 표준화란 말은 PCA의 입력데이터로 공분산행렬이 아니라 상관계수행렬을 사용한다는 뜻이다.

```
data.pca<-PCA(data, quali.sup=5)
```

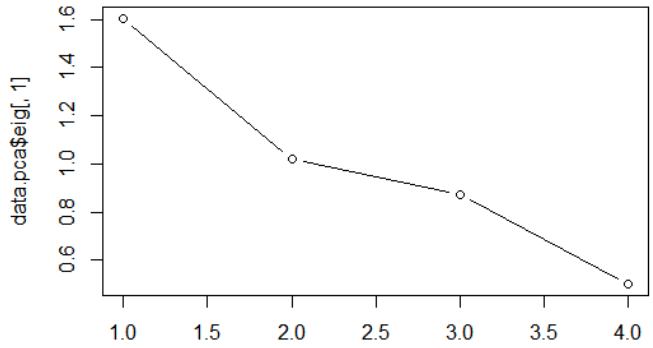
R에서 'summary( )' 기능을 활용하여 PCA 결과를 불러오면 가장 먼저 <표 4>와 같이 고유값과 관련된 분석결과가 제시된다. 여기서 Dim.1, Dim.2 ...는 주성분분석을 통해 찾아낸 새로운 설명변수로 Dimension 1, Dimension 2을 의미하며, 이것이 바로 주성분(PC: principal component)에 해당한다. Variance는 해당 주성분의 고유값을 나타내고 % of var는 해당 주성분이 데이터의 분산을 설명하는 비율, 그리고 Cumulative % of var는 각 주성분의 % of var을 순차적으로 합친 총 비율이 된다. <표 4>를 보면 Dim.1은 데이터 분산도의 40.2%를 대변하고, Dim.2는 25.5%를 대변하며 두 주성분을 합친 비율은 65.7%가 된다.

〈표 4〉 주성분분석 결과 (1)

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4
Variance	1.606	1.020	0.873	0.501
% of var.	40.150	25.508	21.828	12.514
Cumulative % of var.	40.150	65.658	87.486	100.000

〈그래프 1〉 주성분 비율 그래프



이와 같은 PCA 분석결과를 보고 데이터 분석에 몇 개의 주성분을 사용할 것인가를 선택해야 하는데 보통 고유값이 1이 넘는 주성분을 선택한다. 주성분 고유값을 그래프로 그리면 <그래프 1>과 같다. 이런 그래프를 스크릿 그래프(screet plot)라고 하는데, 선이 급격하게 꺾여서 평평해지는 부분(x축 상 2.0)을 주성분 선택 지점으로 잡는다. 본 연구에선 두 기준에 따라 Dim.1과 Dim.2을 데이터 해석에 사용할 주성분으로 선택하기로 한다.

〈표 5〉 각 분석변수가 성분에 기여하는 비율

```
$contrib
        Dim.1     Dim.2     Dim.3      Dim.4
STTR 13.81237 28.051196 57.570574  0.5658623
mSen 47.19547  5.007421  5.973835 41.8232711
Cnn  22.61872 33.741706  5.464963 38.1746146
That 16.37344 33.199678 30.990629 19.4362520
```

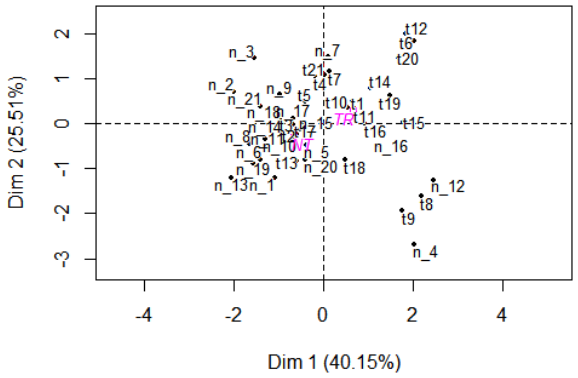
그러면 Dim.1과 Dim.2 두 주성분을 축으로 하여 4가지 분석변수와 번역 및 비번역코퍼스가 어떤 관계를 갖고 있는지를 분석해 보자. 먼저 <표 5>를 보면 4가지 변수가 주성분의 구성에 기여하는 비율이 나와 있다. Dim.1을 보면 mSen의 구성비율은 47.2%, Cnn은 22.6%로 mSen이 가장 큰 대표성을 갖고 있다. 이에 비하여 Dim.2는 Cnn과 That가 33.7%와 33.2%로 비슷한 비율을 보이고 있고 다음으로 STTR이 28%의 구성비를 보여주고 있다.

주성분분석의 핵심은 분석결과를 그래프로 그려서 각 변수와 개별 텍스트 간의 관계를 시각적으로 분석하는 것이다. R에서 PCA 명령어를 실행하면 자동적으로 <그래프 2>, <그래프 3>과 같은 두 개의 그래프가 생성된다. 먼저 <그래프 2>를 보면 PCA를 통해 새로 생성한 두 성분 축이 구성하는 2차원 공간에 개별 텍스트들이 배열되어 있다. 이 그래프에서 x축이 Dim.1에 해당하고 y축이 Dim.2에 해당하며, 축명의 괄호 안의 퍼센트는 각 축이 원래 데이터내의 분산을 설명하는 비율이다.

<그래프 2> 상의 개별 텍스트의 좌표는 <표 6>에 나와 있다. 여기에는 총 42개의 텍스트 중 데이터 프레임 상 처음 7개의 번역문에 관한 정보만 나와 있다. 이 표에는 각 텍스트 별로 구름집단(cloud)의 중심까지의 거리(Dist), 각 Dim별로 텍스트 개별좌표와 Dim형성에 기여하는 기여도(ctr), 그리고 각 Dim

에 대표되는 품질 평가 수치인 코사인 제곱값(cos2)이 담겨있다. 예들 들어 t1번 텍스트의 경우 Dim.1의 좌표값은 1.528, Dim.2 좌표값은 0.489로 이를 <그래프 2>에 맞춰보면 우상단에 있는 t1번 텍스트의 위치와 일치하는 것을 알 수 있다.

〈그래프 2〉 개별 텍스트 그래프



〈표 6〉 개별 텍스트 좌표와 주성분 기여도

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2
t1	1.528	0.636	0.600	0.173	0.489	0.558	0.102
t2	0.668	-0.598	0.530	0.801	-0.238	0.133	0.127
t3	1.153	-0.672	0.670	0.340	-0.005	0.000	0.000
t4	1.647	-0.194	0.056	0.014	1.059	2.619	0.414
t5	1.129	-0.397	0.234	0.124	0.509	0.605	0.203
t6	3.075	2.010	5.987	0.427	1.834	7.845	0.356
t7	1.605	0.125	0.023	0.006	1.169	3.188	0.531

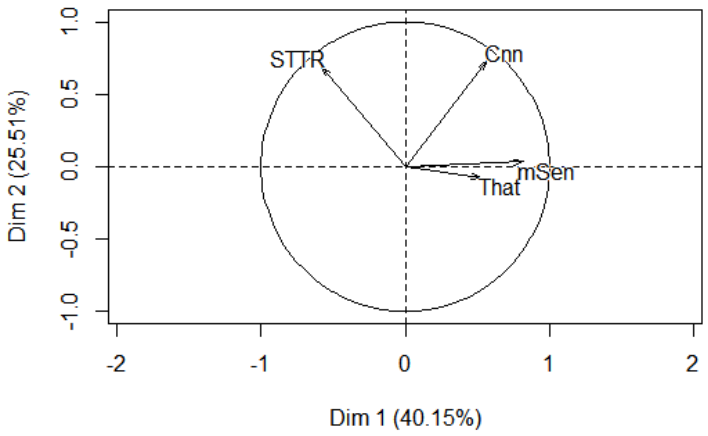
〈표 7〉 개별 변수 좌표와 주성분 기여도

Variables

	Dim.1	ctr	cos2	Dim.2	ctr	cos2
STTR	-0.583	21.192	0.340	0.689	46.584	0.475
mSen	0.818	41.705	0.670	0.034	0.113	0.001
Cnn	0.565	19.847	0.319	0.734	52.744	0.538
That	0.526	17.255	0.277	-0.075	0.558	0.006

숫자와 글자가 겹쳐 잘 보이지 않지만 <그래프 2>의 중앙에는 NT와 TR이란 글자가 적혀 있는데, NT는 비번역문, TR은 번역문의 평균값이 위치한 장소이다. 두 글자가 어느 정도 떨어져 있기 때문에 분석변수 중에 일부에서 양 코퍼스 간에 차이가 있는 것을 확인할 수 있다. 번역문을 의미하는 t숫자와 비번역문을 의미하는 n숫자의 분포를 보면 t는 주로 그래프 우상단에 n은 좌하단에 몰려있는 것이 보인다.

〈그래프 3〉 개별 변수 그래프

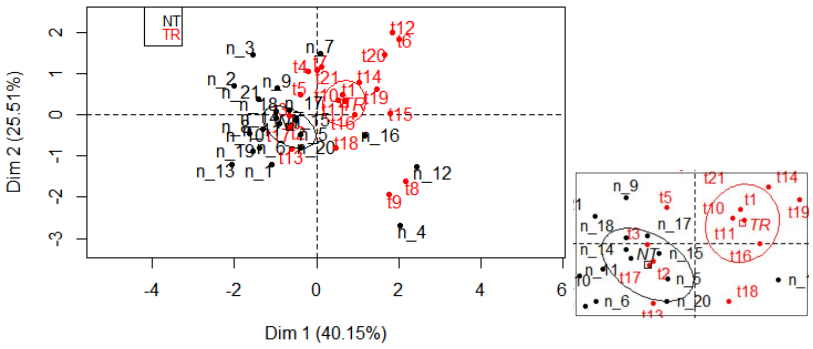


<그래프 3>의 변수 그래프도 주성분분석 결과에 포함된 <표 7>의 좌표값에 따라 그려진 것이다. 이 그래프를 보면 mSen은 Dim.1축과 거의 일치하고 있어 양자 간에 상관관계가 매우 높음을 보여준다. 변수 선의 길이는 상관관계의 강도를 대변한다. Cnn도 Dim.1축과 상관관계가 있는 것을 알 수 있지만 벡터선의 길이로 볼 때 강도는 상대적으로 약하다. 이에 반하여 STTR과 Cnn은 Dim.1과 Dim.2의 중간쯤에 위치해 있지만 Dim.2로 약간 더 기울어 있어 Dim.2와 상관관계가 더 높음을 보여준다.

<그래프 4>는 <그래프 2>를 좀 더 개선한 것으로 번역문과 비번역문이 색으로 구분되며, 중간에 95퍼센트 신뢰타원(95 confidence ellipse)이란 타원이 그려져 있다. 95퍼센트 신뢰타원은 일반 통계의 95퍼센트 신뢰구간(95%

confidence interval)과 비슷한 것으로 두 타원이 겹치지 않는 경우에 통계적으로 유의한 차이가 있는 것이 된다. <그래프 4> 상에서 두 타원이 우상단과 좌하단으로 완전히 떨어져 있어 해당 대각선방향으로 두 코퍼스의 평균치에서 유의미한 차이가 있음을 알 수 있다.

〈그래프 4〉 95%신뢰타원 및 NT와 TR



데이터 상에서 4가지 분석변수와 번역 및 비번역코퍼스 간의 상호 관계를 분석하려면 <그래프 3>과 <그래프 4>를 겹쳐보면 된다. <그래프 4>에선 번역과 비번역코퍼스가 벌어진 대각선 방향과 Cnn의 벡타선의 방향이 거의 일치하고 있다. 좀 더 자세히 보면 번역코퍼스는 Cnn과 같은 방향으로 늘어져 있어 정의 상관관계를 갖고 있는데 반하여, 비번역코퍼스는 반대방향에 포진해 있어 역의 상관관계를 보여준다. 이는 번역코퍼스에서 Cnn이 늘어나면 비번역코퍼스에선 반대로 줄어든다는 뜻이며, 해당 방향으로 95퍼센트 신뢰타원이 겹치지 않기 때문에 양 코퍼스는 평균치에서 통계적으로 유의미한 차이가 있음을 알 수 있다.

Dim.1축과 가깝게 있는 mSen과 That 변수도 마찬가지이다. Dim.1축을 따라 번역과 비번역코퍼스가 좌우로 나뉘고 있기 때문에 양 코퍼스는 mSen과 That에서도 서로 반대되는 상관관계에 있고, 95퍼센트 신뢰타원이 겹치지 않기 때문에 양 코퍼스 간의 평균치의 차이도 통계적으로 유의미할 것임을 보여준다. 이에 비하여 STTR의 벡터선 방향으로는 두 코퍼스를 구별할 수 없다. 이 선을

따라 양쪽에 양 코퍼스의 텍스트들이 비슷하게 늘어서 있고, 95퍼센트 신뢰타원도 해당 벡타선을 따라 겹치기 때문에 큰 차이가 나지 않는다.

이상의 탐구적 데이터분석의 결과는 <표 8>과 같이 간단한 t-검정으로 확인이 가능하다. 주성분분석 그래프에서 분석하였듯이 mSen, Cnn, That의 경우는 번역문과 비번역문 간에 통계적으로 유의미한 평균치의 차이가 존재한다. 특히 Cnn의 경우는 통계적 유의성이 매우 크게 나타났다. STTR은 앞서 분석한대로 비번역문이 번역문에 비하여 평균치에선 미세하게 크지만 통계적으로 유의미하지 않는 것으로 나타났다.

〈표 8〉 4변수에 대한 번역. 비번역 평균차이 t-검정

분석 변수	평균값	t-검정 결과
	번역문 비번역문	
STTR	42.9 < 43.5	p-value=0.5064(t=0.672, df=32.076)
mSen	12.4 >10.9	p-value=0.01436(t=-2.5654, df=38.131)*
Cnn	10.4 > 6.1	p-value=0.0004786(t=-3.8406, df=36.006)***
That	17.8 > 8.0	p-value=0.02178(t=-2.4053, df=33.847)*

이상의 분석결과를 놓고 보면 That과 Cnn 변수의 경우는 비번역문에 비하여 번역문에서 접속사사용 빈도와 'that' 명시화 비율이 더 높을 것이라는 자오(2010: 23)와 올로한과 베이커(2000)의 연구와 일치한다. 즉, 원어영어소설에 비하여 한국어에서 영어로 번역한 소설에서 접속사를 더 많이 사용하고 say동사 뒤에 'that'을 명시하화는 경향이 통계적으로 유의미할 정도로 크다. 이와 같은 결과는 명시화란 변역의 특징일수도 있지만 원문의 영향도 배제할 수 없다. 원문에서 접속사를 많이 쓴다면 번역문에서도 그만큼 접속사를 많이 쓸 수밖에 없을 것이기 때문이다. 이에 반해 'say'동사 뒤에 'that'을 명시하는 것은 한국어와 영어간의 언어 구조적 차이 때문에 번역사의 판단에 따른 결과로 볼 수 있다. 다만 앞서 언급했듯이 베커(2010: 11)는 프랑스어나 이탈리아어에서처럼 인용동사 뒤에 'that'과 같은 접속사를 의무적으로 붙여야 하는 경우 원어의 간섭현상이 일어날 수 있다고 했는데 영어와 같은 어족이 아니더라도 한국어에서

도 인용동사 앞에 '(-다)고'와 같은 인용조사를 의무적으로 붙여야하기 때문에 이것이 'that' 명시화에 영향을 미쳤을 가능성, 즉 원문간섭현상의 가능성을 완전히 배제할 수 없다.

다음으로 mSen에서 번역문이 더 높게 나온 것은 비번역문에 비하여 번역문의 문장이 짧다는 단순화 가설과 배치된다. <그래프 4>를 보면 mSen과 거의 일치하는 Dim.1 축의 영점을 기점으로 mSen방향으로는 비번역문이, 그 반대방향에 번역문이 존재한다. 원어영어소설이 번역소설에 비하여 더 긴 문장을 사용한다는 의미이다. 특히, mSen 벡터선 상에 가장 오른쪽에 위치한 t8, t9와 같은 번역소설은 평균문장길이가 14.12와 14.37로 전체 코퍼스 평균치인 11.7에 비하여 매우 크다. 이에 반하여 반대쪽 끝에 위치한 원어소설인 n13의 평균문장길이는 7.8로 매우 짧다. 따라서 자오(2010: 20-2)가 지적한 대로 평균문장길이는 단순화의 지표가 될 수 없다는 주장이 설득력이 있어 보인다. 이와 같은 분석결과는 원문인 한국소설에서 평균적으로 긴 문장을 쓰거나 도착어인 영어소설에서 짧은 문장을 선호하는 언어와 장르적 특성 때문일 가능성이 크다. 즉 비번역문에 비하여 번역문의 문장이 짧거나 긴 것은 번역보편소 때문이 아니라 언어나 장르의 특성을 반영하는 것으로 설명하는 것이 더 설득력이 있어 보인다.

STTR도 마찬가지이다. 자오(2010: 19)의 번역 중국어 연구에서 STTR은 비번역중국어와 비교하여 차이가 없게 나왔는데 본 연구에서도 같은 결과가 나왔다. 이 역시 STTR도 단순화 가설을 입증하는 언어지표로 부적합한 것임을 보여준다.

4. 결론

3절의 분석결과를 요약하면 명시화의 언어지표로 연구되어 온 접속사와 'say'동사 뒤의 'that' 명시화는 비번역코퍼스에 비하여 번역코퍼스에서 확실히 높게 나타나 타당성이 입증되었다. 그러나 단순화의 언어지표로 사용되어 온 평균문장길이와 STTR은 타당성을 입증하는데 실패하였다. 특히 평균문장길이의 경우는 예상과 반대로 비번역어에서 더 짧게 나타남으로서 평균문장길이는

단순화의 지표가 될 수 없다는 주장이 설득력이 있어 보인다. 결론적으로 자오(2010: 10)가 언급한대로 번역보편소로서 단순화는 현재까지 언급된 언어지표들로는 실증적으로 뒷받침되기 힘들어 보인다.

동 사례연구의 보다 중요한 의미는 코퍼스 기반 번역연구에서 주성분분석법과 같은 다변수, 다차원 데이터분석법의 유용성을 예시하였다는 점이다. 일반적인 발생빈도 비교 통계 검증은 데이터에서 관찰된 차이나 상관관계의 통계적 유의성을 판단할 뿐 데이터 자체를 이해하고 해석하는 데는 큰 도움을 주지 못한다. 이에 반하여 본 연구에서 사용한 주성분분석이나 범주형 변수 분석에 사용하는 대응분석(correspondence analysis), 여러 변수를 상관관계에 따라 나뭇가지형태로 묶어 보여주는 군집분석(cluster analysis)과 같은 다차원 데이터 탐구분석기법은 연구자에게 데이터 내에 존재하는 다양한 관계에 대한 심층 정보와 통찰력을 제공한다. 또한 개별 텍스트가 전체 텍스트 군집 내에 어떤 위치를 차지하는가도 쉽게 파악할 수 있다. 가령, <그래프 4>를 보면 번역문 중 t2과 t3, t20과 t21은 각각 동일 번역가(들)가 번역한 작품이다. 따라서 번역보편소가 사실이라면 이들 번역가들의 작품에선 어느 정도 동일한 경향이 관찰되어야 한다. 이는 <그래프 5>상에서 해당 작품들은 서로 가깝게 위치해야한다는 말과 같다. 그런데 글자가 겹쳐 잘 안보이지만 t2와 t3은 좌하단 중앙에 가깝게 서로 붙어 있다. 그러나 t20과 t21은 우상단에 큰 거리를 두고 떨어져 있다. 평균문장길이만 놓고 봐도 41.91과 46.26으로 큰 차이가 난다. 어떤 이유 때문에 동일번역가 작품 간에 이와 같은 차이가 발생하거나 발생하지 않는 것일까? 이런 문제는 번역보편소 외에 다른 설명을 요구하는 것으로 추가적인 심층 연구의 주제가 될 수 있다.

결론적으로 번역보편소도 모든 언어현상과 마찬가지로 다양한 언어 요소들의 상호작용 속에서 분석하고 이해할 필요가 있다. 본 연구에서 사용한 주성분분석뿐만 아니라 대응분석, 군집분석, 요소분석 같은 통계분석법을 사용하면 여러 변수에 대하여 일관적인 통합 분석이 가능하다. 이런 통계분석은 R 뿐만 아니라 SPSS와 같은 상용통계프로그램에서도 실행할 수 있기 때문에 번역보편소 연구를 포함한 향후 코퍼스 기반 번역연구에서 다차원 데이터분석법의 활용도가 높아지기를 기대한다.

참고문헌

Baayen, R. H (2008) *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R,* Cambridge: Cambridge University Press.

Baker, Mona (1993) 'Corpus linguistics and translation studies: implications and applications.' In M. Baker, G. Francis. & E. Tognini-Bonell (eds.), *Text and Technology: in honor of John Sinclair* 233-50, Amsterdam: John Benjamins Publishing Co.

Baker, Mona (1995) 'Corpora in translation studies: an overview and some suggestions for future research,' *Target 7*(2), 223-43

Baker, Mona (1996) 'Corpus-based translation studies: the challenges that lie ahead' In H. Somers (ed.) *Terminology, LSP and Translation, Studies in Language Engineering in honour of Iuan C. Sager* 75-186, Amsterdam: John Benjamins Publishing Co.

Baker, Mona (2004) 'A Corpus-based view of similarity and difference in translation,' *International Journal of Corpus Linguistics 9*(2), 167-93.

Becher, Viktor (2010) 'Abandoning the notion of 'translation-inherent' explicitation: against a dogma of translation studies,' A*cross Languages and Cultures 11*(1), 1-28.

Biber, Douglas (1988) *Variation across Speech and Writin*g, Cambridge: Cambridge University Press.

Chen, Wallace (2006) *Explicitation through the Use of Connectives in Translated Chinese: A corpus-based Study* (unpublished doctoral dissertation), University of Manchester, Manchester.

De Sutter, Gert, Isabelle Delaere and Koen Plevoets (2012) 'Lexical lectometry in corpus-based translation studies: combining profile-based correspondence analysis and logistic regression modeling' In Michael P. Oakes and Meng Ji (eds.), *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research* 325–346, Amsterdam: John Benjamins Publishing Co.

Eskola, Sari (2000) 'Untypical frequencies in translated language: a corpus-based study on a literary corpus of translated and non-translated Finnish' In Anna Mauranen and Pekka Kujamäki (eds.), *Translation Universals: Do They Exist?* 83-99, Amsterdam: John Benjamins Publishing Co.

House, Juliane (2008) 'Beyond intervention: universals in translation?' *Trans-kom 1*(1), 6-19.

Jenset, Gard B. and Barbara McGillivray (2012) 'Multivariate analyses of affix productivity in translated English' In Michael P. Oakes, and Meng Ji (eds.), *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research* 301-24, Amsterdam: John Benjamins Publishing Co.

Kenny, Dorothy (2000) 'Translators at play: exploitations of collocational norms in German-English translation' In Bill Dodd (ed.), *Working with German Corpora* 143-60, Birmingham: University of Birmingham Press.

Kenny, Dorothy (2001) *Lexis and Creativity in Translation: A Corpus-based Study,* Manchester: St. Jerome Publishing.

Laviosa-Braithwaite, Sara (1995) 'Comparable corpora: towards a corpus linguistic methodology for the empirical study of translation' In Marcel Thelen and Barbara Lewandowska-Tomaszczyck (eds.), *Proceedings of the Maastricht Session of the 2nd International Maastricht-Lodz Duo Colloquium on Translation and Meaning* 153-63, Maastricht: Hogeschool Maastricht.

Laviosa-Braithwaite, Sara (1996) *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation* (unpublished doctoral dissertation), University of Manchester, Manchester.

Malmkjær, Kirsten (1997) 'Punctuation in Hans Christian Andersen's stories and their translations into English' In Fernando Poyatos (ed.), *Nonverbal Communication and Translation: New Perspectives and Challenges in Literature, Interpretation and the Media* 151-62, Amsterdam: John Benjamins Publishing Co.

Olohan, Maeve & Mona Baker (2000) 'Reporting that in translated English: Evidence for subconscious processes of explicitation?' *Across Languages and Cultures 1*(2), 141-58.

Puurtinen, Tiina (2003) 'Nonfinite constructions in Finnish children's literature: Features of translationese contradicting translation universals?' In Sylvianc Granger. Jacques Lerot and Stephanie Petch-Tyson (eds.), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies* 141-54, Amsterdam: Rodopi.

Øver s, Linn (1998) 'In search of the third code: An investigation of norms in literary translation,' *Meta 43*(4), 557 – 70.

Teich, Elke (2001) 'Towards a model for the description of cross-linguistic divergence and commonality in translation' In E. Steiner and C. Yallop (eds.), *Exploring Translation and Multilingual Text Production: Beyond Content,* 191-27. Berlin: Mouton de Gruyter.

Teich, Elke (2003) *Cross-linguistic Variation In System And Text: A Methodology for the Investigation of Translations and Comparable Texts,* Berlin: Walter & Gruyter.

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

Xiao, Richard (2010) 'How different is translated Chinese from native Chinese?: a corpus-based study of translation universals.' *International Journal of Corpus Linguistics, 15*(1), 5-35.

[Abstract]

Multidimensional Explanatory Analysis of Translation Universals

Lee, Changsoo
(Hankuk University of Foreign Studies)

The main objective of the paper is to demonstrate the usefulness of mutlivariate explanatory data analysis in analyzing datasets with multiple variables in corpus-based translation studies, particularly those geared toward testing various translation universal hypotheses. The paper accomplishes this objective by carrying out a case study in which principal component analysis (PCA) is employed to test the validity of four linguistic features which are alleged in the literature to be representative of simplification and explication hypotheses, using a comparable corpus of English translations of Korean fiction and authentic English fiction. The case study renders empircal support to the 'conjunctive' and 'that' explication hypothesis, while finding the others irrelevant as features of translation universals. In the process of the case study, the relevant steps and procedures of using PCA are illustrated, and its merits are discussed in terms of allowing an in-depth integrated explanatory analysis of multiple variables as opposed to the traditional confirmatory statistical methods that simply focus on testing statistical significance.

▸ Key Words: corpus-based translation studies, explanatory data analysis, principal component analysis, translation universals, Korean-English literary translation.

이창수
한국외국어대학교 통역번역대학원
soolee@hanmail.net
관심분야: 문학번역연구, 코퍼스언어학, 코퍼스기반 번역연구, 담화분석, 기호학

논문투고일: 2014년 8월 4일
심사완료일: 2014년 8월 29일
게재확정일: 2014년 9월 15일