번역학연구
2015년 여름 제16권2호

최빈도 어휘를 활용한 동일 원문의 번역물 간 번역문체 연구*

이 창 수
(한국외국어대)

1. 연구 목적 및 배경

본 연구의 목적은 최빈도 어휘(most frequent words: MFW)를 분석 자질로 하여 컴퓨터문체 분석 기법을 활용하여 동일 원문에 기초한 두 개의 번역문의 문체 차이를 정량 분석 연구하는 것이다.

컴퓨터 문체 분석(stylometry)이란 코퍼스에서 다양한 언어 자질에 관한 빈도 데이터를 추출하여, 탐구적 다변량 분석(exploratory multivariate analysis)이나 서포트 벡타머신(support vector machine)과 같은 기계학습 알고리즘 분석 기법을 활용하여 일반적 관찰이나 수작업으로는 찾아내기 힘든 문서에 숨겨진 저자의 문체 흔적을 찾아내는 작업을 의미한다. 이와 같은 기법은 주로 저자가 불분명하거나 논란이 되고 있는 문서의 저자를 판별하는 데 사용되기 때문에

* 본 연구는 2015년도 교내연구비 지원을 받아 작성되었음.

저자판별(author attribution)연구라고 부르기도 한다(Juola 2006: 238; Oakes 2014: 1). 컴퓨터 문체 분석은 코퍼스에서 기능어와 같이 저자의 의식적 통제를 벗어나는 언어 자질을 분석하면 개인 간의 문체를 분리 식별할 수 있다는 전제에서 출발한다. 이와 같은 문체 자질을 인간의 지문에 비유하여 '저자 지문(authorial fingerprint)'이라고 칭하는데, 텍스트에는 이와 같은 저자 지문이 심어져 있기 때문에 이를 찾아낸다면 저자가 불분명한 문서의 저자를 판별할 수 있다는 것이다(Oakes 2014: 1).

컴퓨터 문체 분석기법이 번역 문체 연구에 접목된 것은 2000년대에 들어서다. 번역 문체와 관련하여 헤르만즈(Hermans 1996)는 일찍이 번역자의 목소리(translator's voice)를 언급한바 있다. 헤르만즈(2009: 97)에 따르면 번역자는 번역물에 언어적으로 특이한 자기만의 서명(linguistically idiosyncratic singature)을 남긴다. 베이커(2000: 244) 또한 사람이 물건을 만지면 해당 물건에 지문이 남듯이 번역자는 번역물에 문체 지문을 남긴다고 주장하였다. 이와 같은 배경에서 최근 들어 컴퓨터 문체 분석기법을 활용하여 번역물과 비번역물 간의 문체 차이를 분석하거나 번역물에서 원저자나 번역자의 문체를 찾고자 하는 연구가 활발하게 진행되고 있다(Baroni and Bernardini 2006; Covington et al. 2014; De Sutter et al. 2012; Forsyth and Lam 2014; Grabowski 2013; Hedegaard & Simonsen 2011; Ilise 2013; Ilise and Inkpen 2011; Jenset & McGillivray 2012; Ji 2010; Lynch 2014; Rybicki 2008, 2012; Rybicki & Heydel 2013; Volansky et al. 2013).

본 연구에서는 컴퓨터 문체 분석 기법을 동일 원문에서 비롯된 다른 번역물의 문체 차이를 분석하는 데 적용하고자 한다. 아직 문헌에는 이와 같은 방식으로 번역 문체를 분석한 연구가 보고되지 않고 있다. 구체적으로 본 연구에서는 최빈도 어휘가 동일 원문에서 출발한 서로 다른 번역물의 문체를 구분하는 문체 표지가 될 수 있는지, 있다면 어떤 어휘들이 특징적으로 차이가 나며, 그와 같은 어휘들이 시사하는 문체의 특징은 무엇인지를 분석하고자 한다.

2. 최빈도 어휘를 활용한 컴퓨터 문체 분석 연구

컴퓨터 문체 분석을 활용한 저자판별의 역사는 1964년 모스텔러와 왈레스(Mosteller and Wallace 1964/1984)가 미국의 Federalist Papers란 텍스트 군에서 저자가 논란이 되고 있는 텍스트의 저자를 판별한 연구로 거슬러 올라간다. Federalist Papers는 88건의 신문 에세이로 구성되어 있는데 이중 12건의 저자가 불확실하다. 동 연구에서 모스텔러와 왈레스는 접속사, 전치사, 관사 등 30종류의 기능어의 발생빈도를 분석하여 저자가 불확실한 문서는 저자가 알려진 문서의 저자들 중 특정 저자의 문서와 유사하다는 점을 밝혀냈다(Juola 2006: 242). 그러나 컴퓨터 문체 분석 연구가 본격적으로 꽃을 피게 된 것은 1990년대 들어서다. 인터넷의 발달로 전자문서 데이터에서 정보를 추출하기 위한 기계학습이나 자연어 처리와 같은 첨단 컴퓨터 기법들이 등장하고, 컴퓨터 성능의 향상과 더불어 주성분분석(PCA)과 같은 다변량 통계기법의 확산으로 수천 개의 어휘를 통합적으로 분석하는 것이 가능해진 것이 결정적 동기가 되었다(Stamatatos 2009: 539; Juola 006: 239-240). 저자판별과 관련된 주요 연구 성과는 주올라(Juola 2006), 코펠 외(Koppel et al. 2009), 오크스(Oakes 2014), 스타마타토스(Stamatatos 2009) 등을 참고하기 바란다.

저자판별 연구에서 사용되는 언어자질의 수는 천여 개에 달할 정도로 매우 다양하다(Rudman 1998). 그렇지만 가장 많이 그리고 가장 성공적으로 사용되어 온 언어자질은 최빈도 어휘이다(Oakes & Pichler 2013: 224). 문서에서 가장 빈도수가 높은 어휘는 전치사와 같은 기능어이기 때문에 앞서 언급했던 모스텔러와 왈레스의 Federalist Papers 연구도 실질적으로는 최빈도 어휘를 분석 자질로 사용한 연구라고 할 수 있다. 저자 판별에 사용되는 문체 자질은 주제 내용이나 장르 그리고 저자의 의식적 언어 통제에서 독립적이어야 하는데, 기능어는 문장 내의 문법적 관계를 표현하는 어휘이고 그 발생 빈도가 높다는 점에서 이와 같은 요건을 충족한다. in, of 같은 전치사나 can과 같은 조동사, 또는 when과 같은 접속사들은 그 자체만으로는 큰 의미가 없을지 모르지만 이 같은 최빈도 어휘들이 많이 등장하는 문서와 그렇지 않은 문서 간에는 문장이나 문법 구조에서 큰 차이가 있을 수 있으며(McKenna & Antonia 2001: 354), 그럴 경우에 이와 같은 최빈도 어휘는 문서를 구분 짓는 중요한 문체 표지(style

marker)가 된다.

최빈도 어휘를 사용한 저자판별 연구의 가장 대표적인 예는 버로우(Burrow 2002a, 2003)의 영국 왕정복고시대의 시를 대상으로 한 연구이다. 동 연구에서 버로우는 30개의 최빈도 어휘를 사용하여 버로우의 델타(Burrow's Delta)값을 구하였다. 버로우의 델타는 저자가 알려진 문서 군의 어휘의 발생빈도 z값과 테스트 문서 군의 z값 간의 절댓값 차이의 평균치이다. 이 수치가 작을수록 텍스트 문서는 해당 저자의 작품일 가능성이 높게 된다. 이 같은 델타값을 사용한 버로우의 연구에서 저자판별 정확도는 거의 100퍼센트에 도달했다. 이외에도 최빈도 어휘의 발생빈도 간 상관관계 계수를 사용하여 주성분분석을 통해 2차원 지면에 문서 위치를 표시하고 문서간 거리를 따져 문서 간의 저자 동일성을 판단하는 방식이 있다. 이 방식은 문서 후보자 군이 명확한 경우에 특히 효과적이다(Rybicki 2012: 232).

최빈도 어휘에 기초한 컴퓨터문체 분석을 번역연구에 적용한 사례는 아직 매우 제한적이다. 버로우(2002a)는 자신의 버로우 델타값을 사용하여 1946년에서 1967년 사이 라틴어 시인 주베날(Juvenal)이 쓴 시를 영어로 번역한 왕정복고시대 영국 시인의 판별을 시도하였다. 연구 결과 번역을 한 영국 시인 저자가 밝혀진 경우도 있지만, 그렇지 않은 경우도 있었는데, 이를 두고 버로우는 일부 번역자는 번역작업을 의식하여 저자의 문체 뒤에 자신의 문체를 숨겼다고 해석하였다. 그라보우스키(Grabowski 2013)는 1,000개의 최빈도 어휘를 사용하여 주성분분석과 군집분석을 통해 번역문과 비번역문의 차이를 연구하였다. 리비키와 헤이델(Ribicky & Heydel 2013)은 최빈도 어휘를 사용한 합의나무(consensus tree)분석을 통해 버지니아 울프의 「밤과 낮(Night and Day)」을 덴마크어로 공동 번역한 번역물에서 두 번역자의 문체를 분리하는 연구를 진행하였다. 또한 리비키(2012)는 앞서 사용한 것과 같은 연구방법을 통하여 번역물에서 번역사와 원저자의 문체 중 어느 것이 드러나는가를 연구하였다. 분석 결과 번역물이 원저자 중심으로 분류가 이뤄져 번역사의 문체는 드러나지 않았다.

3. 연구방법

앞서 연구의 목적에서도 밝혔듯이 본 연구의 목적은 최빈도 어휘를 분석 자질로 하고 컴퓨터 문체 분석 기법을 활용하여 동일 원문의 복수 번역물 간의 문체의 차이를 연구하는 것이다. 연구에 사용된 데이터는 1960년에 출판된 황순원 저 「나무들 비탈에 서다」라는 제목의 한국어 소설과 이를 영어로 번역한 번역본 2권이다. 「나무들 비탈에 서다」는 1960년에 '사상계'에 연재된 장편소설로 6.25 전쟁에 참전하였던 동호, 현태, 운구란 세 젊은이들의 파멸적 인생을 통해 전쟁의 비극성을 고발한 작품이다. 두 영어번역본은 1980년에 출판된 Trees on the Cliff(Chang Wang-rok 번역)와 2005년에 출판된 Trees on a Slope(Bruce Fulton과 Ju-Chan Fulton 번역)이다. 편의상 1980년 번역본은 TT80, 2005년 번역본은 TT05로 부르기로 한다.

TT80과 TT05에서 최빈도 어휘를 추출하는 작업은 R 통계프로그램의 컴퓨터 문체 분석용 패키지인 stylo 패키지(Eder et al. 2013)를 활용하였다. 이같이 추출한 어휘목록에서 상위 50개 어휘를 선택하여 합의나무(consensus tree)분석과 주성분분석(principal component analysis: PCA)을 시행하였다. 최빈도 어휘 분석에 몇 개의 단어를 사용하느냐는 특별하게 정해진 규칙은 없다. 앞서 언급한 버로우(2002a, 2003)는 30개의 단어를 사용하였고, 그라보우스키(2013)는 1,000개의 단어를 사용하였다. 일반적으로 저자판별 목적으로는 단어 수가 많을수록 효과적이라고 알려져 있지만 본 연구에서는 번역본 간의 최상위 발생빈도 어휘를 비교해 보는 것이 목적이기 때문에 50개로 한정하였다.

합의나무분석에는 stylo 패키지에서 제공하는 기능을 사용하였고 주성분분석은 R 통계프로그램의 FactoMineR 패키지를 사용하여 시행하였다. stylo 패키지에도 주성분분석을 수행할 수 있는 기능이 포함되어 있지만 FactoMineR 패키지를 사용하면 보다 다양한 분석이 가능하다.

4. 분석 결과와 논의

4.1. 1단계 분석: 최빈도 어휘 분석의 효과성

먼저 최빈도 어휘 분석이 동일 원문의 다른 번역문 간의 문체 차이를 연구하는 데 타당한 분석 자질이 되는지를 검토해보기로 한다. 이를 위해서 먼저 TT80과 TT05를 각각 4개의 샘플로 잘라서 TT80a, TT80b, TT80c, TT80d, TT05a, TT05b, TT05c, TT05d 등 8개의 분석 샘플을 만들었다. 그 이유는 본 연구에서 시도하고자 하는 통계분석은 3개 이상의 텍스트 샘플을 요하기 때문이다. 다음으로 stylo 패키지를 사용하여 8개의 샘플 텍스트에서 발생빈도 순으로 어휘목록을 추출하였다. 그 결과는 <표 1>과 같다.

〈표 1〉 샘플 텍스트의 최빈도 어휘 데이터

	the	to	a	and	of	was	it
TT05a	5.575926	2.655203	2.452902	2.339107	2.364395	1.308636	1.220129
TT05b	5.300243	2.587766	2.313400	2.594001	1.970443	1.471597	1.465361
TT05c	4.875948	2.733747	2.544222	2.101999	1.728693	1.320928	1.148633
TT05d	5.894297	2.918141	2.494634	2.239369	1.879677	1.548993	1.200905
TT80a	5.676521	2.879765	2.343401	1.966669	2.030522	1.577166	1.283443
TT80b	5.467717	2.727559	2.085039	2.280315	1.971654	1.581102	1.385827
TT80c	4.744410	2.717426	2.552733	1.615253	1.678596	1.399886	1.254196
TT80d	5.924197	2.789819	2.239635	1.983995	1.817272	1.456041	1.172613

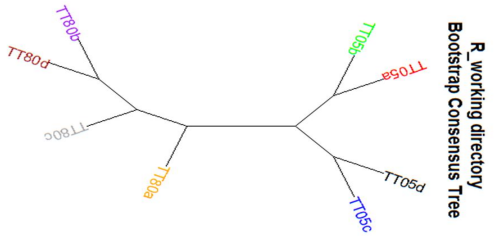
<표 1>에서 맨 왼쪽 열은 8개의 샘플 텍스트 이름을 나타내며 맨 위쪽 행은 발생빈도순으로 배열된 단어를 보여준다. 실제 데이터는 샘플 텍스트 군에서 발생하는 모든 단어를 포함하고 있기 때문에 옆으로 수천 개의 단어가 늘어서있지만 <표 1>에는 공간 제약 상 7개의 단어만 예시되어 있다. 각 단어 아래는 개별 샘플 텍스트에 대한 해당 단어의 발생빈도 수치가 기록되어 있다. 이 숫자는 실 발생건수가 아니라 z값으로 표준화된 표준값이다.

본격적인 분석에 들어가서 먼저 <표 1>의 어휘 데이터 중 상위 50개를 선택하여 합의나무 분석을 시행하였다. 합의나무란 기본적으로 군집분석(cluster analysis)과 같이 분석변수(최빈도 어휘)들을 사용하여 샘플 텍스트 간의 거리를 분석한 후 근접 거리에 있는 샘플들을 군집으로 묶어 덴드로그램(dendrogram)이라고 불리는 나뭇가지 형태의 그래프를 생성한다. 다만 일반 군집분석에서는 일회 분석 결과를 사용하지만, 합의나무 분석에서는 군집분석을 시행한 후 데

이터의 일부를 샘플로 사용하여 별도 군집분석을 시행하여 두 덴드로그램이 일치하는지 여부를 확인한다. 이와 같은 분석을 여러 차례 반복 시행하여 여러 분석 간에 일치되는 부분을 중심으로 최종 덴드로그램을 생성한다(Baayen 2008: 146-148). 따라서 일반 군집분석보다 신뢰도가 높다. 본 연구에서는 앞서 2절에서 언급한 버로우의 델타값을 거리 데이터(distance data)로 사용하여 합의나무 분석을 실시하였다.

만약에 최빈도 어휘가 TT80과 TT05를 구분하는데 유효한 언어 자질이라면 각 번역문에서 파생된 4개의 샘플 텍스트는 최빈도 어휘 분석에서 상호 간에는 유사성을 띠면서 다른 4개의 샘플 텍스트와는 분명한 차이를 보여야 한다. 다시 말하면 각 번역문의 샘플 텍스트는 거리상 서로 가까운 곳에 위치하여 하나의 군집을 형성하며 다른 번역문의 샘플과는 서로 떨어져 있어야 한다. 그림 1의 합의나무 분석 결과를 보면 바로 그와 같은 상황이 드러나고 있다. 이 도표를 보면 8개의 샘플 텍스트는 중앙을 중심으로 좌-우로 뻗은 나뭇가지의 양단에 몰려 군집을 형성하면서 확실하게 분리되어 있다. 이는 최빈도 어휘 50개만으로도 두 번역문을 확실하게 분리하여 식별할 수 있다는 의미이다.

그림 1 TT80과 TT05의 8개 샘플에 대한 합의나무분석 결과



다음으로 같은 최빈도 어휘 50개와 8개의 샘플 텍스트를 사용하여 주성분분석을 시행해보았다. 주성분분석은 다수의 분석변수 간에 상호 연관성을 분석하여 이를 두 세 개 정도의 새로운 변수로 압축하는 통계분석방식으로 이렇게 새롭게 도출한 변수를 주성분이라고 한다.[1] FactoMineR 패키지의 PCA 명령어

를 사용하여 주성분분석을 시행한 결과는 그림 2에 나와 있다. 이를 보면 x축에 해당하는 Dim 1와 y축에 해당하는 Dim 2로 이뤄진 2차원 공간에 8개의 샘플 텍스트가 서로 다른 거리를 두고 배열되어 있다. Dim 1과 Dim 2는 PCA 분석을 통해 새롭게 찾아낸 변수, 즉 제 1 주성분과 제 2 주성분에 해당한다. 이 지도를 보면 두 번역본 샘플 텍스트들은 Dim 1 축을 중심으로 좌우로 독립된 군집을 형성하여 상호 분리되어 있다. 따라서 Dim 1은 두 번역본을 구분하는 역할을 하는 주성분이라고 판단할 수 있다.

그림 2 샘플 텍스트 PCA 지도

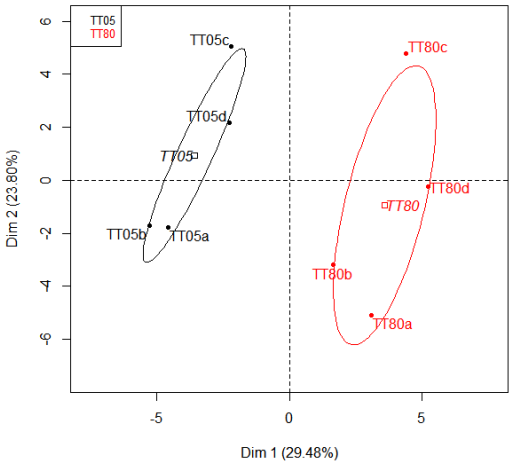


그림 2에 있는 두 개의 타원은 95% 신뢰타원이라고 부르는데 일반 통계의 95% 신뢰구간과 유사하여 그림 2에서처럼 두 타원이 겹치지 않으면 양 샘플 군 간에 통계적으로 유의미한 차이가 있다는 것을 의미한다. 이렇게 Dim 1 축을 따라 TT80과 TT05의 샘플 텍스트들이 양쪽으로 명확히 분리되어 있는 모습은 그림 (1)의 합의나무분석 결과와 일치한다. 이와 같은 분석 결과에서 최빈도 어휘는 동일 원문의 다른 번역문을 구별하는 데에도 유효한 언어지표라는

1) 주성분분석에 대한 보다 자세한 설명은 저자의 2014년 논문(이창수 2014)을 참고하기 바란다.

것을 재차 확인할 수 있다.

4.2. 2단계 분석: 두 번역문을 구분하는 핵심 어휘 분석

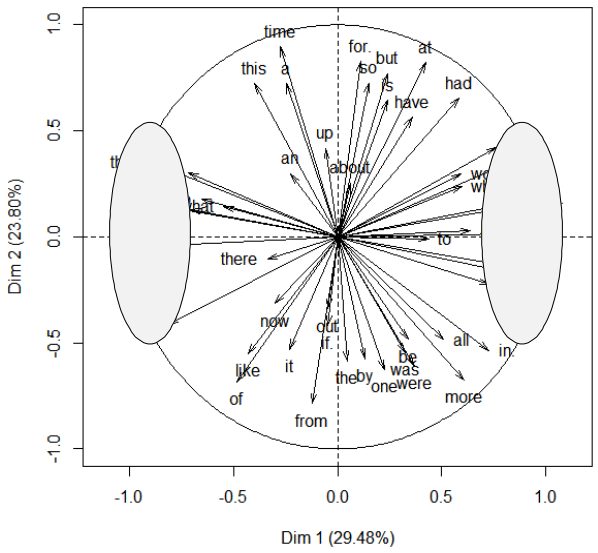
4.1 절의 합의나무와 주성분분석에선 최빈도 어휘 50개만으로도 두 번역문을 확실히 구분할 수 있다는 점을 확인하였다. 그러나 50개의 어휘 모두가 같은 변별력을 가지고 있는 것은 아니다. 그렇다면 어떤 어휘가 두 번역문을 구분하는데 핵심 역할을 하고 있나? 이 질문에 답하기 위해선 PCA분석에서 생성된 또 다른 도표인 그림 3을 살펴볼 필요가 있다.

그림 3은 그림 2와 똑같은 2차원 주성분 공간에 분석변수인 50개의 단어들을 배열해 놓은 것이다. 도표 상의 화살표는 해당 방향으로의 상관관계 강도를 나타내며 바깥 원은 상관관계가 1이 되는 지점을 나타낸다. 따라서 두 그래프를 서로 겹쳐 보면 어떤 단어들이 어떤 텍스트 샘플과 상관관계가 있는지를 알 수 있다.

앞서 4.1 절의 분석에서 두 번역문을 구분 짓는 주성분은 Dim 1으로 파악되었다. 따라서 어떤 어휘들이 Dim 1과 상관관계가 높은지를 분석하면 두 번역문을 구분하는 핵심 어휘를 찾아낼 수 있다. Dim 1과 상관관계가 높은 어휘들은 그림 3에서 Dim 1에 근접해 있으면서 동시에 양 끝으로 화살표가 가장 길게 뻗어 있는 단어들이다. 이 단어들은 Dim 1 양쪽 끝에 타원형으로 묶어 표시해 놓았다.

이들 어휘를 보면 우선 Dim 1의 오른쪽 방향에 위치한 TT80과 높은 상관관계를 보이는 단어들은 with, not, said, after, could, been, who, on, would, when, as등이다. 이와 반대로 Dim 1의 왼쪽 방향에 위치한 TT05와 상관관계가 강한 어휘들은 then, s, and, t, woman, that 등이 있다. 바로 이 단어들이 두 번역문을 구분 짓는 최빈도 어휘, 즉 문체 표지(style marker)이다. 이들 어휘 대부분은 내용 보다는 문법과 관련된 기능어라는 점이 특징적이다.

그림 3 최빈도 어휘 PCA 지도



4.3. 3단계 분석: 최빈도 어휘와 문체 특징

이번에는 4.2 절 분석에서 밝혀낸 TT80과 TT05를 구분하는 최빈도 어휘의 특징을 분석하여 이것이 양 번역문의 문체와 어떤 연관이 있는지를 논해보자. 그 전에 먼저 TT80과 TT05에 대한 일반적인 텍스트 정보를 비교해 보기로 한다. 표 (2)를 보면 TT80은 TT05보다 총 어휘 수는 더 크지만 문장 수는 적고, 따라서 평균 문장길이가 더 길다. 그리고 어휘다양도를 보여주는 STTR은 상대적으로 작다. 이는 TT80이 상대적으로 보다 길고 복잡한 문장을 사용하며 특정 어휘에 대한 의존도가 높다는 점을 시사한다.

〈표 2〉 TT80과 TT05의 텍스트 정보

	총어휘수	문장수	평균문장길이	STTR
TT80	63,728	5,085	12.53단어	43.22
TT05	62,696	5,369	11.68단어	44.57

이 같은 텍스트 정보를 염두에 두고 4.2 절에서 밝혀낸 두 번역문의 핵심 최빈도 어휘가 갖는 문체적 의미를 살펴보자. 우선 TT80과 높은 상관관계를 갖는 어휘 수는 10개로 TT05의 6개 보다 훨씬 많다. 이는 TT80이 상대적으로 특정 단어에 의존하는 경향이 더 크다는 것을 나타낸다. 이 점은 <표 2>의 STTR 수치 비교에서 TT80이 상대적으로 어휘 다양도가 떨어지는 것과 일치한다.

TT80에서 특징적으로 많이 나타나는 어휘는 with, not, said, after, could, been, who, on, would, when, as이며, TT50의 경우는 then, and, s, t, woman, that 등이었다. 이들 어휘를 놓고 여러 다양한 분석이 가능하지만 지면 제약 상 가장 두드러지는 특징 두 가지만 논하기로 한다. 먼저 TT80의 어휘 중 after, when, who, as등은 문장 내에서 구나 절을 연결하는 역할을 한다는 공통점이 있다. when은 종속접속사이며 who는 의문사로 쓰이기도 하지만 관계대명사절을 이끄는 단어이기도 하다. as와 after도 종속접속사로 사용되는 어휘이다. 이는 상대적으로 TT80이 절 간의 관계를 종속 또는 내포의 관계로 묘사하는 경향이 강하다는 점을 시사한다. 이와 대조적으로 TT50의 핵심 어휘인 then과 and는 절을 등위관계로 표현하는 데 사용되는 단어들이다.

위 어휘 중 when을 예로 삼아 두 번역문 간에 실제 발생빈도를 비교해보면 TT80은 286개, TT05는 242개이다. 빈도수 40의 차이는 그리 커 보이지 않지만 when은 절을 연결하는 접속사로 문장 단위로 발생하기 때문에 전체 문장 수에 대비해 비교해보면 확실한 차이가 난다. 두 번역문에서 when의 발생 건수를 when이 포함된 문장 수로 본다면, 이를 총 문장 수에서 빼면 when이 포함되지 않은 문장 수를 구할 수 있다. 이와 같이 계산하여 두 번역문의 문장 수 대비 when의 발생건수의 차이에 대해 카이제곱검정을 해보면 p-value = 0.009142 (X-squared = 6.795, df = 1)로 통계적으로 매우 유의한 차이로 나타난다.

예문 (1)

ST: 현태가 역시 바람벽에 바짝 등을 붙이고 문짝을 핵 잡아 젖히면서, 꼼짝 말어!했을 때, 방 안에서 사람의 기척이 났던 것이다.

TT80: When Hyontae, pressing himself to the wall, shouted “Don’t move!” and flung the door open as he had at the other houses, someone stirred in the room.

TT05: Hyont'ae flattened himself against the side of the house, jerked open the door, and shouted "Freeze!" Someone stirred inside.

이와 같은 문체의 차이는 예문 (1)에서 잘 드러난다. 여기서 TT80은 ST의 "-했을 때"란 구문을 종속접속사 When을 써서 종속절로 표현하고 있다. 이에 덧붙여 TT80은 as란 종속접속사를 추가하고 있다. 이에 반하여 TT05는 원문의 절간 종속 관계를 풀어서 and를 사용한 독립절 3개로 연결하고 있다. and가 TT05의 핵심 최빈도 어휘에 포함된 이유도 이 같은 배경에서 이해할 수 있다.

그런데 TT50의 어휘에도 that절에 사용되는 that가 포함되어 있어, TT50에도 복문의 수가 적지 않을 것이라고 추정할 수 있다. 그러나 that은 대명사로도 많이 사용되기 때문에 양 번역문에서 that가 어떤 용도로 사용되었는지를 조사해봐야 한다. 양 번역문에서 that의 실 발생건수는 TT80은 686, TT05는 738로 TT80이 약간 많은 수준이다. 이 많은 수의 that의 사용 환경을 수작업으로 조사하는 것은 비현실적이다. 따라서 <표 3>과 <표 4>의 연어관계 표에서 that과 주로 어울리는 단어들에서 힌트를 얻어 보자. 워드스미스 프로그램을 이용해 생성한 두 표에서 각 열 별로 가장 큰 차이가 있는 것은 R1 열에 있는 HE이다. 이 경우 that의 사용 환경은 바로 뒤에 대명사 he가 연결되는 관계대명사절이다. 여기서 발생 건수는 TT50은 59건인데 반하여 TT80은 101건으로 거의 두 배 가까이 많다. 또한 같은 행의 L2열을 보면 HE + (?) + THAT로 이어지는 사용 환경이 나타난다. 이 경우 (?) 자리에는 '말하다'는 SAY류의 동사가 들어간다. 따라서 이 경우도 that은 that 절을 이끄는 역할을 한다. 여기서도 발생건수가 20대 48로 TT80이 훨씬 많다. 이 두 경우에서 유추해보면 that의 실 발생건수에서는 TT05가 조금 높지만 that가 that 절을 이끄는 용도로 사용되는 상황은 TT80이 훨씬 많다고 할 수 있다. 반대로 TT05의 경우 that은 지시대명사로 더 많이 사용되고 있을 가능성이 높다. 따라서 TT80에서 접속사를 사용한 복문이 더 많다는 분석은 여전히 유효하다. 결론적으로 TT80은 TT05에 비하여 문장구조가 좀 더 복잡하고 길 가능성이 크다. 이와 같은 해석은 <표 2>에서 TT80의 평균문장길이가 상대적으로 크게 나타난 사실로 뒷받침된다.

〈표 3〉 TT80의 that 연어관계

N	Word	Total	Total Left	Total Right	L5	L4	L3	L2	L1	Centre	R1	R2
1	THAT	719	15	15	4	1	5	4	1	689	1	4
2	THE	235	114	121	22	21	30	41			39	16
3	HE	230	105	125	16	20	21	48			101	12
4	WAS	181	79	102	18	16	16	21	8		14	54
5	TO	170	94	76	9	16	20	45	4			4

〈표 4〉 TT05의 that 연어관계

N	Word	With on △	Total	Total Left	Total Right	L5	L4	L3	L2	L1	Centre	R1	R2
1	THAT	that 0	752	7	7	4	2		1		738		1
2	THE	that 0	266	134	132	25	32	32	45			30	19
3	TO	that 0	168	91	77	16	15	20	34	6		1	16
4	HE	that 0	159	69	90	14	19	16	20			59	11
5	WAS	that 0	152	63	89	7	16	16	9	15		20	33

이번에는 반대로 TT05와 관련된 최빈도 어휘에서 t, s에 주목해보자. 최빈도 어휘 목록에서 t, s란 알파벳으로 나타난 어휘는 실은 't나 's처럼 아포스트로피가 붙은 형태이다. 영어에서 't는 항상 not이 축약된 형태이다. 따라서 't가 TT05와 연관되어 있다는 말은 TT05에서는 not을 축약해서 사용하는 경향이 강한 반면 반대로 TT80에서는 not을 독립적인 단어로 사용하는 경향이 강하다는 것을 의미한다. 이는 TT80과 연관된 최빈도 어휘 속에 not이 포함되어 있다는 사실에서도 확인할 수 있다. 실제로 축약형 't의 발생빈도는 TT08은 394건인데 반하여 TT05는 478번으로 크게 차이가 난다. 반대로 축약이 되지 않은 형태인 not의 경우는 TT08은 325건인데 반하여 TT05는 209건으로 적다. 이는 통계적으로 유의미한 차이이다(p-value = 1.145e-08, X-squared = 32.5788, df = 1).

's는 Tongho's feet과 같이 일반 명사의 소유격으로 사용되기도 하고 is와 has가 줄어든 축약형에서도 나타난다. 's의 성격을 규명하기 위하여 워드스미스의 콘코던스 기능을 활용하여 's 축약형의 한 형태인 that's의 발생빈도를 조사해보았더니 TT80에서는 50번, TT05에서는 84번으로 TT05에서 훨씬 더 많이 발생하였다. 이를 총 단어 수에 대비하여 카이스퀘어 검정을 실시해보니 p-value = 0.00241 (X-squared = 9.2012, df = 1)로 통계적으로 매우 유의미한 차이로 나타났다. 따라서 't와 's의 경우를 종합해보면 TT80에 비하여 TT05가 축약형 표현을 특징적으로 많이 사용하고 있는 것은 분명하다.

그렇다면 두 번역문간에 축약형 빈도에서 차이가 나는 이유는 무엇일까? 축약형은 기본적으로 대화문에서 발생한다. 그렇지만 두 번역문이 동일 원작에 기반을 두고 있기 때문에 어느 한쪽이 특별히 더 많은 대화문을 포함하고 있을 가능성은 매우 적다. 따라서 TT80이 대화문이 아닌 경우에도 축약형을 사용하고 있을 가능성이 크다. 실제로 축약형이 사용된 예문들을 조사해 보니 원작 소설의 서술문을 번역하는 과정에서 TT80은 비축약형을, TT05는 축약형을 사용하는 차이가 발견되었다. 예문 (2)에서 ST는 소설에서 내레이션에 해당하는 서술문이다. 원문의 "-할 수 없었다"는 부분을 양 번역문 다 조동사 would와 부정사 not을 사용하여 표현하고 있는데 TT80은 두 단어를 독립적으로 사용하고 있는데 반하여 TT05는 축약해서 사용하고 있다. 즉, TT80은 대화문에서만 축약형을 사용하는데 반하여 TT05는 서술문에서까지 축약형을 사용하는 문체적 특징이 드러난 것이다.

예문 (2)
ST: 그러나 동호는 오금이 말을 듣지 않아 그리 달려갈 수 없었다.
TT80: His legs would not let him.
TT50: But Tongho's legs wouldn't obey.

이상의 분석을 종합해보면 TT80의 문체는 상대적으로 만연체와 격식체 성격이 강한데 반하여 TT05는 간결하면서도 구어체 성격이 강하다고 판단할 수 있다.

5. 결론

4 절의 분석결과에서 다음과 같은 결론을 이끌어낼 수 있다. 첫째, 컴퓨터 문체 분석에서 저자 판별 연구에 많이 사용되는 최빈도 어휘는 동일 원문의 서로 다른 번역문을 구분하는 데에도 효과적인 언어 자질이다. 이는 본 연구에서 합의나무분석이나 주성분분석의 그래프 상에서 TT80과 TT05의 텍스트 샘플들이 뚜렷한 군집을 형성하며 서로 분리되어 나타난 점에서 확인할 수 있다.

둘째, 주성분분석을 통해 번역문을 구분하는 핵심 어휘들은 찾아낸다면, 이를 통해 번역문 간의 문체의 차이를 구체적으로 밝힐 수 있다. TT80과 TT05를 특징짓는 최빈도 어휘는 대부분 절이나 구를 형성하고 연결하는 역할을 하는 기능어들이다. 따라서 이들 기능어의 차이는 곧 문장 구성의 차이로 연결되고 문체에 영향을 준다. 이와 같은 맥락에서 4.3 절에서 예시한 것처럼 이들 기능어들의 특징과 사용 환경을 분석한다면 번역문 간의 구체적인 문체의 차이를 밝혀낼 수 있다.

이와 같은 문체의 차이는 통계적 패턴에서 드러나는 것이기 때문에 시각적 관찰만으로 발견하기는 매우 어렵다. 따라서 저자판별 기법에서 사용되는 다변량통계나 본 연구에서는 다루지 않았지만 기계학습과 같은 최신 통계분석기법을 활용한다면 제한된 관찰과 주관적 통찰력에 의존하던 기존의 문체 연구를 좀 더 객관적이고 과학적인 차원으로 끌어 올릴 수 있다. 여기에 최빈도 어휘 외에 품사주석을 단 코퍼스에서 추출한 품사별 빈도나 품사연쇄(POS n-grams)와 같이 다양한 언어자질을 분석 데이터로 활용하고 4.3 절 분석에서처럼 콘코던스 검색이나 연어분석과 같은 기존의 코퍼스 분석기법을 추가로 사용할 경우 번역문체 연구에 대한 컴퓨터 문체 분석기법의 잠재력은 매우 크다. 이런 관점에서 본 연구가 코퍼스 기법을 활용한 번역 문체 연구에 관심을 갖고 있는 연구자들에게 새로운 연구 방법을 찾아가는 길잡이 역할을 할 수 있기를 기대해 본다.

참고문헌

이창수 (2014)「다차원통계분석법을 활용한 번역보편소 사례연구」,『번역학연구』15(3): 211-232.

Baroni, Marco and Silva Bernardini (2006) ‘A New Approach to the study of translationese: machine-learning the difference between original and translated text,’ *Literary and Linguistic Computing* 21(3): 259-274.

Burrows, John (2002a) ‘The Englishing of Juvenal: computational stylistics and translated texts.’ *Style* 36(4): 677-699.

Burrows, John (2002b) ‘‘Delta’: A measure of stylistic difference and a guide to likely authorship,’ *Literary and Linguistic Computing* 17(3): 267-87.

Burrows, John (2003) ‘Questions of authorship: Attribution and beyond: a lecture delivered on the occasion of the Roberto Busa Award ACH-ALLC 2001 New York,’ *Computers and the Humanities* 37(1): 5-32.

Covington, Michael A, Iris Potter and Tony Snodgrass (2014) ‘Stylometric classification of different translations of the same text into the same language,’ *Literary and Linguistic Computing*, doi:10.1093/llc/fqu008.

De Sutter, Gert. I, 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.

Forsyth, Ricahrd S. and Phoenix. W. Y. Lam (2014) ‘Found in translation: To what extent is authorial discriminability preserved by translators?’ *Literary and Linguistic Computing* 29(7): 199-217.

Grabowski, Łukaz (2013) ‘Interfacing corpus linguistics and computational stylistics: Translation universals in translational literary Polish.’ *International Journal of Corpus Linguistics* 18(2): 254–280.

Hedegaard, Steffen and Jakob G. Simonsen (2011) ‘Lost in Translation: authorship attribution using frame semantics.’ In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* 65-70. Stroudsburg, PA: Assocaition for Computational Linguistics.

Hermans, Theo (1996) ‘The Translator’s voice in translated narrative,’ *Target* 8(1): 24-46.

Hermans, Theo (2009) ‘Translation, ethics, politics.’ In Jeremy Munday (ed) *The Routledge Companion to Translation Studies* 93-105, Abingdon, Oxon: Routledge.

Ilisei, Iustina (2013) *A Machine Learning Approach to the Identification of Translational Language: An Inquiry into Translationese Learning Models*. Unpublished Ph. D Thesis, Wolverhampton: University of Wolverhampton.

Ilisei, Iustina and Diana Inkpen (2011) 'Translationese traits in Romanian newspapers: a machine learning approach,' *International Journal of Computational Linguistics and Applications* 2(1-2): 319-332.

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* 301-324, Amsterdam: John Benjamins.

Ji, Meng (2010) *Phraseology in Corpus-Based Translation Studies*, Bern, Switzerland: Peter Lang AG.

Juola, Patrick (2006) 'Authorship attribution.' *Foundations and Trends in Information Retrieval 1*(3). 233-334.

Koppel, Moshe, Jonathan Schler and Shlomo Argamon (2009) 'Computational methods in authorship attribution,' *Journal of the American Society for Information Science and Technology* 60(1): 9-26.

Lynch, Gerad (2014) 'A Supervised learning approach towards profiling the preservation of authorial style in literary translations.' In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Paper* 376–386. Available at http://anthology.aclweb.org/C/C14/

McKenna, C. W. F. and A. Antonia (2001) 'The Statistical analysis of style: reflections on form, meaning, and ideology in the 'Nausicaa' Episode of Ulysses.' *Literary and Linguistic Computing* 16(4): 353-373.

Mosteller, Frederick and David L. Wallace (1964/1984) *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*. New York: Springer-Verlag.

Oakes, Michael. P. (2014) *Literary Detective Work on the Computer,*

Amsterdam: John Benjamins B.V.

Oakes, Michael. P. and Alois Pichler (2013) 'Computational slylometry of Wittgensteins "Diktat f r Schlick",' *Bergen Language and Linguistics (BcLLs)* 3(1): 221-240.

Rudman, Joseph (2005) 'The non-traditional case for the authorship of the twelve disputed Federalist Papers: a monument built on sand.' In *Proceedings of ACH/ALLC 2005*. Available at http://www.is.informatik.uni-duisburg.de/bib/docs/ACH_ALLC_05.html.en

Rybicki, Jan (2008) 'Burrowing into Translation: character idiolects in Henryk Sienkiewicz's Trilogy and its two English translations,' *Literary and Linguistic Computing* 21(1): 91-103.

Rybicki, Jan (2012) 'The great mystery of the (almost) invisible translator: stylometry in translation.' In Michael P Oakes and Meng Ji (eds) *Quantitative Methods in Corpus-Based Translation Studies* 231-248. Amsterdam: John Benjamins Publishing Co.

Rybicki, Jan and Magda Heydel (2013) 'The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish.' *Literary and Linguistic Computing* 28(4): 708-717.

Stamatatos, Efstathios (2009). 'A survey of modern authorship attribution methods,' *Journal of the American Society for Information Science and Technology* 60(3): 538-556.

Volansky, Vered, Noam Ordan and Shuy Wintner (2013) 'On the Features of Translationese,' *Literary and linguistic computing*, doi:10.1093llcfqt031.

[Abstract]

An MFW-based Computational Analysis of Translation Style in Two English Translations of a Korean Novel

Chang-soo Lee
(Hankuk University of Foreign Studies)

The purpose of the paper is to test the usefulness of MFW(most frequent words)-based computational analysis for investigating stylistic characteristics of different translations of the same original literary work. For this purpose, the study employs consensus tree and principal component analysis to analyze 50 most frequent words in two English translations of a Korean literary classic by Hwang Sun-won. The analyses were successful in separating the two translations in consensus tree plots and PCA maps. PCA maps for features were used to identify the MFWs most distinctly associated with the two translations respectively. Then, the contextual environments of these potential style markers were analyzed to uncover important stylistic distinctions between the two translations.

▸ Key Words: computational style analysis, authorship attribution, translating style, corpus-based translation research, Korean-English literary translation.

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

논문투고일: 2015년 4월 30일
심사완료일: 2015년 5월 31일
게재확정일: 2015년 6월 8일