특집 4

빅 데이터 시대의 신어

김일환
성신여자대학교 국어국문학과 교수

1. 도입

우리는 데이터 시대에 살고 있다. 아니, 우리는 늘 데이터와 함께 살아왔다. 데이터는 과거에도 있었고 현재에도 존재한다. 디지털 혁명을 겪으면서 데이터의 규모와 생산 속도가 급격히 증가함에 따라 우리는 이전에는 경험할 수 없었던 대규모의 데이터를 확보하게 되었을 뿐이다. 소위 '빅 데이터'(big data)가 출현하게 된 것이다.

빅 데이터는 우리가 이전에 경험하지 못했던 많은 정보를 새로운 시각에서 제공해 준다는 점에서 가히 혁명적이라 할 만하다. 특히 빅 데이터에 기반한 연구는 데이터의 규모가 크건 작건 대상을 새로운 관점에서 볼 수 있는 시야를 제공해 줄 뿐 아니라 우리의 직관이 가진 한계를 극복하고, 사용자의 '솔직한' 마음을 구체적으로 포착해 낼 수 있다는 점에서 그 중요성은 나날이 증대되고 있다.

이 연구는 이러한 빅 데이터의 시대 속에서 '신어'를 논의의 주제로 하지만 구체적으로 빅 데이터 시대의 신어 자체에 대해서는 다루지 않는다. 빅 데이터 시대라고 해서 신어가 폭발적으로 증가한다든가 기존과는 확연히 구별되는 신어의 조어 유형이 등장하는 것은 아니기 때문이다. 즉 이 글에서는 빅 데이터 시대의 신어를 포착하기 위한 여러 방법에 초점을 두고 논의를 진행하고자 한다.

2. 신어 연구의 경향

지금까지의 신어 연구는 크게 다음의 몇 가지 유형으로 분류된다.

먼저 특정한 시기의 신어를 제시하고, 이들의 특성을 설명하는 연구들이 있다. 문금현(1999)를 비롯한 많은 연구가 이러한 유형에 포함된다. 이러한 연구에서는 발견된 신어들을 제시하고, 이들의 조어적인 특성, 사용 배경 등에 관심이 많았다. 이를테면 노명희(2006)에서와 같이 합성과 파생, 혼성 등과 같은 조어 유형으로 신어들을 분류하고 이들이 가지는 특성을 밝히는 데 논의를 집중한다. 신어의 조어 유형상 특징으로는 주로 혼성과 축약이 언급되기도 하였다.

두 번째로는 신어의 추출 방법과 관련한 연구이다. 신어를 직접 포착하고 그 결과를 중심으로 논의를 진행한 연구에서는 추출 방법에 대한 언급이 빠질 수 없다. 이수진·김예니(2014)에서는 ≪우리말샘≫의 '신어 추출기'를 활용하여 신어 후보를 추출하고 그 결과를 걸러 내어 최종적인 신어 목록을 확보하는 내용이 소개되기도 하였다. 김일환(2014)에서도 2000년대 신문 기사로 구성된 '물결21' 말뭉치(코퍼스)로부터 기존 사전 표제어와의 비교를 통해 신어 후보를 산출하는 과정에 대해 논의한 바 있다. 이들은 '신어 추출기'를 직접 활용하였느냐의 여부만 다를 뿐 언론 매체 중심의 텍스트 자료를 대상으로 하고, 기존 사전의 표제어와 비교, 검토하는 과정을 거쳤다는 점에서 방법론적으로 유사한 것으로 볼 수 있다.

세 번째 유형으로는 신어의 정착 과정에 대한 연구를 들 수 있다. 사실 신어의 정착 과정을 살피는 것은 이전에는 현실적으로 매우 어려운 일이었다. 생성된 신어의 쓰임을 지속적으로 관찰하는 것이 거의 불가능했기 때문이다. 그러나 최근 구축된 빅 데이터, 특히 시기별 정보가 포함된 데이터를 통해 특정 신어의 사용 여부를 추적하는 것이 가능해졌고 이를 통해 생성된 신어의 정착 여부를 논의할 수 있게 된 것이다. 이러한 연구에는 정한데로(2017),

김일환(2014) 등이 포함된다.

이러한 논의를 종합하면 신어에 대한 최근 연구는 신어(주로 국립국어원에서 조사, 발표한 신어)의 조어상의 특성을 밝히려 하거나, 신어를 새롭게 추출, 발굴하거나 아니면 사용 양상을 추적하여 정착 여부를 규명하려는 것으로 정리된다.

이러한 성과들을 정리하는 과정에서 생겨나는 의문은 과연 신어만이 가진 조어상의 독특한 특징이 있느냐 하는 점이다. 신어도 국어 단어의 하나일 뿐이라는 점을 고려하면 신어만이 가진 조어상의 독특한 특징은 그리 두드러지지 않을 수 있을 것이다. 여기에 단순히 조사된 신어만을 나열하는 것도 사전 편찬이나 신어 조사와 같은 특수 목적이 아니라면 별다른 흥미를 끌기도 어렵다. 독자 입장에서는 생소한 단어들을 접하는 것도 부담스러운 데다가 신어 후보들이 무더기로 제시되기만 한다면 이를 참을성을 갖고 확인하는 것은 여간한 인내력으로는 어림도 없는 일이기 때문이다.

3. 신어의 탐지

기존의 신어 추출은 '형태'를 중심으로 이루어져 왔다. 즉 기존에 출현하지 않은 어형을 포착하는 데 초점이 있었다. 이는 신어 조사 방법과도 관련이 있다. 형태는 변하지 않고 새로운 의미가 추가되거나 기존 의미에 변화가 생긴 단어까지 신어에 포함한다면 이들을 포착하는 것이 단순하지 않기 때문이다.

지금까지의 형태를 기준으로 신어를 추출하는 방법은 다음과 같이 비교적 단순한 과정을 거친다.

먼저 신어를 추출할 대상 텍스트를 수집하고, 이 텍스트로부터 어절의 유형과 빈도 통계를 산출한다. 이 어절 유형을 기존의 사전 표제어와 비교하여 사전 표제어에 등재되지 않은 어형을 신어 후보로 추출한다.

이러한 방식으로 신어를 추출한 김일환(2014)를 통해 좀 더 구체적으로 논의해 보자.

[1] '물결21' 말뭉치로부터 연도별 명사 사용 목록 추출(673,117개 명사)
[2] 기존 사전의 명사 목록과 비교 검토
[3] 미등재 일반 명사 확보(563,789개 명사 추출)
[4] 분석 오류, 고유 명사, 축약어, 일부 파생 명사 배제
[5] 미등재어와 신어 판별(6,698개 선정)
[6] 신어 후보 선정(1,240개 신어 명사)

김일환(2014)에서는 2000년대 신문 기사(2000~2012)로 구성된 '물결21' 말뭉치로부터 신어 후보를 추출, 선정하는 과정을 자세히 밝히고 있다. 67만여 개의 일반 명사 중 최종적으로 신어 후보로 선정된 단어는 1,200개를 조금 넘는 수준이다. 중간 과정인 [3]과 [4]의 단계를 거치면서 신어 후보는 급격히 그 규모가 축소된다(미등재 일반 명사의 0.22%).

이러한 과정은 이수진·김예니(2014)에서도 유사하게 적용된 바 있다. 단지 이수진·김예니(2014)에서는 ≪우리말샘≫의 신어 추출기를 사용했다는 점만 다를 뿐이다. 즉 이수진·김예니(2014)에서도 64,360개의 신어 후보를 확보한 후 신어 여부에 대한 판별 과정을 거쳐 최종적으로는 전체 신어 후보의 약 0.35%인 227개만을 신어로 판정한 바 있다.

이러한 신어 추출 방법은 대규모의 언론 매체 텍스트를 신어 추출의 대상으로 삼았다는 점, 이후 사전 표제어와의 비교를 통해 미등재어를 추출하고, 이를 걸러 내서 신어 후보로 삼았다는 점에서 공통점이 있다. 또한 이들은 자동 분석 과정을 일부 도입하였으며 추출된 결과를 일일이 검토하는 과정도 포함하였다. 이를 통해 해당 시기의 신어 자료를 확보할 수 있었다는 성과가 있었으나 일차로 추출된 미등재어를 일일이 검토해야 한다는 점, 그리고 전

체 미등재어 가운데 최종적인 신어 후보로 0.2~0.3%의 신어만을 확보할 수 있었다는 점은 한계로 지적될 수 있다. 즉 최종 신어 후보를 산출하는 과정이 너무 비효율적이라는 것이다. 특히 미등재어 가운데에는 조사, 어미 결합형이 상당수 포함되어 있다는 점은 미등재어의 정밀한 형태 분석을 통해 해결될 수 있을 것이다.

신어 추출을 좀 더 효율적으로 수행할 수는 없을까?

먼저 우리는 위에서 제시한 기존의 신어 추출 방법을 검토하는 과정에서 많은 수의 신어 후보가 중간 단계에서 걸러진다는 것을 확인하였다. 이들을 모두 신어에서 배제하는 것이 정당한 것일까?

신어 관련 연구들에서 살펴볼 수 있는 것은 신어에 대한 다소 엄격한 기준이다. 이들은 신어와 미등재어, 유행어, 임시어를 엄격히 구분하는 경향을 보인다.

그러나 신어는 엄밀히 말하면 새롭게 조어된 단어를 지칭하는 것이므로 유행어, 임시어와는 같은 차원에서 논의될 수 없다. 유행어, 임시어도 새롭게 조어된 것이라면 신어에 포함하지 못할 이유가 없다. 이때 기존의 신어 관련 연구에서는 해당 언어에 정착할 만한, 임시적이거나 유행적인 단어를 배제하고 신어를 선정한다. 즉 새롭게 조어된 단어라고 하더라도 모어 화자의 언어 직관 혹은 사전 편집자의 판단에 따라 신어가 되기도 하고 그렇지 않기도 한다. 이는 유행어나 임시어의 입장에서는 다소 억울한 판정일 수 있다. 신어에는 시기적인 요인도 포함된다. 작년에 새롭게 만들어진 단어는 올해의 신어가 되지 못한다. 즉 신어의 판정에는 시간성이 개입한다.

결과적으로 신어의 판정을 위한 복잡한 기준을 모두 도입하는 것보다는 우선 '신어' 후보를 정확히 판정하는 것이 더욱 중요한 것으로 보인다. 즉 미등재어를 비롯하여 유행어, 임시어 등도 모두 신어 후보로 목록화하고 이들의 사용 양상을 추적, 관찰하는 과정을 추가하는 것이 효율적일 것이다.

이와 관련하여 필자는 신어의 추출이라는 개념보다는 '탐지'의 개념을 도

입할 것을 제안한다.

'추출(extraction)'은 신어 후보가 되는 단어를 연구자가 일차적으로 탐색하고, 그 결과를 검토하여 최종적인 신어 후보를 확보하는 방식의 개념이라면 '탐지(detection)'는 기계적, 통계적 방법을 적용하여 신어 후보를 자동적으로 탐지하여 연구자에게 제공하는 방식의 개념이다. 즉 '추출'은 비용이 많이 드는 전통적인 방식인 반면 '탐지'는 빅 데이터를 활용하여 효율적으로 신어를 포착할 수 있는 방법이다.

신어 탐지를 위해서는 몇 가지 선결될 것이 있다.

우선 빅 데이터가 필요하다. 신어의 추출에서도 빅 데이터가 필요하지만 '탐지'를 위해서는 빅 데이터의 존재가 필수적이다. 특히 신어 자료 추출을 위한 텍스트 빅 데이터는 품사 정보가 부착되어야 한다. 신어 후보들이 사전에 등재되지 않은 단어들이므로 이들의 품사를 정확히 분석하여 정보를 부착하기 위해서는 미등재어까지 처리할 수 있는 높은 정확성을 가진 품사 부착 도구가 확보되어야 할 것이다.

또한 신어 탐지를 위해서는 텍스트의 생산 시점을 확인할 수 있어야 하므로 단순히 규모만 큰 빅 데이터가 아니라 시기별로 균형 잡힌 롱 데이터(long data)가 필요하다. 대규모의 시기별 데이터에 품사 정보까지 부착된다면 신어 탐지를 위한 기반이 확보될 수 있을 것이다(빅 데이터에 대한 자세한 논의는 4장을 참조).

한편 신어 탐지의 개념을 확장하면 기존의 '형태' 차원의 신어뿐 아니라 의미에 변화가 생기거나 새로운 의미가 추가된, 의미 차원의 신어도 포착할 수 있는 길이 열린다. 즉 기존에 없는 새로운 어형이 나타난 것이 아니라 형태에는 변화가 없지만 의미적으로 신어에 해당하는 단어들을 탐지해 낼 수 있다.

이는 대상어의 문맥 변화를 추적함으로써 가능하다.

다음 (1)~(2)는 '텐트'와 '천막'의 예문들로서 이들은 모두 2010년 신문기사에서 추출된 것이다.

(1) 가. 노조는 이날 오후 부산 영도 공장 안마당에서 조합원 집회를 연 뒤 본관 앞에 **천막**을 치고 간부들이 농성에 들어갔다.(한겨레신문 2010. 1. 6.)
나. **천막**에 참가자 다섯 명이 촛불 여섯 개를 쓸쓸히 들고 있더라고요.(동아일보 2010. 5. 25.)

(2) 가. 뚝딱뚝딱 **텐트**를 치고, 야생에서의 저녁 식사 준비를 거뜬히 해내고, 아이들과 손잡고 자연 속을 거니는 아빠를 보며 아이들은 '우리 아빠 최고'를 외친다.(중앙일보 2010. 4. 30.)
나. 코오롱스포츠의 '메가팰리스'는 실속형 오토캠핑용 **텐트**로 5인용 이너(inner) **텐트**의 탈부착이 가능해 내부 공간 활용성이 뛰어나다.(동아일보 2010. 5. 20.)

예문에서 확인할 수 있는 바와 같이 '천막'과 '텐트(tent)'는 최근 들어 그 쓰임이 구별되는 경향이 강하다. 즉 취미, 여가를 위한 야외 활동으로는 '텐트'를, 시위나 농성, 구호를 위해서는 '천막'이 사용된다는 것이다. 이러한 변화는 두 단어의 주요 공기어를 네트워크로 시각화한 〈그림 1〉에서 더욱 극적으로 나타난다(김일환, 2019).

〈그림 1〉만을 보면 '천막'과 '텐트'는 '설치'라는 명사만 주요 공기어로 공유할 뿐 공기하는 명사가 서로 배타적인 양상을 보인다.

장기간에 걸쳐 텍스트 빅 데이터를 확보할 수 있다면 각 시기별 단어의 공기어 벡터를 계산하고 이들이 시기별로 어떻게 변화하는지를 추적함으로써 의미적인 신어의 탐지가 전방위적으로 수행될 수도 있다.

〈그림 2〉는 1960년대부터 2010년대까지 10년 단위로 일반 명사의 공기어 벡터를 구성하고, 이들의 변화 양상을 살펴본 사례 중 하나를 보인 것으로써, '서클'이 시기적으로 공기어의 변화가 매우 큰 단어라는 점을 공기어 벡터의 유사도 계산 결과가 제시해 준다(〈그림 2〉는 그 계산 결과를 시각화해서 표현한 것).

그림 1 ‘천막’과 ‘텐트’의 공기어 네트워크(2010년 기사)

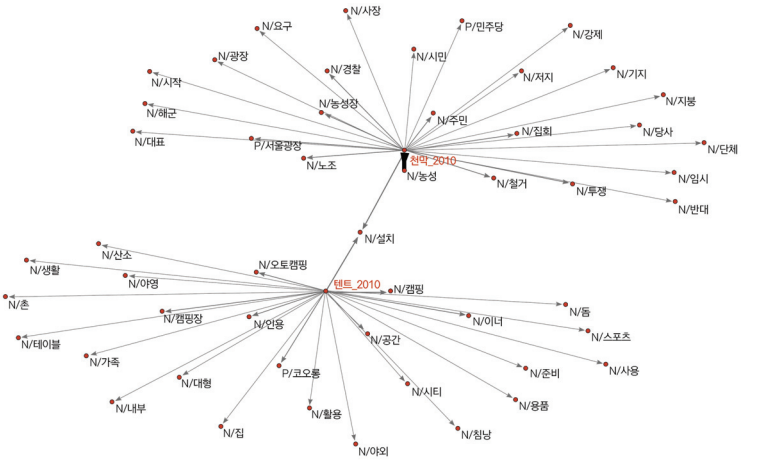
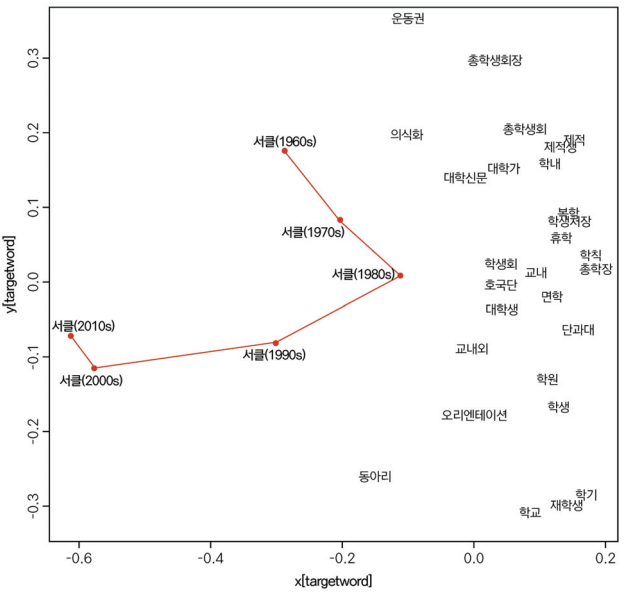


그림 2 ‘서클’의 주요 공기어 변화



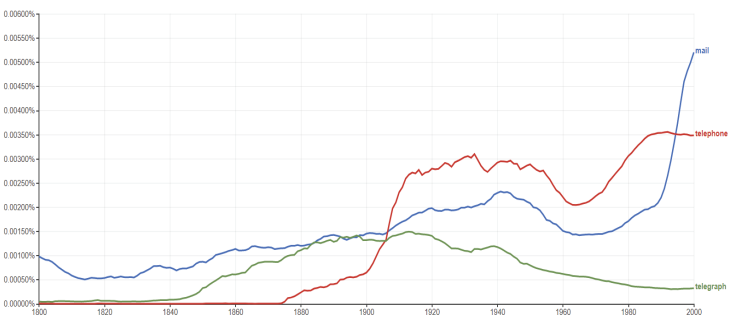
1960년대와 1970년대에는 '서클'이 주로 운동권과 관련한 단어들이 주요 공기어를 구성하다가 1990년대에 이르러서는 동아리와 유사한 단어로 사용되고 있는데 이는 '서클'이란 단어의 쓰임에 많은 변화가 있었음을 보여준다.

4. 신어와 빅 데이터

신어의 추출이나 탐지를 위해서는 국어의 언어 사용 양상을 포괄적으로 담고 있는 대규모의 텍스트 빅 데이터가 필수적이다. 특히 '빅 데이터' 시대에서 규모의 문제는 무시할 수 없는 중요한 이슈가 된다. 구글(Google)에서는 1800년대 이후 출간된 책을 모두 디지털화하여 대규모의 텍스트 빅 데이터를 구축하고 이를 다양하게 활용하고 있다. 구글이 서비스하는 엔그램 뷰어(Ngram Viewer)는 주로 언어 연쇄의 사용 추이를 보여 주지만 이는 비단 언어 문제에만 활용되는 것은 아니다.

다음의 〈그림 3〉은 'mail(메일)', 'telegraph(텔레그래프)', 'telephone(텔레폰)'의 시기별 사용 빈도를 엔그램 뷰어에서 검색한 결과이다.

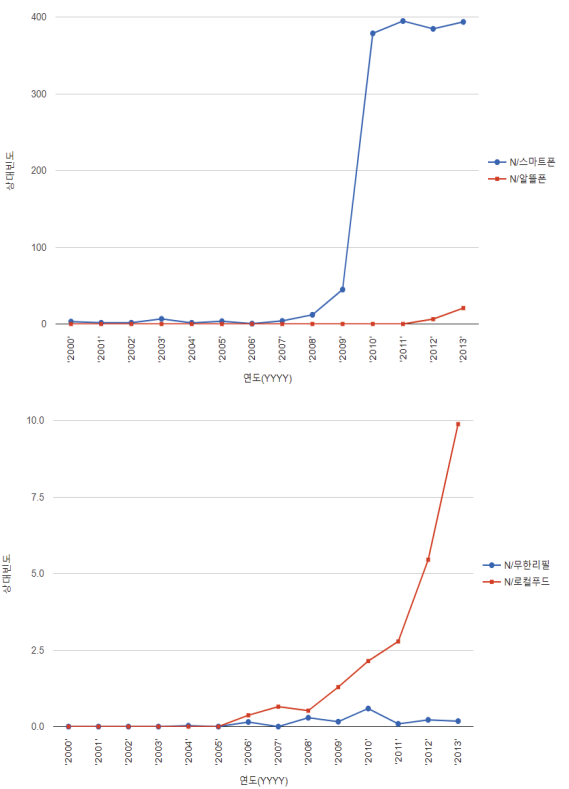
그림 3 구글(Google)의 엔그램 뷰어(Ngram Viewer)로 확인한 단어 사용의 양상



〈그림 3〉에서는 1800년부터 2000년까지 약 200년 동안의 세 단어의 사용 추이를 보여 주고 있는데, 이에 의하면 'telephone(텔레폰)'은 1870년대 후반에 신어로 포착되었을 법하다. 200년 동안 인류가 생산해 낸 단행본 서적을 디지털로 구축하여 서비스하겠다는 구글의 엄청난 포부는 다양한 분야에서 폭넓게 활용되고 있다.

국내에서는 '물결21' 말뭉치가 시기별 언어 사용 추이를 살펴보기 위한 자료로 활용된다.

그림 4 **'스마트폰', '알뜰폰'과 '무한리필', '로컬푸드'의 사용 빈도 추이**



〈그림 4〉는 '물결 21' 코퍼스를 활용하기 위해 개발된 '웹 기반 코퍼스 분석 도구'에서 '스마트폰', '알뜰폰'과 '무한리필', '로컬푸드'의 사용 빈도 추이를 보여 준다. 이들은 모두 특정 시점에서 출현이 포착되고 있다는 점에서 2000년대의 신어에 포함될 수 있다. 물론 이들의 사용 양상은 차이가 있다. '스마트폰'과 '로컬푸드'는 빈도가 크게 증가하고 있는 추이를 보이는 반면, '알뜰폰'과 '무한리필'은 그렇지 않다는 점이다.

'물결 21' 말뭉치는 비록 2000년대 주요 일간지의 신문 기사로 구성되어 있다는 한계는 있지만 시기별로 단어의 사용 추이를 관찰할 수 있는 국내의 거의 유일한 자료이다. 대상 자료의 유형을 신문 기사를 넘어서 다양하게 확장하고, 시기별로도 범위가 확대된다면 신어 연구뿐 아니라 다양한 분야에서 활용될 수 있을 것이다. 적어도 신문에 국한해서 본다면 100년(2020년은 조선일보, 동아일보의 창간 100주년이 되는 해) 동안의 신문 기사 빅 데이터를 구성하는 것도 가능하며 이를 통해 100년 동안의 신문의 언어 사용 추이를 보다 쉽고 정밀하게 관찰할 수 있을 것이다.

그렇다고 해서 규모가 모든 것을 해결해 주지는 않는다. 즉 '다다익선'이 능사는 아니다. 검색 엔진에서 다소 후발 주자였던 구글이 경쟁사들을 물리치고 독보적인 위치에 오른 배경에는 단순한 데이터의 규모를 넘어 질적으로 우수한 데이터의 확보가 있었다는 점을 우리는 기억해야 할 것이다. 즉 신어 조사와 관련한 텍스트 데이터는 규모에 대한 고려뿐 아니라 다양한 종류의 텍스트 데이터를 적절히 포함할 수 있도록 설계, 구축되어야 할 것이다.

5. 맺음말

이상으로 신어에 대한 최근 연구 경향을 간략히 살펴보고 텍스트 빅 데이터를 대상으로 신어의 추출과 탐지를 위한 몇 가지 방법과 사례를 제시하였

다. 또한 신어 연구를 위한 빅 데이터와 롱 데이터의 필요성에 대해서도 간략하게 논의해 보았다.

신어를 연구할 때에 단순히 신어 후보를 추출하고 이를 일일이 비교하는 방법은 효율성이 크게 떨어질 뿐 아니라 데이터의 규모와 범위를 확장하게 되면 더 이상 유지되기 어렵다는 점을 지적하였으며 이에 대한 대안으로 신어의 추출이 아닌, 신어 탐지의 개념이 필요함을 제안하였다. 특히 신어의 탐지를 위해서는 시기별 균형이 고려된 대규모의 텍스트 빅 데이터가 필요하며 여기에는 정밀하게 품사 정보가 부착될 필요가 있음을 논의하였다.

4차 산업 혁명은 데이터가 필수적인 시대이다. 특히 이 데이터는 적절한 방식으로 디지털화되어 있어야 하며, 많은 사용자가 쉽게 접근할 수 있도록 개방되어야 한다. 빅 데이터 시대의 신어라고 해서 기존의 신어와 크게 다를 것이 없다. 단지 달라진다면 신어를 추출 또는 탐지하고 이를 확인하는 과정이 기존의 방법과 달라질 것이다. 이러한 과정에는 최근 떠오르는 기계 학습 방법의 한 유형인 딥 러닝(deep learning) 기법이 적용될 수도 있을 것이다. 시기별 텍스트 빅 데이터가 충분히 확보되고 이를 모두가 공유할 수 있게 된다면 사전 편찬자나 연구자의 주요 업무는 신어의 조사와 추출이 아니라 탐지된 신어 후보를 잘 정리하고, 기술하는 것이 될 것이다. 구글 수준의 텍스트 빅 데이터는 아닐지라도 균형 잡히고 시기별로 잘 정리된 텍스트 빅 데이터가 확보될 날이 머지않기를 바랄 뿐이다.

참고문헌

김일환(2014), “신어의 생성과 정착 – 신문의 신어 명사를 중심으로”, ≪한국사전학≫ 24, 98~125쪽.
김일환(2019), “인문학을 위한 신문 빅 데이터와 텍스트 마이닝”, ≪어문논집≫ 78, 중앙어문학회.
노명희(2006), “최근 신어의 조어적 특징”, ≪새국어생활≫ 16-4, 31~45쪽.
도원영(2000), “국어 사전의 신어 처리”, ≪한국어학≫ 34, 21~45쪽.
문금현(1999), “현대국어 신어(新語)의 유형 분류 및 생성 원리”, ≪국어학≫ 33, 295~326쪽.
세스 스티븐스 다비도위츠 저, 이영래 옮김(2018), ≪모두 거짓말을 한다≫, 더퀘스트.
소강춘·이래호·주경미(2012), “『개방형 한국어 지식 대사전』 신어 분과의 표제어 선정과 그 실제”, ≪한국사전학≫ 20, 52~85쪽.
이도길·최재웅(2014), “물결21 코퍼스: 공개 웹 자원 및 활용 도구”, ≪민족문화연구≫, 64, 3~23쪽.
이수진·김예니(2014), “2013년 신어의 추출 방법론과 형태의미적 특성”, ≪한국사전학≫ 23, 232~262쪽.
임홍택(2018), ≪90년생이 온다≫, 웨일북스.
정한데로(2017), “신어의 삶에 관한 탐색”, ≪국어학≫ 83, 119~152쪽.

구글 엔그램 뷰어(Ngram Viewer)(https://books.google.com/ngrams/)
물결21 말뭉치(웹 기반 코퍼스 분석 도구)(corpus.korea.ac.kr)