새국어생활

국립국어원 2016-02-02
정간위 심의필 95-13-4-21

ISSN 1225-7168

새국어생활

제26권 제2호(2016년 여름) Vol. 201

인쇄일 · 발행일 2016년 6월 30일

펴낸이 송철의

편집위원 남길임 · 이광표 · 주세형 · 진정란 · 최경봉

기획 · 편집 이승재 · 김지숙 · 신다원

제작 ㈜늘품플러스

펴낸 곳 국립국어원(www.korean.go.kr)

주소 07511 서울특별시 강서구 금낭화로 154(방화3동 827번지)
National Institute of Korean Language
154, Geumnanghwa-ro, Gangseo-gu, Seoul, Korea

전화 (02) 2669-9775

전송 (02) 2669-9737

※ 정기 구독 신청 및 구독 소감, 건의 사항 등 문의
《새국어생활》 담당자 | (02) 2669-9633 | urimal365@korea.kr

새국어생활

2016년 제26권 제2호·여름

국립국어원

새국어생활

2016년 제26권 제2호 · 여름

차례

[특집] 빅데이터 시대의 국어 자료

특집

빅데이터 시대의 국어 자료

언어 자료의 보고, 빅데이터

이기황
다음소프트

1. 들어가는 말

이른바 빅데이터의 시대가 도래하였다. 거대 정보 기술(IT) 기업 가운데 하나인 아이비엠(IBM)이 최근 조사한 바에 따르면, 매일 2.5엑사바이트(1엑사바이트=1,000,000테라바이트)의 데이터가 생성되고 있다. 더욱 놀라운 사실은 전 세계에서 폭발적으로 생성되는 데이터의 90%가 최근 2년 이내에 생성되었다는 것이다(IBM, 2015). 이렇듯 방대한 양의 데이터는 미세먼지와 오존의 농도를 측정하는 센서, 카카오톡과 같은 메신저 서비스, 주식 거래소 등 다양한 원천에서 쉴 새 없이 생성되고 있다.

최근 빅데이터가 특별한 주목을 받는 이유는 그 규모 때문만은 아니다. 고도로 산업화된 오늘날 우리가 삶 속에서 겪는 여러 가지 문제를 해결하는 데 성공적으로 사용되고 있기 때문이다. 정부 주도의 빅데이터 활용 촉진 기관인 'K-ICT 빅데이터 센터(https://kbig.kr)'의 ≪빅데이터 글로벌 사례집≫(한국정보화진흥원, 2015, 2016)에서는 고객 관리, e-비즈니스, 의료, 제조, 재난·공공 등의 빅데이터 활용 분야를 소개하고 있는데 이는 수많은 빅데이터의 성공적인 적용 사례 중 극히 일부에 불과하다. 또한 최근 많은 화제를 몰고 온 인공 지능 바둑 에이전트 '알파고(AlphaGo)'는 대규모 데이

터의 유용성을 극명히 드러내었다.

주목할 것은 빅데이터의 80% 이상이 텍스트, 음성, 영상 등 구성 요소의 구조적 속성을 명시적으로 규정하기 어려운 반정형, 혹은 비정형 데이터로 구성되어 있으리라고 추정된다는 점이다(Economist, 2015). 여기서 텍스트라 함은 컴퓨터로 처리될 수 있는 형태로 저장된 글, 곧 언어 자료를 뜻한다. 실제로 앞서 소개한 빅데이터의 성공적인 적용 사례 가운데 상당수는 텍스트 자료의 분석을 통해 이루어진 것이다.

이와 같은 상황에서 우리는 빅데이터, 특히 텍스트로 이루어진 빅데이터를 언어의 탐구에 활용할 수 있는 가능성에 대하여 고려하게 된다. 언어 연구에 있어서 대규모 언어 자료인 말뭉치를 이용하는 것은 더 이상 낯선 일이 아니다. 그러므로 빅데이터를 언어 연구에 활용할 수 있는 방안에 대하여 고민하는 것은 매우 당연한 일이다.[1)]

이 글에서는 빅데이터의 개념과 특성을 언어 연구와 연관 지어 살펴보고 빅데이터를 언어 연구에 활용하기 위한 절차를 기술적 요건과 함께 소개하고자 한다. 그러나 자세한 기술적인 사항을 깊이 소개하는 것은 이 글의 범위를 벗어나는 일로 판단되어 개략적인 설명에 그쳤다.[2)] 또한 빅데이터를 언어 연구에 활용하는 일은 아직 걸음마 단계에 있으므로 명확한 방향을 제시하기 어려운 부분도 존재한다.

---

1) 언어 자료가 언어 연구에 유효한가에 대해서는 논쟁이 계속되고 있다. 촘스키는 최근 진행된 면담에서 제기된 빅데이터의 유효성에 관한 질문에 답변하면서 잘 설계된 실험을 통해 축적된 데이터의 사용에 대해서는 긍정적으로 평가하였으나 빅데이터의 유효성은 여전히 매우 부정적으로 평가하였다(뉴스센터, 2016).

2) 빅데이터를 언어 연구에 활용함에 있어서 적절한 기술의 도입과 활용은 필수적이다. 최근 말뭉치 언어학, 전산 언어학 등의 연구가 비교적 활발히 이루어지며 기술의 도입과 활용이 예전에 비해 활발해진 것은 사실이지만 빅데이터를 사용하기 위해서는 한 번의 도약이 더 필요하다.

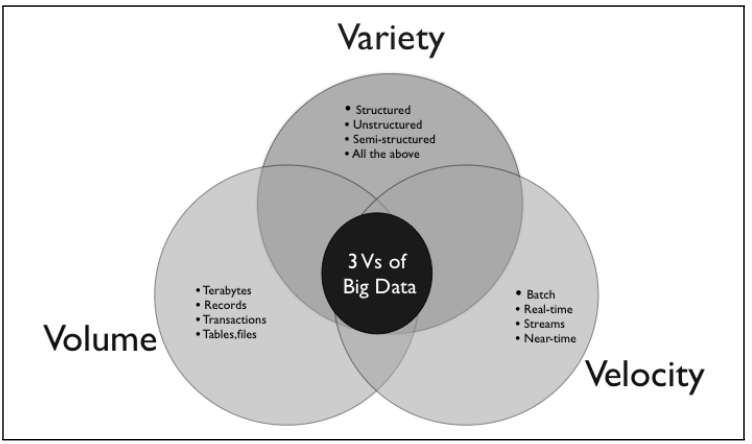
2. 빅데이터란 무엇인가?

2.1. 빅데이터의 정의

'빅데이터'라는 말은 이제 결코 생소한 용어가 아님에 틀림이 없지만, 어떠한 데이터가 빅데이터인지에 대해서는 명확히 규정하기가 쉽지 않은 것이 현실이다.[3] 그럼에도 불구하고 다음에 보이는 가트너(Gartner)의 정의는 가장 포괄적이면서도 고전적인 빅데이터의 정의로 널리 인용된다.

> 빅데이터의 정의(Gartner)
>
> Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

그림 1 빅데이터의 기본 속성: 3V



(https://commons.wikimedia.org/wiki/File:BigDataVs.png).

3) 빅데이터의 소개 자료는 이루 나열하기 힘들 정도로 많다. 다소 비즈니스 편향적이기는 하지만 IDG(2012)는 빅데이터에 대한 간략한 이해에 많은 도움이 된다. 한국소프트웨어기술인협회(2016)는 교과서로 활용될 수 있도록 단권으로 구성된 개론서이다.

위 정의의 핵심은 이른바 3V로 알려진 빅데이터의 기본 속성에 있다. [그림 1]은 빅데이터의 기본 속성 3V를 요약적으로 보여 준다. 이들 속성에 대하여 차례로 살펴보자.

첫 번째 V는 '규모' 혹은 '용량'을 뜻하는 볼륨(Volume)이다. 빅데이터의 '빅'은 규모가 크다는 빅데이터의 속성을 글자 그대로 드러낸다. 데이터 규모가 데이터의 가치를 결정짓는 유일한 속성은 아니지만 어느 정도의 질이 보장된다면 규모가 큰 데이터에서 추출된 정보의 신뢰성이 상대적으로 높다는 것은 일반적으로 알려져 있다. 그러므로 데이터에 기반을 둔 연구에서는 가능한 한 규모가 큰 데이터를 확보하기 위한 노력을 기울인다. 다만 데이터 수집, 저장, 처리 등의 공정에서 맞닥뜨리게 되는 현실적 한계는 무시할 수 없는 문제이다. 그런데 최근 컴퓨터 하드웨어와 소프트웨어의 급격한 발달로 이 한계가 급속히 무너지고 있다. 예를 들어, 2003년 인간 게놈 프로젝트를 통해 30억 개의 염기쌍 해독을 하는 데에 13년간 총 30억 달러의 비용이 들었는데 현재는 약 3.2기가바이트 용량의 인간 게놈 서열을 2시간 내에 1,000달러의 비용으로 해독할 수 있다고 한다 (IDG, 2012).

그렇다면 얼마나 규모가 큰 데이터가 빅데이터인가? 아이비엠이 2012년에 1,000명이 넘는 관련 분야 전문가들을 대상으로 실시한 설문 결과에 따르면 절반이 넘는 응답자가 적어도 1테라바이트는 넘어야 빅데이터라고 부를 수 있다고 하였다(슈렉 외, 2012). 1테라바이트는 디브이디(DVD) 220장의 저장 용량과 맞먹는 규모이다. 그런데 서두에 언급한 대로 매일매일 생성되는 데이터가 엑사바이트급인 오늘날 테라바이트급이 아니라 페타바이트급 데이터도 그리 희귀하지는 않다.

앞서 언급한 대로 데이터의 규모는 기본적으로 상대적인 개념일 수밖에 없다. 점점 더 많은 데이터가 생성될 것이고 이를 저장할 수 있는 저장 매체의 용량도 점점 더 커질 것이다. 또 한 가지 데이터 규모의 상대성에

영향을 미치는 요소는 데이터의 종류이다. 예를 들어, 같은 용량의 데이터라고 해도 그 데이터가 데이터베이스에 저장된 정형 데이터인지 동영상 데이터인지에 따라 전혀 다른 데이터 처리 방법이 요구되므로, 빅데이터의 정의는 특정한 유형의 데이터가 사용되는 산업과 응용 분야에 따라 달라진다.

결국 데이터의 규모는 중요한 빅데이터의 정의 요소 가운데 하나임에 틀림없지만 어느 정도 규모가 빅데이터에 해당한다고 규정하는 것은 의미가 없다. 따라서 앞서 보인 가트너의 정의와 같이 '혁신적인 형태의 자료 처리 방법이 필요할 정도의 규모'라고 정의하는 것이 합리적이라고 결론지을 수 있다. 이러한 생각을 적극적으로 확장하면, 규모로는 스몰데이터이더라도 그 데이터를 바라보는 새로운 관점과 시각이 동반된 새로운 방식의 자료 처리와 해석이 더해진다면 빅데이터로 볼 수 있을 것이라는 주장도 가능할 수 있다.

두 번째 V는 '속도'를 뜻하는 벨로시티(Velocity)이다. 빅데이터는 그 규모가 클 뿐만 아니라 매우 빠른 속도로 생성되는 데이터를 말한다. 앞서 보인 바와 같이 빅데이터는 다양한 원천으로부터 생성되는데 이들 원천의 공통된 특징은 데이터 생성 속도가 매우 빠르다는 것이다. 기상 정보 측정 장치는 시시각각으로 변하는 기상 정보를 측정하여 기록하는데 그 데이터 생성 속도는 오로지 미리 정한 데이터 측정 사이클에 의해 결정된다. 미국의 대형 마트 체인인 '월마트'에서는 시간당 100만 건 이상의 거래 정보를 처리한다고 한다(쿠키어, 2010).

소셜 네트워크 서비스(SNS)에서 생성되는 메시지 역시 매우 빠른 속도로 생성된다. 우리는 소셜 네트워크 서비스를 통하여 엄청난 속도로 소식이 퍼져 나가는 것을 여러 번 목격한 바 있다. 소식의 확산은 바로 데이터의 생성에 의해 이루어지는 것이다. 오늘날 스마트폰으로 대표되는 휴대 가능한 기기의 확산으로 개인화된 데이터가 엄청난 속도와 양으로 생성되고 있음은 널리 알려진 사실이다.

이렇듯 데이터의 엄청난 대응 속도에 반응하여 데이터의 분석 또한 실시간으로 이루어져야 하는 요구가 발생하였다. 실시간 교통 안내 시스템은 그러한 예의 하나다. 지속적으로 수집되는 교통량과 통행 정보를 바탕으로 한, 실시간으로 최적화된 교통 안내를 할 수 없는 시스템은 아무런 쓸모가 없다. 유용한 정보를 제공하는 빅데이터의 실시간 분석을 위해서는 새로운 기술의 개발이 필수적으로 요구된다. 그런 기술의 목표는 고속으로 생성되어 사라져 갈 수밖에 없는 데이터로부터 실시간으로 유의미한 정보와 지식을 산출해 내는 것이다.

세 번째 V는 '다양성'을 뜻하는 버라이어티(Variety)이다. 역시 앞서 언급한 대로 빅데이터는 다양한 원천으로부터 생성되기에 다양한 형태를 지닌다. 과거 컴퓨터를 이용한 데이터 처리에 있어서 처리 대상은 각종 측정치, 계산값 등을 기록한 수치 데이터가 주종을 이루었으며, 텍스트가 포함되었다고 해도 가로와 세로가 잘 구성된 목록형 데이터, 다른 말로 정형 데이터가 대부분이었다. 그런데 컴퓨터의 활용 분야가 확대되면서 전자우편, 소셜 미디어 포스팅과 같은 비정형 텍스트, 그리고 방대한 음성과 영상 데이터가 축적되고 있다. 글머리에서 언급한 바와 같이 이러한 반정형, 혹은 비정형 데이터가 오늘날 생성되고 있는 데이터의 80% 이상을 차지할 것이라 추정되고 있다.

다양한 형태의 데이터에 대한 관심은 빅데이터가 유행하기 전에도 존재하였다. 그런데 다양성이 빅데이터의 정의 요소로까지 중요해진 것은 최근의 기술적 진보와 무관하지 않다. 예를 들어, 최근 이미지 처리 기술이 급격히 발전하여 안면 인식을 통해 고객의 성별과 나이 등을 파악하고 이를 마케팅에 활용하는 일이 현실화되기 시작했다(간도미와 하이더, 2015). 즉 종전에는 축적되기는 하여도 제대로 활용할 수 없었던, 특수한 처리 기술이 요구되는 다양한 형태의 데이터가 풍부하고도 유용한 정보를 제공할 수 있는 소중한 자원의 지위를 갖게 되었다.

이상으로 가트너의 '빅데이터의 정의'에 나타난 3V, 즉 '볼륨(Volume), 벨로시티(Velocity), 그리고 버라이어티(Variety)'에 대하여 알아보았다. 이들 속성은 빅데이터에만 있는 고유 속성이라기보다는 빅데이터라는 '현상'을 이해하기 위한 상대적인 속성으로 이해해야 한다. 이 상대적 속성을 드러내는 핵심 요소는 혁신적인 데이터 처리와 해석 방법에 있다.

한편 위의 3V에 더하여 몇몇 기업들이 다음과 같은 V를 추가로 제시하였다(간도미와 하이더, 2015).

- Veracity: 아이비엠(IBM)이 추가한 네 번째 V로 '진실성'을 뜻한다. 이는 데이터 원천의 특성상 어느 정도 존재할 수밖에 없는 데이터의 불확실성, 비신뢰성 등을 지적한 것이다. 특히 소셜 미디어 등에 나타난 소비자의 의견 등은 어느 정도의 불확실성을 가질 수밖에 없다. 그러나 이러한 데이터의 유용성 자체를 부정할 수는 없다.
- Variability: 새스(SAS)는 빅데이터의 추가 속성으로 variability와 complexity, 즉 '가변성'과 '복잡성'을 제시하였다. 가변성은 데이터 생성 속도가 변할 수 있음을 지적한 것이고 복잡성은 데이터가 단일 원천으로부터가 아니라 여러 원천이 복잡하게 뒤엉켜 있는 상태에서 생성될 수 있음을 지적한 것이다.
- Value: 오라클(Oracle)이 추가한 것으로 '가치'를 뜻한다. 오라클에 의하면 빅데이터에 포함되는 원본 데이터들은 상대적으로 규모에 비해 가치가 적다. 그런데 이와 같이 '저가치 밀도'의 데이터를 대량으로 분석했을 때에 큰 가치가 창출될 수 있다는 것이다.

2.2. 소셜 빅데이터

앞서 빅데이터의 80% 이상이 비정형 데이터로 추정되며, 이 가운데 특히 텍스트 데이터가 매우 중요한 위치를 차지하고 있음을 언급하였다. 빅데이터를 구성하는 텍스트 데이터의 주요 원천은 소셜 미디어이다.[4)]

소셜 미디어는 인간의 사고와 행위를 인간 스스로 기록하여 생성하는 장이라는 점에서 특수한 가치를 지니고 있다. 사용자들은 소셜 미디어를 통해 연결되어 온라인 환경에서 새로운 공동체를 형성하고 서로의 생각과 일상을 공유하며 그 기록을 텍스트로 남긴다. 이와 같은 현상은 온라인 공동체의 형태를 지닌 온라인 카페나 게시판에서도 관찰된다. 나아가 포털 서비스의 포스팅, 뉴스 기사 등에 대한 댓글 또한 온라인 공간에서의 미디어 소비와 여론 생성 현장을 고스란히 기록하고 있다.

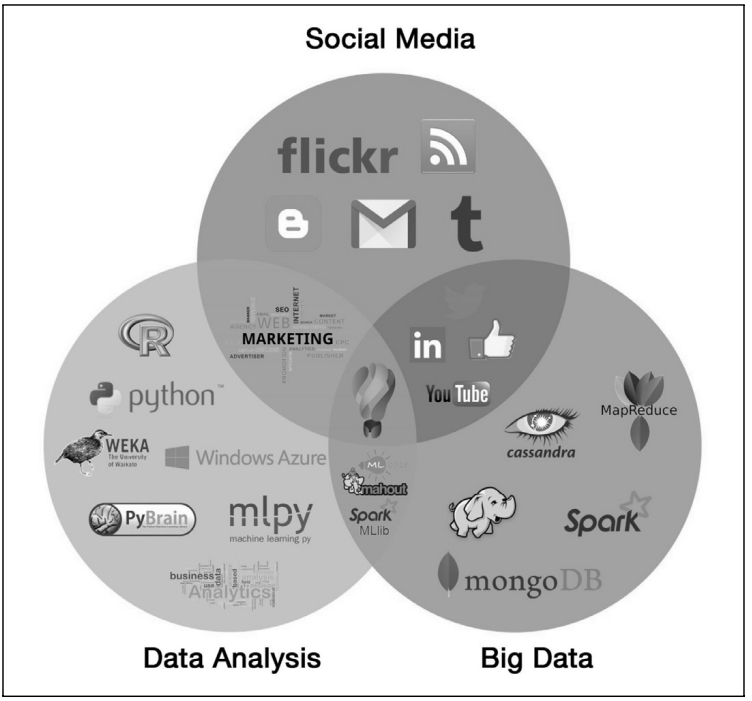
위와 같은 배경에서 미디어, 혹은 플랫폼으로서의 소셜 미디어와 빅데이터 현상이 결합된 소셜 빅데이터라는 개념이 등장하였다(송길영, 2012, 2015; 벨로-오르가즈 외, 2016). 언어 자료로서의 빅데이터를 이야기할 때 소셜 빅데이터의 개념을 다루지 않을 수 없다. 그러므로 이 글에서 빅데이터라 함은 곧 소셜 빅데이터를 가리킨다. [그림 2]는 벨로-오르가즈 외(2016)에서 보인 소셜 빅데이터의 개념을 나타내는 그림이다.

벨로-오르가즈 외(2016)가 [그림 2]를 통해 특히 강조하고자 하는 것은 소셜 빅데이터 분석이 근본적으로 학제적이라는 것이다. 이 논문에서는 관련된 분야로 데이터 마이닝, 기계 학습, 통계학, 그래프 마이닝, 정보 검색, 언어학, 자연 언어 처리, 시맨틱 웹, 온톨로지, 빅데이터 컴퓨팅을 나열하고 있다. 즉 수치 데이터와 정형 데이터 중심의 일반적인 빅데이터에 비해 비정

4) 소셜 미디어는 "개방, 참여, 공유의 가치로 요약되는 웹 2.0 시대의 도래에 따라 소셜 네트워크의 기반 위에서 개인의 생각이나 의견, 경험, 정보 등을 서로 공유하고 타인과의 관계를 생성 또는 확장할 수 있는 개방화된 온라인 플랫폼"으로 정의된다(위키피디아). 소셜 미디어에는 블로그, 소셜 네트워크 서비스, 위키, 마이크로 블로그 등이 포함된다.

형의 텍스트 중심이 되는 소셜 빅데이터의 분석에서는 언어학, 자연 언어 처리 분야 등의 참여가 필수적으로 요구되는 것이다. 바꾸어 말하면 소셜 빅데이터야말로 언어학의 연구 대상이라고 할 수 있다.

그림 2 소셜 빅데이터의 개념(벨로-오르가즈 외, 2016)



3. 언어 자료로서의 빅데이터

3.1. 균형 말뭉치와 모니터 말뭉치

언어 연구를 위한 대표적 언어 자료인 말뭉치는 여러 가지 기준에 따라 다양한 유형으로 분류할 수 있다. 그 가운데 하나는 원자료의 수집 방식에

따라 말뭉치를 균형, 혹은 표본 말뭉치와 모니터 말뭉치로 구분하는 것이다(맥에너리 외, 2011).

균형 말뭉치란 탐구 대상이 되는 언어 전체, 즉 모집단의 표본 자료로서의 역할에 충실하기 위해, 즉 대표성을 극대화하기 위해 말뭉치에 포함되는 원자료들을 다양한 매체, 장르에 걸쳐 선정하고 이들의 포함 비율을 적절히 조절하여 구성한 말뭉치를 말한다(맥에너리 외, 2006). 그러므로 균형 말뭉치는 조심스럽게 준비된 원칙에 따라 한 번 구성되면 그 내용과 규모가 고정된다.

이에 비해 모니터 말뭉치는 시간이 흐름에 따라 규모가 점점 커지는 말뭉치를 말한다. 원자료의 추가는 연간, 월간, 그리고 일간으로 이루어질 수도 있다. 말뭉치의 구성을 미리 설계할 수 없으므로 포함된 원자료의 균형성을 보장하는 것은 불가능하다. 모니터 말뭉치의 개념은 싱클레어(1991)에서 처음 주창된 것으로 균형 말뭉치의 폐쇄성이 살아 있는 언어 현상의 발굴, 그리고 매우 드물게 발생하지만 유의미한 언어 현상의 관찰에 적합하지 않다는 점을 지적하며 고안된 것이다.

맥에너리 외(2006)에서는 모니터 말뭉치의 강점으로 언어 변화의 관찰이 가능하다는 점을 들면서 신어의 등장과 사멸 등에 대한 연구를 예로 들었다. 또한 모니터 말뭉치가 매우 오랜 기간 축적되면 문법의 변화 등도 추적이 가능할 것이라고 보았다. 그러나 모니터 말뭉치는 균형성을 보장할 수 없으므로 신뢰도가 높은 통계 정보의 추출이 불가능하고 내용과 규모가 고정되지 않으므로 연구 결과의 비교가 불가능하다는 문제점을 지적하였다.

3.2. 말뭉치로서의 웹

모니터 말뭉치와 유사한 개념으로 말뭉치로서의 웹이 있다(킬가리프와 그레펜슈테트, 2003). 현재 주어진 가장 방대한 언어 자료임에 틀림없는

웹을 언어 연구에 활용하려는 시도이다. 웹에 존재하는 모든 언어 자료를 오프라인 사용을 위해 저장하는 것은 불가능한 일이므로 이 접근에서는 구글 등의 상용 검색 엔진, 그리고 언어 연구를 위해 특별히 고안된 인터페이스인 '웹콥(WebCorp)'(르누프, 2003)을 사용한다.

말뭉치로서 웹을 사용하는 가장 큰 장점은 역시 방대한 규모로부터 온다. 항상 그러한 것은 아니지만 보통의 균형 말뭉치에서 그 용례를 찾기 어려운 비교적 희귀한 언어 현상의 경우에도 웹에서는 상당히 많은 건수의 용례를 찾을 가능성이 있다. 또한 모니터 말뭉치의 경우와 마찬가지로 새로이 생성되는 데이터의 반영이 매우 빠르므로 새로이 등장한 단어나 표현의 추적에도 매우 유용하다.

그러나 말뭉치로서 웹을 사용할 때에는 검색 엔진 등의 제한적인 방법을 사용할 수밖에 없다는 근본적인 한계에서 벗어나기 어려워 그 용도가 제한될 수밖에 없다. 또한 검색 엔진의 검색 결과 수의 표시가 어떤 과정을 통해 생성되는지 알 수 없기 때문에 안정적인 통계 데이터를 얻을 수 있다는 보장을 하기 어렵다.

3.3. 말뭉치로서의 빅데이터

위의 원자료 수집 방식에 따른 말뭉치의 유형 분류로 보면 빅데이터는 모니터 말뭉치의 부류에 속한다. 한편 웹에서 생성된 데이터로 구성되었다는 점에서 말뭉치로서의 웹의 성격도 어느 정도 지니고 있다. 다만 빅데이터는 분석을 위해 모든 데이터를 데이터 원천으로부터 수집, 저장하여 사용하기 때문에 말뭉치로서의 웹이 갖는 한계는 없다.

말뭉치로서 빅데이터가 갖는 첫 번째 가치는 그 규모이다. 글쓴이의 일터에서 경험한 바에 따르면 대표적인 마이크로 블로깅 서비스인 트위터에서 생성되는 한국어 작성 트윗은 하루에 최소 500만 건에 이르며 대표적인

블로그 서비스인 네이버 블로그에서 생성되는 블로그 포스트는 하루 최소 50만 건에 이른다. 이 규모는 물론 말뭉치로서의 웹의 규모에는 훨씬 미치지 못하지만 그 어느 한국어 말뭉치보다도 규모가 크다.[5]

물론 아무리 규모가 크다고 해도 모든 한국어 사용자가 트위터, 혹은 블로그 서비스를 이용하는 것이 아니며 여기에 한국어의 양상이 모두 반영되어 있다고 할 수는 없다. 그럼에도 불구하고 한국어의 언어적 특성에 대한 탐구에 있어서 기존 말뭉치가 주지 못하는 풍부한 용례를 제공할 수 있다. 또한 맥에너리 외(2006)가 모니터 말뭉치에 대하여 지적한 대로 빅데이터는 균형성과 대표성에 있어서 문제가 있다고 볼 수 있다. 균형성과 대표성은 통계적 유의성에 기반을 둔 일반적인 통계적 연구 방법, 즉 모집단으로부터 추출한 비교적 작은 규모의 표본에서 통계적 유의성을 바탕으로 결론을 도출하고 이를 모집단으로 일반화하는 연구 방법에서 제기되는 문제이다. 그런데 모집단은 아닐지라도 모집단의 상당 부분을 포함하는 빅데이터에 있어서는 통계적 유의성이 그렇게 큰 의미를 갖지 못한다(간도미와 하이더, 2015). 그러므로 빅데이터로부터 통계적 정보를 얻기 위해서는 기존의 통계적 방법이 아닌 새로운 방법의 개발이 요구된다.[6]

대규모 말뭉치로서 빅데이터가 지니는 진정한 가치는 데이터의 원천인 소셜 미디어의 특성에서 찾아야 할 것이다. 예를 들어, 빅데이터는 언어 사용의 맥락과 언어 공동체에 대한 새로운 시각을 제공할 수 있을 것이다. 빅데이터에 포함된 언어 자료를 생성한 사람들은 넓게 보면 한국어라는 특정한 언어를 사용하며 이 시대를 살아가는 언어 공동체의 일원이다. 그러나 자료 생성자들은 지역, 직업, 연령, 관심사 등에 따라 각자 다른 맥락에서 한국어를 사용한다.

---

5) 한국어 트윗과 네이버 블로그 포스트의 일일 생성량은 글쓴이의 일터에서 측정한 것으로 실제 생성량과는 차이가 있을 것이다.

6) 특히 최근에는 베이지안 통계 기법의 활용이 여러 분야에서 시도되고 있다(알렌비 외, 2014; 스콧 외, 2016).

빅데이터는 언어 사용의 현장에서 동떨어지고 고립되어 존재하는 언어의 조각들이 아닌 무한히 확장될 수 있는 맥락 속의 언어를 들여다볼 수 있게 해 준다. 이를 통해 진정으로 동적인 언어 공동체의 생성과 발전의 양상을 살펴볼 수 있을 것이다. 이는 빅데이터가 단발적인 언어 사용을 담는 것에서 그치는 것이 아니라 다양한 환경에 처한 매우 많은 언어 사용자들의 언어 사용 양상을 비교적 장시간 지속적으로 담을 수 있기에 가능한 일이다.

말뭉치로서 빅데이터가 갖는 또 하나의 가치는 앞서 논의한 데이터 생성 속도와 관련이 있다. 소셜 미디어 서비스, 특히 마이크로 블로깅 서비스인 트위터에서는 초 단위로 새로운 트윗이 생성된다. 이를 통해 언어 연구자들은 언어 현장의 시간성을 정확히 파악할 수 있다. 특정한 발화가 이루어진 계절, 날짜, 시간은 물론이고 그 발화에 영향을 미쳤을 수도 있는 언어 외적 요소들에 대한 추적도 어느 정도 가능하다. 예를 들어, 우리 사회에 큰 영향을 미친 사건이 사람들의 언어 사용에 끼친 영향들을 관찰할 수 있을 것이다. 또한 특정한 언어 사용 양상이 사람들 사이에서 어떻게 펴져 나가는지, 즉 언어 사용 양상의 확산에 관한 연구도 가능할 것이다.

앞서 언급한 대로 빅데이터의 주요 속성 가운데 하나는 그 형식의 다양성이다. 이제까지의 언어 연구는 어쩔 수 없는 기술적, 또는 자료 수집의 제약으로 글말 중심으로 이루어져 왔다. 그런데 최근 기술의 발전 양상을 볼 때에 동영상을 언어 연구에 적극적으로 활용하게 될 날이 그리 멀지 않아 보인다. 먼저는 동영상에 포함된 음성의 인식이 가능하게 될 것이다. 이어서 동영상의 배경과 참여자를 인식하여 수많은 동영상을 자동으로 분류하고 이를 맥락화하는 일이 가능해질 것이다. 이는 언어 연구의 방법론과 대상에 있어서 작지 않은 변혁을 불러올 것으로 기대된다.[7]

---

7) 앞서 기술한 대로 빅데이터는 언어 연구의 대상과 방법에 상당한 변화를 가져올 것으로 보인다. 글쓴이는 한걸음 더 나아가 빅데이터가 기존 언어학의 확장이 아닌 전혀 새로운 시각의 언어학 출현, 즉 패러다임의 변화를 불러일으키지 않을까 조심스럽게 점쳐 본다.

4. 빅데이터 활용의 절차의 기술적 요건

앞에서 언급하였듯이 빅데이터를 언어 자료로 활용하기 위해서는 일정한 절차를 거쳐야 하며 각 절차에는 적절한 기술적 요건이 따른다. 이 글에서는 라브리니디스 외(2012)에서 도식화한 빅데이터 분석의 과정을 언어 연구의 관점에 맞추어 설명한다.

4.1. 데이터 수집

빅데이터를 언어 연구에 활용하기 위한 가장 첫 단계는 데이터 수집 단계이다. 소셜 미디어 서비스로부터의 데이터 수집에는 크게 세 가지 방법을 이용할 수 있다.

(1) 데이터 제공 서비스 이용

소셜 미디어 서비스로부터 데이터 제공 업무를 대행하는 업체의 서비스를 이용하는 방법으로, 가장 안정적으로 데이터를 수집할 수 있다. 대표적인 서비스 업체로는 트위터 데이터를 공급하는 '지닙(GNIP, www.gnip.com)'이 있다. 이 업체의 서비스를 이용하면 실시간으로 생성되는 모든 트윗, 혹은 표본 데이터를 수집할 수 있다. 이 업체에서 제공하는 가장 특징적인 서비스는 과거에 작성된 트윗에 접근할 수 있도록 해 주는 서비스이다. 과거

현존하는 가장 영향력 있는 과학 철학자 중 한 사람인 이언 해킹은 그의 저서 ≪우연을 길들이다≫에서, 19세기 초까지 모든 과학을 지배하던 결정론적인 믿음을 뚫고 다른 어느 법칙이나 원리로 환원될 수 없는 '우연'이라는 개념이 받아들여지는 과정을 보였다. 해킹은 우연과 확률은 과학에 있어서 거대한 사고의 전환을 가져 왔으며, 오늘날 가장 엄정한 과학으로 인정받는 양자론의 근간을 불확정성의 원리가 이루게 되었음을 논증하였다.
빅데이터는 자연 과학이 경험한 패러다임의 변화를 언어학도 마찬가지로 경험하게 될 것이라 믿는다. 이세돌과 알파고의 바둑 대국을 보면서 과연 알파고가 바둑을 이해하고 있는지에 대한 논쟁이 벌어졌던 것처럼 인간과 대화를 나누고 소설을 쓰는 컴퓨터가 과연 인간의 언어를 정확히 이해하고 있는지에 대한 논쟁이 벌어지는 날이 올 것이고, 그때 우리는 언어, 그리고 언어 연구에 대한 생각을 많이 바꾸어야 할지도 모른다.

에 생성된 트윗을 수집할 수 있는 유일한 방법은 이 서비스를 이용하는 것이다. 이와 같은 장점을 지닌 이 서비스를 사용하는 데에 있어서 가장 큰 난관은 사용료이다.

(2) 오픈 에이피아이(Open API) 사용

두 번째 방법은 소셜 미디어 서비스 업체에서 제공하는 오픈 에이피아이(Open API)를 이용하여 데이터를 수집하는 방법이다. 소셜 미디어 서비스는 다른 서비스와의 연동이 매우 중요하므로 서비스 업체에서는 다양한 형태로 데이터를 생성하거나 데이터에 접근할 수 있는 오픈 에이피아이를 제공한다. 오픈 에이피아이를 사용하기 위해서는 이를 주어진 규격에 따라 사용하는 컴퓨터 프로그램을 작성해야 한다.[8]

트위터의 경우 트윗의 수집에 이용할 수 있는 샘플 에이피아이, 검색 에이피아이, 스트리밍 에이피아이, 그리고 레스트 에이피아이를 제공한다. 이 가운데 스트리밍 에이피아이는 검색어를 지정하여 실시간으로 생성되는 트윗들을 수집할 수 있도록 해 준다. 한 번 실행할 때에 지정할 수 있는 검색어의 수에 제한이 있고 에이피아이 호출 간격에도 시간제한이 있기 때문에 대량의 트윗 수집을 위해서는 여러 컴퓨터에서 수집 프로그램을 구동해야 한다.

(3) 웹 접근 수집

마지막 방법은 오픈 에이피아이가 제공되지 않는 자료원으로부터 데이터를 수집할 때에 사용하는 방법으로, 인간이 웹브라우저를 통해 해당 서비스를 이용하는 것을 흉내 내는 프로그램을 작성하여 데이터를 수집하는 것이다.

---

8) 트위터 오픈 에이피아이(Open API)를 쉽게 사용할 수 있도록 도와주는 라이브러리들이 프로그래밍 언어별로 존재한다.

데이터 접근 스케줄링을 비롯한 많은 고려 사항이 따르는 방법이나 에이피아이가 제공되지 않는 서비스에 대한 유일한 데이터 수집 방법이다.

4.2. 데이터 정제와 정보 추출

많은 경우에 수집된 자료는 바로 사용할 수가 없고 일정한 정제 과정을 거쳐야 한다.

(1) 필터링

필터링이란 연구 목적에 부합하지 않거나, 나아가 연구 목적 성취에 방해가 되는 데이터를 걸러내는 과정이다. 트위터의 경우 자동으로 트윗을 생성하는 '봇'의 트윗을 제거한다든지, 이벤트성 트윗을 제거한다든지 등의 처리를 할 수 있다. 블로그의 경우 상당수를 차지하는 광고성 포스트를 제거할 수 있다. 물론 이 과정은 연구 목적에 따라 다른 접근을 하게 될 수도 있다.

(2) 중복 제거

소셜 미디어에서 생성된 데이터는 다양한 형태의 데이터 중복이 존재한다. 트위터의 경우에는 '리트윗'이라는 형태의 적극적인 데이터 전파 기능이 있어서 데이터 중복이 발생한다. 블로그의 경우에도 소위 '퍼나르기'에 의한 데이터 중복이 발생한다. 이러한 데이터 중복을 어떻게 처리할 것인가도 연구 목적에 따라 결정된다.

(3) 가공

데이터 가공은 오픈 에이피아이가 아닌 웹 접근 수집에 모아진 데이터일 경우 주로 이루어져야 하는 일이다. 즉 렌더링을 위해 부가된 에이치티엠엘(HTML) 태그 등을 제거하고 순수 텍스트만 추출하는 과정을 거쳐야 한다.

단순히 제거할 뿐만 아니라 최소한의 구조적 정보인 포스트의 제목, 본문을 구분하고 작성자, 작성 날짜와 시간, 태그 등을 분절해야 한다.

(4) 언어 처리

정보 추출 단계에서 이루어져야 할 일은 언어 처리이다. 언어 처리라 함은 자동화된 언어의 형식적 분석을 말하는데 현실적으로 한국어 데이터에 대하여 할 수 있는 언어 처리는 형태소 분석이다. 형태소 분석이 이루어지지 않은 데이터를 언어 연구에 이용하는 일은 불가능하지는 않다. 그러나 많은 경우에 형태소 분석은 효과적인 언어 연구를 위한 최소한의 언어 처리 단계일 것이다.

과거에는 일반 연구자들이 자동화된 데이터의 처리에 사용할 수 있는 형태소 분석기가 거의 없었지만 최근에는 무료로 사용할 수 있는 공개 형태소 분석기들이 등장하여 많은 연구자에게 큰 도움이 되고 있다. 그러나 형태소 분석기를 연구 목적에 맞게 조절하여 사용하는 일은 결코 쉽지 않은 일이다.

4.3. 데이터의 구조화와 통합

데이터의 구조화는 연구자들이 언어 처리가 적용된 데이터에 쉽고도 효과적으로 접근할 수 있도록 해 주는 일이다. 즉 자소, 음절, 형태소, 어절, 연어, 구 등의 언어 단위별로 다양한 질의 조건을 부가하여 데이터에 접근할 수 있어야 한다.

또한 데이터 통합에 의해 다양한 원천으로부터 수집된 데이터를 하나로 통합하여 접근할 수 있어야 하며 각종 메타 데이터에도 접근이 가능해야 한다.

매우 방대한 양의 데이터를 효율적으로 저장해야 하기 때문에 여러 대의

컴퓨터로 이루어진 분산 파일 시스템이나 분산 데이터베이스를 사용해야 하는 경우가 있다.[9]

4.4. 데이터 모델링과 분석

이 단계에서는 구조화된 데이터로부터 데이터를 효율적으로 질의하여 데이터에 대한 분석이 이루어져야 한다. 예를 들어, 특정 단어의 의미 변화에 대한 연구를 수행한다면 그 단어의 의미를 파악할 수 있는 실마리 문맥을 분류하고 그 변화를 추적할 수 있어야 한다.

빅데이터를 활용할 때에는 매우 많은 양의 데이터를 사용하게 되므로 자동화된 데이터 마이닝 기법의 도움을 받지 않을 수 없다. 다양한 데이터 마이닝 기법이 언어 연구에 어떻게 접목될 수 있는지에 대해서는 다양한 실험과 검증을 통해 밝혀져야 할 것이다.

이 단계에서는 통계적 분석도 수행하게 된다. 앞서 언급한 바와 같이 통계적 유의성에 기반을 둔 전통적 통계 분석 방법이 빅데이터에서는 큰 의미가 없다는 지적이 있다. 그러나 그 대안은 아직 마련되지 않았다.

한편 웹 규모의 빅데이터를 이용한 언어 처리의 경험을 간략히 요약한 해일러비 외(2009)는 빅데이터를 이용한 언어 연구에서도 참고할 만하다. 이 논문에서는 다음과 같은 '교훈'을 역설한다.

- "존재하지 않는 주석된 데이터를 기대하지 말고 존재하는 대규모의 데이터를 이용하라." 데이터를 이용한 연구에서는 탐구 대상 데이터를 해석하고 이용하기에 편리한 주석을 중요하게 여긴다. 나아가 주석된

9) 빅데이터의 분산 저장과 처리에 관련하여 많은 기술적 진보가 있었고 지금도 진행 중이다. 특히 아파치 하둡(http://hadoop.apache.org)과 아파치 스파크(http://spark.apache.org)는 오늘날 빅데이터 처리의 핵심 기반 기술이다.

데이터의 부재가 연구의 발전을 가로막는 장애임을 지적하기도 한다. 언어 연구에 있어서 주석된 데이터라 함은 형태소, 단어, 구, 문장 등의 언어 단위의 분절과 최소한의 해석이 이루어진 데이터를 말할 것이다. 이러한 언어 주석 데이터가 언어 처리와 언어 연구에 큰 도움이 됨은 틀림이 없다. 그러나 이러한 주석 데이터를 구축하는 데에는 엄청난 비용과 시간이 소요되며, 일반적인 규모를 훨씬 뛰어넘는 빅데이터에 주석을 부가하는 일은 비현실적이다. 그러므로 주어질 가능성이 거의 없는 주석 데이터에 의존하지 않고 대규모로 주어지는 원시 데이터를 어떻게 이용할 수 있는지에 대하여 깊이 고민해 보아야 한다.

- "정교하고 일반화된 규칙보다는 개별 사실에 집중하라." 이 논문의 저자들은 최근의 기계 번역에서 기억된(memorized) 개별 번역 사례의 중요성을 예로 들면서 일반화된 규칙보다 개별 사실을 최대한 이용할 것을 권장한다. 이 교훈은 언어 현상을 간명히 설명할 수 있는 일반화된 규칙의 작성에 관심을 두는 언어학 연구에서는 받아들이기 힘들 수도 있다. 다만 소규모 데이터에서 도출된 규칙은 언제든지 그 적용 범위에 한계가 올 수 있다는 점을 알아야 한다는 점에는 동의할 수 있을 것이다. 나아가 지식의 표현이 일차술어논리 형식의 간결한 규칙으로 되어야만 한다는 것 또한 편견일 수 있다는 사실을 인정해야 한다. 수많은 개별 사실과 개별 사실들의 조합으로부터 도출된 확률적 표현 또한 훌륭한 지식 표현의 방법 가운데 하나이다.

4.5. 결과의 해석

가장 어려운 단계이다. 연구 가설이 주어진 연구였다면 빅데이터에 의해 가설이 지지되는지 그러지 않는지를 검증하여야 하며, 연구 가설이 주어지지 않은 탐색적 연구였다면 연구 결과가 다른 연구로 이어질 수 있도록 정리해야

한다.

결과의 해석을 효과적으로 전달하기 위하여 적절한 시각화 기법의 활용을 적극적으로 고려해 볼 필요가 있다. 방대한 데이터로부터 도출된 복잡한 결론을 글로만 표현하는 데에는 한계가 있을 때가 많기 때문이다.

5. 맺는말

이 글에서는 빅데이터의 특성을 먼저 살펴보고, 빅데이터, 특히 소셜 빅데이터가 언어 연구에 새로운 전기를 마련해 줄 수 있는 언어 자원으로서의 가치가 있음을 논하였다. 이어서 빅데이터를 언어 연구에 활용하기 위한 절차를 기술적 요건과 함께 간략히 설명하였다.

앞서 언급한 대로 빅데이터를 언어 연구에 활용하는 일은 아직 걸음마 단계에 있다. 그리고 해결해야 할 문제도 다수 존재한다. 특히 개인 정보 보호의 문제는 연구 윤리에 있어서 매우 중요한 문제이다. 또한 비즈니스의 목적으로 서비스되고 있는 데이터를 이용하기 때문에 데이터의 공유 등에 있어서 자유롭지 못한 부분이 많은 것도 문제이다.

이미 우리는 빅데이터의 시대에 살고 있고 어떠한 형태로든 빅데이터와 연관이 되어 있다. 이러한 시대에 빅데이터를 언어 연구에 활용하는 것은 필연적인 일일 수도 있다. 활발한 토론과 다양한 시도가 이루어지기를 기대해 본다.

참고 문헌

송길영(2012), ≪여기에 당신의 욕망이 보인다≫, 쌤앤파커스.

_____(2015), ≪상상하지 말라≫. 북스톤.

이안 해킹 저·정혜경 역(2012), ≪우연을 길들이다≫, 바다출판사. / Hacking, I.(1990), *The Taming of Chance*, Cambridge University Press.

한국소프트웨어기술인협회 빅데이터전략연구소(2016), ≪빅데이터 개론≫, 광문각.

한국 IDG(2012), 빅 데이터의 이해, ≪IDG Tech Report≫. http://kbig.kr/index.php?sv=title&q=knowledge/pds_&tgt=view&idx=15326/(검색일: 2016. 5. 29.).

한국정보화진흥원 미래전략센터(2015), ≪2015년 빅데이터 글로벌 사례집≫. http://kbig.kr/index.php?sv=title&q=knowledge/pds_&tgt=view&idx=15614&sv=title/(검색일: 2016. 5. 29.).

한국정보화진흥원 ICT융합본부(2016), ≪2016 글로벌 빅데이터 융합 사례집≫. http://kbig.kr/index.php?sv=title&q=knowledge/pds_&tgt=view&idx=16137/(검색일: 2016. 5. 29.).

Allenby, G. M., Bradlow, E. T., George, E. I., Liechty, J. and McCulloch, R. E.(2014), Perspectives on Bayesian Methods and Big Data, *Customer Needs and Solutions*, 1(3): 169~175.

Bello-Orgaz, G., Jung, J. J. and Camacho, D.(2016), Social big data: Recent achievements and new challenges. *Information Fusion*, 28: 45~59.

Cukier, K.(2010). The Economist, Data, data everywhere: A special report on managing information. http://www.economist.com/node/15557443/(검색일: 2016. 5. 29.).

Gandomi, A. and Heider, M.(2014), Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35: 137~144.

Gartner, IT Glossary: Big Data. http://blogs.gartner.com/it-glossary/big-data/(검색일: 2016. 5. 29.).

Halevy, A., Norvig, P. and Pereira, F.(2009), The Unresonable Effectiveness of Data, *IEEE Intelligent Systems*, 8~12.

Economist(2015), The data deluge: Five years on. http://www.veritas.com/content/dam/Veritas/docs/reports/EIU-veritas-data-deluge.pdf/(검색일: 2016. 5. 29.).

IBM(2015), Big Data and Analytics. http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html/(검색일: 2016. 5. 29.).

Kilgarriff, A. and Grefenstette, G.(2003), Introduction to the special issue on the Web as Corpus, *Computational Linguistics*, 29(3): 333~347.

Labrinidis, A. and Jagadish, H. V.(2012), Challenges and opportunities with big data, *Proceedings of the VLDB Endowment*, 5(12): 2032~2033.

McEnery, T. and Hardie, A.(2011), *Corpus Linguistics: Method, Theory and Practice*, Cambridge University Press.

McEnery, T., Xiao, R. and Tono, Y.(2006), *Corpus-based Language Studies*, Rutledge.

Newscenter, Conversations on linguistics and politics with Noam Chomsky, 2016년 4월 18일 자. http://www.rochester.edu/newscenter/conversations-on-linguistics-and-politics-with-noam-chomsky-152592/(검색일: 2016. 5. 29.).

Renouf, A.(2003), WebCorp: providing a renewable data source for corpus linguists, S. Granger and S. Petch-Tyson(eds.) *Extending the Scope of Corpus-based Research: New Applications, New Challenges*, Rodopi, 39~58.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D. and Tufano, P.(2012), Analytics: The real-world use of big data. How innovative enterprises extract value from uncertain data, *IBM Institute for Business Value*. http://www-03.ibm.com/systems/hu/resources/therealworduseofbigdata.pdf/(검색일: 2016. 5. 29.).

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. and McCulloch, R. E.(2016), Bayes and big data: The consensus Monte Carlo algorithm, *International Journal of Management Science and Engineering Management*, 11(2): 78~88.

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

빅데이터의 정확한 언어와 모호한 언어

김형석
한국과학기술원(KAIST) 경영대학 교수(경제학 전공)

전주곡: '빅데이터'는 시대정신(Zeitgeist)인가?

바야흐로 '빅데이터' 시대이다. 소위 '빅데이터 전문가'라고 불리는 일군의 무리가 마치 중세 시대 복음서를 무시무시한 방식으로 전파하는, 어떠한 토라도 단다면, '마녀'로 낙인찍혀 재단의 불쏘시개가 될 것이라고 엄중한 예언을 하는, 구글 신(Google 神)은 알고 있다며 '구글 신, 구글 신 구글 신' 염불하는, '구글 신'이 못 박힌 십자가를 앞세워 이교도의 군대로 겁 없이 행군하는 '구글교' 순례자들의 환영을 재현한다는 의미에서의 '빅데이터' 시대이다. 좀 더 직접적으로 표현한다면 전쟁터이건 유대인 수용소이건 베를린의 음악당에서건 '바그너(Wagner)'를 합창하는 나치의 순례자들처럼, 국어학자의 학술지이건, 회장님의 집무실이건, 회장 사모님들의 미술관 '포럼'에서건, '빅데이터'를 화학조미료 뿌리듯이 뿌리는 '포스트모던'한 '셰프'의 감각과도 같다는 의미에서의 '빅데이터' 시대이다. 그나마 유대인의 목숨이 아니라 기업가의 돈주머니와 학생들의 돈주머니를 노린다는 점에서, '스와티카'와 히틀러의 ≪나의 투쟁≫이 아니라 미국 명문 M대학의 권위를 업은 명망가들이 작성한 '자연(Nature) 방법서'와 '과학(Science) 방법서'를 중심으로 대물주의(大物主義) 광기를 수출한다는 점에서, 나치보다는 인본주의적이다. 또한 이 글을

작성하는 필자의 목숨에 지장을 주지 않는다는 점에서 '빅데이터' 시대는 참으로 존경의 예를 갖출 만하다. 그렇다고 1945년 5월 1일 베를린 함락전의 나치처럼, 아주 미약한 존재라고 보기는 어렵다.[1] '구글교'의 순례자들이 '극악무도'하게 불쑥 내민 '빅데이터'라는 십자가 앞에서 '뱀파이어'라고 낙인찍힐 수 있다는 두려움을, 적어도 인문학자는 체감할 수 있기 때문이다. 그 두려움의 실체를 정확히 밝힘으로써 인문학자가 '뱀파이어'가 아님을 소명(疏明)하는 의미에서 아마도 필자에게 이 글의 의뢰가 왔을지도 모르겠다. 의뢰자의 '교환동기'의 가치를 중시하는 경제학자답게, 이 글에서는 '빅데이터'의 실체를 해부할 뿐만 아니라, 인문학자의 온전한 두려움 또한 해체하는 것을 목표로 한다.

'빅데이터'의 정의(定義)

인문학 우상 숭배 타파를 외치는 '빅데이터'의 홍위병들은 여러분의 돈주머니와 영혼주머니를 노리지만, 마오쩌둥(毛澤東)의 홍위병들이 그랬듯이, 그들이 외치는 구호 '빅데이터' 자체의 의미는 가늠하지 못한다. '빅데이터'란 단어 자체가 주는 '사전적' 의미는 단순히 말해 "데이터가 크구나!"라고 외치는 감탄문의 범주를 벗어나지 못한다. '남모를 열등감의 원인'이 '크기'에서 비롯될 때, "so big!"이라는 외침은 홍위병의 집단의식을 일깨우기 마련이다.[2] 따라서 전통적으로 왜소 콤플렉스의 악몽을 겪어 왔던 한국적 상황에서, 홍위병들이 외치는 '빅데이터'는 자생적인 재구조화 없이, 대물주의(大物主義)와의 야합(野合)을 통하여 시대의 총아, 새로운 패러다임의 재래로 위치가 격상되었다.

---

1) 베를린의 공식 함락일은 1945년 5월 2일이다.

2) '남모를 열등감의 원인'이라는 표현은 한강의 ≪채식주의자≫(창비, 2007)에서 발췌하였다.

그렇다면, 실제 '빅데이터'란 무엇을 의미하는가? '빅데이터'의 사전적 의미는 정보 수집 기법의 발전에 따른 관찰된 정보량의 비약적 증가, 그렇게 증가된 정보량의 저장 기술 확대를 의미한다. 특히, 여기에서의 정보량은 '디지털'화한 정보량, ≪표준국어대사전≫의 순화된 말로 표현한다면, 0과 1 두 개의 수치로 '환원'된 '수치적' 정보량을 말한다. 다시 말하면, 관찰 횟수와 저장 규모 두 개의 차원에서 보았을 때, '빅데이터'라 불리는 정보량의 물리적 크기는 말 그대로 "so big!"이다. 예를 들어, 한 사람의 사회적 경험에 대한 정보량을 얻기 위해, '빅데이터' 연구에서는 보통 한 사람에 관한 '사회적 행동 변수' 30개를 정하고, 그 30개 변수를 6분마다 18개월 동안 관찰한다.[3] 즉 한 사람에게서 얻는 정보량 또는 데이터는 대략 '30×240×30×18=3,888,000'이다. 주의해야 할 것은 이 데이터 수는 한 사람에게서만 얻어낸 것이라는 점이다. 그러나 '빅데이터'를 정의할 때 무엇보다 주목해야 할 것은, 그 정보량이 무엇에 관한 정보량인가? 즉 데이터가 지향하고 있는 지점이 어디인가를 파악하는 것이다. 사실 3,888,000이라는 숫자가 원자 또는 소립자에 대한 데이터 숫자라면 "so big!"이라는 감탄문은 '위선적'인 감탄에 가깝다. "외계인이 크다."라고 말하는 것은, 큰 것은 알지만 우리의 크기와 상관없기에 냉소적으로 말하는 "커요."에 가깝다. 우리의 감탄이 머무는 지점이 바로 인간의 삶에 관한 정보를 담고 있기에, 우리는 비로소 위선의 가면을 내려놓고 "정말로 커요."라고 말한다. 다시 말하면, '빅데이터'는 인간의 삶에 대한 수치화된 정보량이다. 인간의 삶에 관한 수치화된 정보량이 크기에, 우리는 이것을 '빅데이터'라고 부른다. 또한 30개의 '행동 변수'가 선사하는 3,888,000개의 숫자가 한 개인의 삶을 이해할 수 있다는 '믿음'도 '빅데이터'라는 언어에 반영되어 있다. '양으로 승부해서 질로 승화할 수 있다'는 '소박한 공대생'의

3) 알렉스 펜틀런드의 ≪창조적인 사람들은 어떻게 행동하는가≫(박세연 옮김, 와이즈베리, 2014) 참조.

믿음을 '빅데이터'라는 '시적' 용어로 표현한 것이다. 결국 정보 기술의 발전에 의한 방대한 양의 정보를 수집하고 저장할 수 있는 물리적 기술력이라는 '하부 구조(substructure)' 위에, '인간의 삶조차 양으로 승부하면 그 삶의 질을 이해할 수 있다'는 '상부 구조(superstructure)'를 가진 믿음 체계로 볼 수 있겠다.

삶의 이해 방식으로서의 '빅데이터': 공대생의 재도전

일단 믿음 체계가 형성되면, 특히 그 믿음 체계가 사람을 향해 있는 것이라면, 그 믿음 체계 특유의 삶을 이해하는 방식이 드러나기 마련이다. 믿음 체계로서의 '빅데이터' 또한 특유의 이해 방식을 선사하는데, 이 특유의 이해 방식은 삶을 이해하는 전통적인 방식과는 확연한 대척점에 있다.

우리는 우리의 삶을 이야기할 때, '모호하다'고 말한다. 국립국어원 ≪표준국어대사전≫에 의하면, '모호하다'는 "말이나 태도가 흐리터분하여 분명하지 않다."로 정의된다. 따라서 우리의 삶이 '모호하다'고 규정짓는 것은 우리의 삶이 흐리터분하여 분명하지 않고, 이런 이유로 그러한 삶을 '분명한 말'이나 '분명한 태도'라는 매개체(媒介體)로 표현한다 한들, '모호함'의 무게가 너무나 무겁기에, 그 무거움에 짓눌린 우리의 언어나 태도 또한 흐리터분하고 분명함과는 거리가 멀 수밖에 없다. 결국 모호한 우리의 삶을 분명한 언어·말로 이해한다는 것은 '모호함'의 코끼리를 '분명함'이라는 부처님의 손가락으로 지탱하려는 위태로운 곡예처럼 보인다. 가끔 찾아오는 위태로운 곡예는 우리 삶에 신선한 자극을 주고 생기를 북돋아 주지만, 그러한 곡예가 일상이라면, 그 곡예는 우리 삶의 일부가 될 수 없다. 즉 위태로운 곡예는 예외적일 때만 가치가 있다. 따라서 '모호'한 삶을 '분명한' 언어로 이해하다는 것은 우리의 삶을 이해하는 전통적인 방식이 아니다. 우리의 삶을 이해하

는 전통적인 언어 방식은 '모호한' 언어이다. '모호함'의 무게에 짓눌려 질식 상황을 처절하게 분명한 언어로 고발하기보다, 그 '모호함'의 무게를 온전히 받아들이는, 그 처절하게 숨 막히는 상황과 함께 산화하는 '메시아'의 메시지와도 같은 '모호한' 언어가 오히려 이율배반적으로 우리의 '모호한' 삶을 '분명히' 전달하는 것이, 삶을 이해하는 전통적인 언어 방식이다. 그래서 우리의 '모호한' 삶을 '모호한' 언어로 이해하는 전통적인 이해 방식의 정점은 바로 '시'이다. 김소월의 <산유화>라는 시를 통해 우리가 우리의 모호한 삶을 어떻게 '모호한' 언어로 이해하는지 알아보자.

산에
산에
피는 꽃은
저만치 혼자서 피어 있네.

최근 화제가 되었던 ≪시를 잊은 그대에게: 공대생의 가슴을 울린 시 강의≫(휴머니스트, 2015)에서 정재찬 교수가 지적했듯이, 시어 '저만치'가 의미하는 거리는 수치로 치환(置換) 가능한 물리적 거리가 아니다. 이 시의 화자와 꽃 사이의 거리 '저만치'를 1미터, 5미터, 10미터, 200미터로 이해하는 것은 삶의 '모호함'이라는 무게에 짓눌려 압사해 버린, '분명한 언어'란 이름을 가진 사내의, 피에 버무린 살점이 꽃잎처럼 산화하는 것을 보는 것과 같다. 시어 '저만치'는 '모호함의 천공(天空)'을 완연(完然)히 짊어질 수 있는, 모호한 존재였던 거인족(巨人族) 아틀라스처럼, 모호함의 무게를 원망 없이, 속절없이 감내하는 모호한 언어의 극치를 보여준다. 시어 '저만치'의 거리는 그냥 지나치기는 아쉬운 거리, 그러나 내 손길을 뻗친다 한들, 인연의 끈을 잡을 수 없는 거리, 아련히 다가오는, 그러나 이내 시나브로 멀어져 가는 '복학생 오빠'와의 거리, 그래서 가슴 설레는, 한(恨)이 이슬처럼 망울지는, '마음'의

‘모호함’을 드러내는 거리이다. 모호한 언어야말로 모호한 삶을 ‘분명히’ 전달할 수 있는 언어임을 김소월의 시어 ‘저만치’가 오히려 ‘정확히’ 보여준다. 화자와 꽃 사이에 벌어지는 삶의 모호한 만상(萬象)을 1미터, 5미터, 10미터, 200미터 등과 같은 ‘정확한 언어’로 이해하는 것은 언어의 폭력성 또는 언어술사의 재능 없음을 드러내는 ‘분명한’ 언어 방식이다. 그렇기에 ‘모호한’ 삶을 ‘모호한’ 언어로 ‘정확히’ 전달하는 시라는 이해 방식이야말로 인문학의 정점이자 극치라는 것이 전통적인 견해이다.

삶의 이해 방식으로서의 ‘빅데이터’는 바로 이 전통적인 견해가 선험적으로 배제해 왔던, ‘분명하고 정확한’ 언어 방식으로 화자와 꽃 사이의 거리를 ‘정확한 잣대’로 재고자 하는 시도이다. 화자와 꽃 사이에 벌어지는 삶의 모호한 만상을 1미터, 5미터, 10미터, 200미터 등과 같은 ‘정확한 언어’로 재단하는 것이 삶의 민낯을 드러낼 수 있고, 그러한 몸짓이 의미 있음을 보이고자 하는 시도이다. ‘양으로 승부하여 질로 승화할 수 있음’을 보이고자 하는 공대생의 재도전이기도 하다. 즉 ‘시’가 추구하는 방식과는 달리, ‘정확한’ 언어로 ‘모호한’ 삶을 이해하고자 하는 시도이다. ‘모호한 천공’을 수많은 숫자로 조합하고 빚어낸, 인공족(人工族) 아틀라스로 지탱하고자 하는 시도인 셈이다.

이제 ‘시’의 화자와 꽃 사이의 거리를 어떻게 ‘정확한 언어’, 즉 수치화된 정보량을 활용하여 이해할 수 있는지, ‘모호한 언어’를 대표하는 시어 ‘저만치’를 어떻게 ‘정확한 언어’로 대체할 수 있는지 ‘빅데이터’ 방식으로 알아보자. ‘빅데이터’ 방식으로 시어 ‘저만치’를 ‘명확한 빅데이터 언어’로 대체하기에 앞서, 먼저 일반적 의미에서의 ‘빅데이터 분석’을 소개하고자 한다.

문법적 은유(grammatical metaphor)로서의 '빅데이터 분석'

'빅데이터' 방식으로 삶을 이해한다는 것은 기본적으로 인간의 삶을 분석(分析)한다는 것이다. ≪표준국어대사전≫에 의하면, 분석이란 '얽혀 있거나 복잡한 것을 풀어서 개별적인 요소나 성질로 나눔'을 뜻한다. 즉 분석이란 분석의 주체가 분석의 대상 또는 객체를 얽혀 있거나 복잡한 그 무엇으로 전제한 후, 대상의 복잡함과 대비되는 '단순함'을 표상하는 개별 요소 및 성분으로 분해하여 대상의 실체를 '까발리는' 것을 의미한다. 따라서 '빅데이터'가 이해하는 삶은 분석의 대상이자, 얽혀 있고 복잡한 것이다. 인간의 모호한 삶은 '모호한' 시어로만 이해되는 신비로운 천공이 아니라, 단순히 얽혀 있어 복잡한, 그래서 우리가 풀어내 까발려야 하는 '실타래'일 뿐이다. 그 '실타래'를 풀기 위한 '가위'를 성분 또는 요소라 말할 수 있는데, '빅데이터 분석'에서는 그 '가위'를 '행동 변수'라 부른다. 다시 말해서, 인간의 '모호한' 삶을 '행동 변수'라는 '명확한 언어'로 분해하는 것이다. 결국 '빅데이터 분석'이란 '빅데이터'의 사전적 의미, 그 사전적 의미에서 파생된 신념 체계, 다시 신념 체계에서 파생한 삶을 이해하는 방식, 그 이해하는 방식이 보편적 분석의 범주와 결합된 것을 의미한다. 즉 '빅데이터 분석'이란 할리데이(Michael Halliday)의 언어 이론을 차용하자면, 지금까지 논의한 '빅데이터'의 일차적(congruent) 표현을 문법적으로 은유화(grammatical metaphor)한 것이다.

'행동 변수'란 관찰 가능한 '보편적' 인간의 직접적 행동과 '보편적' 인간의 의도, 선호 등 직접적으로 관찰 불가능한 주관적 요인을 간접적·반복적으로 관찰할 수 있는 신호들(signals)을 의미한다. 반복적으로 관찰 가능하기에 '행동 변수'는 사전적 의미에서 '빅데이터'화 할 수 있다. 즉 인간의 특정 행동 및 특정 선호를 나타내는 '행동 변수'는 '수리 언어화' 또는 '수치화'된 후 '빅데이터'란 이름의 정보량으로 기억되는 것이다. 더 나아가, 행동 변수와

행동 변수 간에 인과율 또는 인과관계를 확정할 수 있다면, 우리는 '행동 변수'를 통해 인간의 특정 행동이 '예측 가능하다'고 말할 수 있다.

여기까지 논의한 '빅데이터 분석'은 사실 보편적 과학 분석의 속성에서 크게 벗어나지 않는다. 보편적 과학 분석과 빅데이터 분석의 결정적인 차이점은 우리가 관찰한 '행동 변수'를 조종하는 '우리의 그림자'가 존재한다는 점이다. 다시 말해서, '행동 변수'는 '보편적' 인간의 어떤 선택의 결과로 파생된다.

굳이 비유하자면 '빅데이터 분석'은 피아니스트의 피아노 연주를 감상하고 비평하는 '음악 평론가의 평론(critique)'과 같다. 분석의 주체가 관찰하는 하나의 '행동 변수'는 무명의 피아니스트가 이름 모르는 작곡가의 악보에서 반복적으로 뽑아 오는 특정한 음높이에 위치해 있는 '도' 음을 듣는 것과 같다. '분명히' 피아니스트가 '미치광이가 아닌 사람'인 것은 알건만, '빅데이터 분석가'라는 명함을 가진 음악 평론가는 오직 특정한 높이의 '도' 음만을 관찰하는 것과 같다. 여러 높이의 '도' 음을 반복적으로 듣고, 또한 여러 높이의 '미', '솔' 음을 구별해서 피아니스트가 들려주는 '미지(未知)의 곡'과 이를 파악하는 음악 평론가의 작업을 '빅데이터 분석'이라 부를 수 있겠다. 관찰된 데이터 또는 수치화된 정보량에서는 '행동 변수'로 명명된 '도', '미', '솔', 또는 '도', '파', '라'만 기록될 뿐, 그 행동 변수 '도'를 어느 시점에서 선택한 피아니스트의 의도, 목적 등은 데이터에 기록되지 않는다. 그러나 데이터에서 기록되지 않는다고 해서, 피아니스트의 존재가 무시될 수는 없다. '분명히' 피아니스트가 '행동 변수'로 피아노 건반을 '감촉'함으로써 체계적으로 선택하고 있다는 것을 알기 때문이다. 특히 이 피아니스트가 들려주는 곡의 악보를 사전적(事前的, ex ante)으로, 또는 선험적(先驗的, a priori)으로 알지 못할 때, '행동 변수'를 조종하는 '우리의 그림자'로서의 피아니스트의 존재는 '빅데이터 분석'에서 절대 무시될 수 없다.

과학 분석에서의 '행동 변수'는 행위의 시작이자 끝이지, 그 행위를 조종하

는 '그림자'를 논하지 않는다. '원자의 속도'는 바로 관찰 행위의 시작이면서 마지막에 불과하다. 그 '원자의 속도'를 조종하는 것이 있다면, 그것은 또 다른 '행동 변수'이어야 하지, 행동 변수가 아닌 '그림자'일 수는 없다. 칸트의 인과율 형식에 관한 논쟁에서 '방사성 원자 라듐 B가 전자를 방출하고 라듐 C로 변하는 과정에는 어떠한 원인도 없다'는 하이젠베르크(Werner Karl Heisenberg)의 주장은 자연 과학 분석과 인간의 삶을 대상으로 삼는 '빅데이터 분석'의 본질적 차이를 오히려 극명히 드러낸다. 반면, '빅데이터'의 분석 대상이 인간의 삶이기에, '행동 변수' 수립 시 항상 그 '행동 변수'를 선택하는, 그러나 '행동 변수'는 아닌, 그렇기에 직접적으로 관찰 가능하지 않은, '우리의 그림자'이자 피아니스트를 염두에 두어야 한다.

실전 연습: OK 큐피드의 사례[4]

앞 장에서 '빅데이터 분석'의 의미와 '모형적 사고'에 기인한 '빅데이터 분석' 자체의 모호함까지 살펴보았다. 이제 '모호한 언어'를 대표하는 시어 '저만치'를 어떻게 '정확한 언어'로 대체 가능한가를 시현해 보고자 한다.

먼저 '빅데이터 분석'에서 가장 중요한 시발점은 관찰 가능한 '행동 변수'의 존재 여부를 파악하는 것이다. 첫째, '맥락'에 맞는 수치화 가능한 '행동 변수'를 설정할 수 있어야 하고, 그 '행동 변수'에 대응하는 반복 관찰 가능한 데이터가 존재해야 한다. 앞 장에서 논의했듯이, '행동 변수'란 반복적으로 관찰 가능한 '보편적' 인간의 직접적 행동과 '보편적' 인간의 의도, 선호 등 직접적으로 관찰 불가능한 주관적 요인을 간접적·반복적으로 관찰할 수 있는 신호들을 포함한다. 따라서 '행동 변수'는 '보편적' 인간의 의도, 선호,

4) 이 절의 분석 내용은 크리스티안 루더의 ≪빅데이터 인간을 해석하다≫(이가영 옮김, 다른, 2015)를 참조하였다. 특히 [그림 1], [그림 2], [그림 3]은 해당 문헌에서 발췌하였다.

동기에 의한 제반 행위를 포괄하기 때문에, 김소월의 시 <산유화>에서 꽃과 화자의 심적 거리 '저만치'를 '행동 변수'화하는 것이 결코 불가능하지 않다. 먼저 '저만치' 거리에 있는 대상 '꽃'은 '화자'에게 연모(戀慕)의 대상이다. '전통적'으로 연모의 대상은 이성 상대이기에, '저만치'에 피어 있는 '꽃'은 화자인 '내'가 연모의 감정을 느끼는 이성 상대로 분석할 수 있다. 연모의 감정이란 결국 수많은 이성 상대 가운데 하나를 선호하는 것이다. 여기서 우리는 '보편적 화자인 나'를 분석하는 것이지, 카사노바를 분석하는 것이 아니다. 따라서 '보편적 화자인 나'는 수많은 상대에서 오로지 하나의 '꽃'을 선호하고 선택할 수 있다는 행동 강령의 '제1원리(the first principle)'를 부여할 수 있을 것이다. 만일 '수많은 상대', '꽃', 화자인 '나'를 숫자로 표시할 수 있다면, 또한 그러한 숫자를 모을 수 있다면, 우리는 '빅데이터 분석'을 통해 성공적으로 '저만치'를 명확한 언어로 대체할 수 있을 것이다. 따라서 먼저 '수많은 이성 상대'를 수치화하는 것이 가능할지 여부를 따져야 할 것이다. 만일 반대편에 있는 선택 가능한 이성 상대를 숫자로 표현할 수 있다면, '내'가 선택하는 '꽃'도 선택 가능할 것이고, '꽃' 입장에서 보면, 화자인 '나'도 '수많은 이성 상대' 중 하나이므로 수치로 표현 가능할 것이다. 물론 '꽃' 또한 반대편 성(性)을 대표하는 카사노바인 경우는 배제한다. 다시 말하면, '모호한' 언어 '저만치'는 '명확한' 언어 '선호도'로 대체 가능할 것이다. 물론 '저만치'의 상대를 어떤 숫자로 대표해야 할지 결정하는 것은 쉬운 문제가 아닐뿐더러, 설사 대표하는 숫자를 찾았다 한들, 그 숫자를 기록한 데이터를 모으는 것도 쉬운 문제는 아니다. '양으로 승부하여 질로 승화'하려는 공대생의 재도전은 여전히 험난한 길인 것처럼 보인다.

그러나 미국의 데이트 사이트 'OK 큐피드' 운영자이자 '데이터 과학자' 크리스티안 루더(Christian Rudder)는 놀라운 발상으로 '데이터 과학자'가 '양으로 승부하여 질로 승화'하는 돌파구를 찾아냈다. 일단 매년 1,000만 명 이상이 방문하는 사이트를 운영하기 때문에, 적절한 행동 변수만 정해진다

면, 5년간 자료 5,500만 개를 생성하는 것이 가능했다. 또한 남녀 간의 데이트를 성사시키는 사이트를 운영하기 때문에 남녀 간 선호에 관한 정보를 비교적 쉽게 수집할 수 있었다. 크리스티안 루더가 고안해 낸 '행동 변수'는 아래와 같다.

1) 여성은 어떤 나이의 남성을 가장 매력적으로 생각하는가?
2) 남성은 어떤 나이의 여성을 가장 매력적으로 생각하는가?

[그림 1]은 OK 큐피드 자료를 활용하여 조사한 연령별 여성의 남성에 대한 선호 연령을 표로 나타낸 것이다. 예를 들어, 20세인 여성이 가장 매력적으로 느끼는 남성의 연령은 23세인 것을 의미한다. 또는 20세인 여성은 23세인 남성을 '선호'한다. 시어 '저만치'는 여성의 남성 '연령 선호도'로 대체된다. 표에서 보듯이, 21세인 여성은 23세의 남성을, 22세인 여성은 24세의 남성을, 23세의 여성은 25세의 남성을 가장 선호하는 것을 알 수 있다. 시어 '저만치'의 거리는 그냥 지나치기는 아쉬운 거리, 그러나 내 손길을 뻗친다 한들, 인연의 끈을 잡을 수 없는 거리, 아련히 다가오는, 그러나 이내 시나브로 멀어져 가는 '복학생 오빠'와의 거리, 그래서 가슴 설레는, 한(恨)이 이슬처럼 망울지는, '마음'의 '모호함'을 드러내는 거리이지만, 이 연령 선호도는 왜 '복학생 오빠'와의 거리가 그토록 한이 이슬처럼 망울지는 거리인지, 숫자로 정확히 보여준다. '정확한' 언어로 말하자면, 20세 초반의 여성은 2~3살 많은 남성을 '체계적'으로 선호하는 경향이 있기에, '복학생 오빠'가 '저만치' 거리에 자리를 차지할 가능성이 매우 높다.

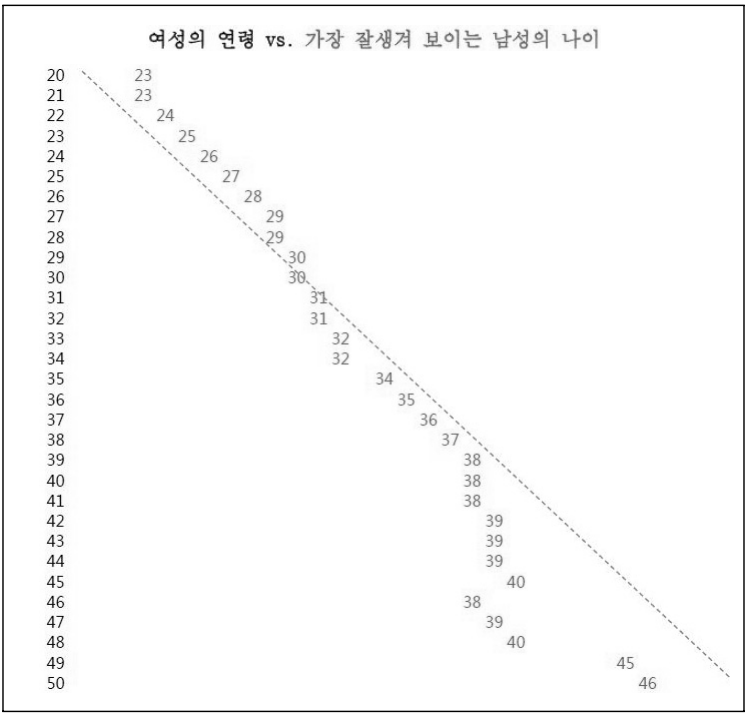
그림 1

여성의 연령	vs. 가장 잘생겨 보이는 남성의 나이
20	23
21	23
22	24
23	25
24	25
25	26
26	27
27	28
28	29
29	29
30	30
31	31
32	31
33	32
34	32
35	34
36	35
37	36
38	37
39	38
40	38
41	38
42	39
43	39
44	39
45	40
46	38
47	39
48	40
49	45
50	46

또한 [그림 1]은 여성의 나이가 증가함에 따라, 선호하는 남성도 비교적 '비례적'으로 올라간다는 것을 알 수 있다. 예를 들어 29세의 여성은 29세 남성을, 30세 여성은 30세 남성을, 31세 여성은 31세 남성을 선호함을 알 수 있다. 그러나 여성의 나이가 32세를 넘어감에 따라, 연하의 남성을 선호하는 경향이 강해지고 있음을 알 수 있다. 놀랍게도, 작금의 '송중기' 열풍 사태를 이 도표는 '정확한' 숫자로 보여 준다. 지금까지의 논의를, 크리스티안 루더가 명명한 '나이 비교선'을 가지고 재구성할 수 있다. [그림 2]는

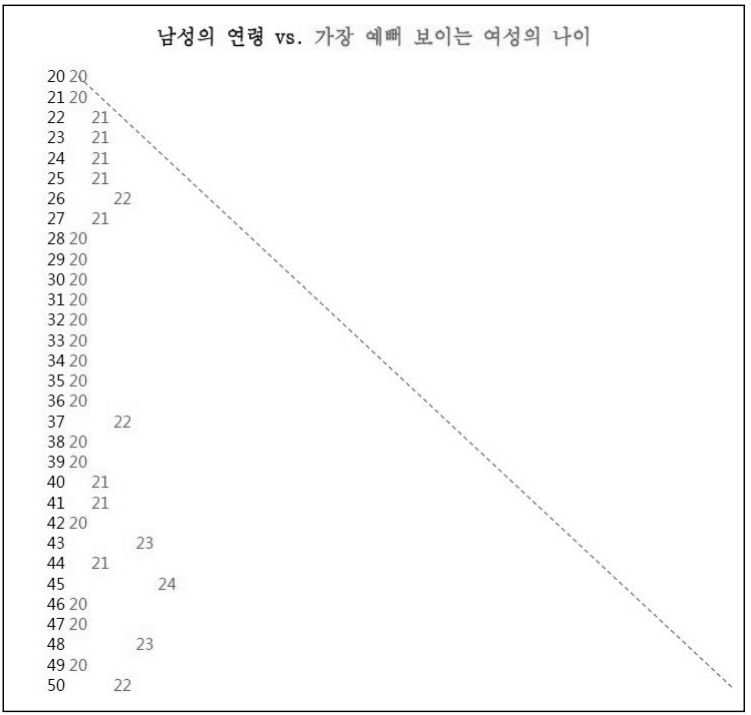
[그림 1]을 여성의 나이와 남성의 나이가 같은 '나이 비교선'으로 재구성한 것이다.

그림 2



대각선으로 그려진 '나이 비교선'을 보면, '복학생 오빠'와의 거리, '송중기' 열풍 사태 등을 보다 명확히 파악할 수 있다. 이제 남성의 연령 선호도를 알아보고자 한다. [그림 3] 역시 OK 큐피드 자료를 활용해서 작성되었으며, 대각선의 점선은 '나이 비교선'을 나타낸 것이다.

그림 3



남성의 경우, '나이 비교선'은 의미가 없다. 남성은 오로지 20대 초반의 여성만이 '저만치' 거리에 피어 있는 '꽃'이기 때문이다. 저만치의 '꽃'이 20대 초반으로 고정되어 있다면, 과연 그 거리를 내 손길이 뻗친다 한들, 인연의 끈을 잡을 수 없는, 아련히 다가오는, 한이 이슬처럼 망울지는 모호하고, 미묘한 거리라고 말할 수 있을까? 남성에게 애당초 같이 늙어가는 원숙미에 대한 존경은 있지도 않다. 남성의 '꽃'에 대한 반응은 지극히 생화학적 반응처럼 보인다. 20세와 21세의 남성은 '저만치' 한이 이슬처럼 망울지는 '시감(詩感)'을 만끽할 수도 있을 것이다. '빅데이터 분석'에 의하면, 20세, 21세의 남성이 한이 맺힐 정도로 아끼는 20세의 여인은 오직 23세 '복학생 오빠' 쪽으로

시선이 고정되어 있다는 것을 '빅데이터 분석'이 보여주기 때문이다.

크리스티안 루더가 지적했듯이, 이런 연구 결과는 '진실의 이면에 가려져 있는 우리의 허영심과 취약점'을 공공연히 드러낸다. '모호한 천공'을 수많은 숫자로 조합하고 빚어낸, 인공족 아틀라스로 지탱했을 때, 우리는 까발려진다. 이런 의미에서 '명확한 언어'는 '폭력적'이다.

'빅데이터 분석'의 명과 암: 구글 독감 예측 프로그램

OK 큐피드 사례에서 보여주듯이, 인간의 모호한 삶은 '모호한' 시어로만 이해되는 신비로운 천공이 아니라, 단순히 얽혀 있어 복잡한, 그래서 우리가 풀어내 까발려야 하는 '실타래'일 뿐이다. '양으로 승부하여 질로 승화한다'는 공대생의 무모한 재도전은 성공적인 것처럼 보인다. 인공족 아틀라스의 사촌뻘인 구글 신은 여기서 더 나아가, '빅데이터 기법'을 활용하여 전 세계의 독감 환자 수를 예측할 수 있(다고 주장하)는 '빅데이터 분석'을 예시하였다.[5]

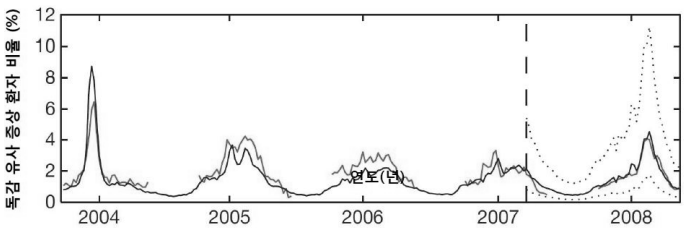
OK 큐피드 사례에서와 같이, 독감 환자 수를 예측하기 위해서는 독감 환자를 식별하는 '행동 변수'를 찾아내는 것이 중요하다. 특히 '우리의 그림자'가 조종하는 '행동 변수'의 모습을 보여주는 것이 중요하다. 구글이 고안해 낸 '행동 변수'는 아래와 같다.

> "사람들은 독감 증상이 있는 경우, '기침', '고열', '해열제' 등과 같은 독감 증상 관련 용어를 검색 엔진을 통해 검색할 것이다."

5) 구글의 독감 예방 프로그램에 보다 관심 있는 독자는 정하웅·김동섭·이해웅 저의 ≪구글 신은 모든 것을 알고 있다≫(사이언스북스, 2014)를 참조하기 바란다.

그림 4

◆ 구글 독감 트렌드 : 검색에 근거한 구글의 예측과 실제 독감 환자 데이터 비교 (출처: Ginsberg et al. "Detecting influenza epidemics using search engine query data," Nature, 457, 2009)

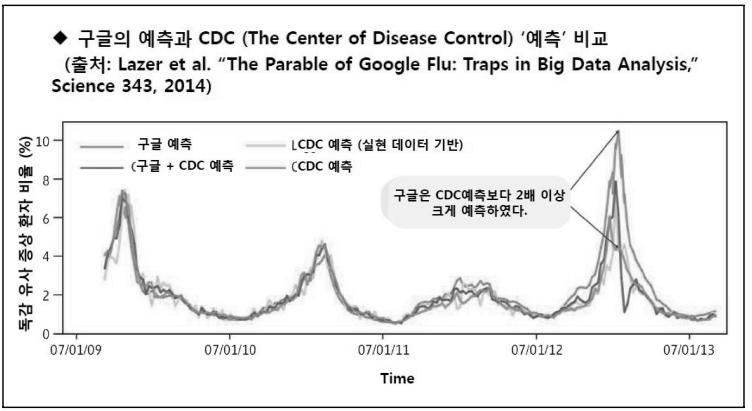


구글 자체가 검색 엔진을 운용하기 때문에, 어느 시점에, 어느 지역에서 독감 증상 관련 검색어가 '증폭'하는지 구글 서버를 통해 파악할 수 있고, 또한 데이터로 기록할 수 있다. 실제 구글은 독감 환자 수와 가장 연관성이 높은 검색어 50개를 선정하고, 2003년에서 2007년 사이의 데이터를 수집·분석한 후, 2008년 환자 예측치를 발표하였다. [그림 4]는 구글의 2008년 예측치를 실제 독감 환자 집계 수와 비교한 것이다. [그림 4]에서 보듯이, 구글 예방 프로그램은 거의 완벽하게 독감 환자 수를 예측했다! '모호한 천공'을 수많은 숫자로 조합하고 빚어낸, 인공족 구글로 지탱했을 때, '모호한 천공'은 더 이상 그렇게 무거워 보이지 않는다. 구글 신을 숭배하기만 한다면, '저만치' 거리에 있는 20세 여인에게 가슴 설렜던, 그러나 그 감정을 시어 '저만치'로 세련되게 표현할 재능이 없는 공대생조차 '모호한 천공'을 지탱할 수 있음을 보여주는 사례인 셈이다.

이제 구글은 거인족 아틀라스를 대신해, '모호한 천공'을 짊어질 수 있어 보인다. 구글은 '모호한 천공' 정도는 한 손가락으로도 지탱할 수 있는 '폭력성'을 자랑이라도 하는 듯, 2012년 독감 환자 수 예측치를 발표하였고, 이 예측치는 실제 환자수와 일치할 것이라 장담했다. [그림 5]는 2012년 구글의 예측치와

미국질병통제예방센터가 실제 보고한 독감 환자 수(CDC '예측'이라 명명된 것)를 비교한 것이다.

그림 5



구글 신의 장담과는 달리, 구글 독감 환자 수 예측 프로그램은 '정확히' 실제 환자수보다 2배이상을 과대 예측하였다! 삶의 '모호함'이라는 무게에 짓눌려 압사해 버린, 피에 버무린 살점이 꽃잎처럼 산화한 사내의 이름은 바로 '구글 신'이었다. '모호한 천공'은 여전히 버겁고, 그 버거움을 짊어져야 하는 사내는 거인족 아틀라스이지, 인공족 '구글 신'이 아님을 보여준다. 그렇다면 이러한 구글 예측 실패는 구글이 고안한 '빅데이터 분석'의 실패인가? 아니면 일반적 의미의 '빅데이터 분석'의 한계인가? 다음 장에서 이 질문에 대한 답을 찾고자 한다.

'모형적(模型的) 사고'로서의 '빅데이터 분석'

보편적 과학 분석과 달리, '우리의 그림자'가 '빅데이터 분석' 과정에 투영되어야 하는 이유는, 관찰의 대상이 분석의 주체인 '보편적' 인간, 바로 우리 자신이기 때문이다. 분석의 주체인 '보편적' 인간이 바로 '보편적' 인간을 관찰·분석 대상으로 삼기에, '빅데이터 분석'에서는 항상 분석의 대상과 분석의 주체 간의 거리, 즉 관찰자인 '보편적' 인간과 관찰 대상인 '보편적' 인간 사이의 거리는 시어 '저만치' 정도의 모호한 거리이다. 대부분의 '빅데이터 분석'에 관한 저작(著作)들을 보면, 관찰자인 '보편적' 인간과 관찰 대상인 '보편적' 인간 사이의 거리에 '저만치'만큼의 모호한 거리가 있다는 것을 인식하지 못하거나, 설사 인식한다 하더라도 단순히 '우려'하는 수준에 머물고 있다. 관찰자인 '보편적' 인간과 관찰 대상인 '보편적' 인간 사이에 '저만치' 놓여 있는 모호한 지대(地帶)를 확산시키는 중심축은 소위 '빅데이터 분석가'의 '모형적(模型的) 사고'에 대한 진지한 고찰의 결여에 있다고 진단할 수 있다. 앞의 절에서 본 구글 독감 예측 프로그램의 과대 예측은 바로 구글의 '빅데이터 분석가'가 세심하게 '모형적 사고'를 고찰하지 못한 점에 있다고 말할 수 있다.

'모형적 사고'를 이해하기 위해서는 먼저 '모형'이 무엇인지 이해할 필요가 있다. '모형'이란, 세계의 관측자로서의 '문제의식'을 갖고 있는 관찰자가, 세계의 속성 중 필수 불가결한 요소만 취사선택하는 '추상화'의 과정을 통하여 '실존하는' 세계를 '수리적 언어'로 축소화, 재구성한 작은 세계(microcosm)를 의미한다. '모형적 사고'란 이 축소화되고 재구성된 작은 세계를 가지고 실제 세계를 분석하는 것을 의미한다. 또한 '만일 내가 이 축소화된 세계에 살고 있다면, 나는 어떠한 행동을 할까? 또는 나는 어떠한 행운과 불행을 맞이할 수 있을까?' 머릿속으로 상상하는 것을 '사고 실험(Das Gedankenexperiment)'이라고 하는데, '모형적 사고'는 이 '사고 실험'의 활동도 내포한다.

이 '모형적 사고'의 근저에는 세계의 관측자로서의 인간이 이 '복잡하고 모호한' 세계를 '즉시적'으로 이해하는 것은 불가하지만, 순차적·점진적으로 이해하는 것은 가능할 것이라는 믿음을 암묵적으로 전제하고 있다. 20세기를 대표하는 독일의 수학자 힐버트(David Hilbert)의 묘비명에 적혀 있는 "우리는 알아야 한다. 그리고 우리는 알 것이다.(Wir müssen wissen, Wir werden wissen.)"라는 어구는 바로 '모형적 사고'의 기본적, 암묵적 전제를 함축적으로 표현한 것이라 볼 수 있다. 따라서 '모형적 사고'란 우리의 삶이 '회색'인 것은 알지만, '복잡하고 모호한' 색깔이기에 상대적으로 '분명한' 색깔인 흑색과 백색으로 '일단' 우리의 삶을 이해하는 것을 의미한다. 모호한 회색의 삶을 이해하기 위해, 필수 불가결한 요소인 흑색을 취사선택하는데, 이렇게 흑색을 취사선택하는 과정을 '변수화(變數化)'라고 부른다. 즉 취사선택된 필수 불가결한 요소는 변수(變數, variable)로 표현하는 것이다. 반면에 선택받지 못한, 버려지는 삶의 나머지 요소는 상수화(常數化) 처리를 하는데, 상수(常數, constants)라는 말 자체가 의미하듯이, 취사선택되지 못한 나머지 요소는 관찰자의 관찰 유무와는 독립적으로 항상 변하지 않는 것, 일정한 속성을 유지하는 것으로 간주한다. 변수는 크게 '행동 변수'와 '매개 변수(媒介變數, parameter)'로 나누어질 수 있는데, 행동 변수란 관찰자의 관찰 기간 동안, 계속 관찰되어야 하는 대상이고, 매개 변수란 관찰자의 명백한 관찰 대상이지만, 관찰 기간 동안 일정한 속성을 유지하는 것, 즉 관찰 기간 동안은 변하지 않는 '안정성'을 갖는 변수로 이해된다. 모든 과학적 모형은 행동 변수, 매개 변수, 상수 세 가지로 이루어져 있다. '모형적 사고'란 결국 행동 변수, 매개 변수, 상수, 이 세 가지 성분으로 세상을 이해하는 것이다. '빅데이터 분석'은 우리의 삶을 대상으로 삼기 때문에, '빅데이터 분석'에서의 '모형'은 한 가지 성분을 더 추가하게 되는데, 앞에서 논의했다시피, '우리의 그림자'인 가공의 '인격체'가 부여되어야 한다. 이 가공의 '인격체'는 바로 모형이 펼쳐내는 축소화된 세계에서, 행동 변수와 매개 변수를 조종하는 우리의 '아바타'라 볼 수 있다.

이 인격체인 '아바타'의 존재가 순수 과학에서의 모형과 인간의 삶을 다루는, '빅데이터 분석'을 포함한 사회 과학에서의 모형과의 차이를 만들어 낸다. 요약하면 '빅데이터 분석'에서의 '모형적 사고'란 상수, 매개 변수, 행동 변수, 그리고 변수들을 조종하는 가공의 '인격체' 네 가지 성분으로 세상을 이해하는 것이다. 여기서 반드시 지적되어야 할 것은, '빅데이터 분석'에서 다루는 모형에서 우리의 '아바타', 즉 가공의 '인격체'는 사전적(事前的, ex ante)으로, 또는 선험적(先驗的, a priori)으로 존재하고 행동 변수를 결정하지만, 사후적(事後的, ex post)으로, 또는 경험적(經驗的, a posteriori)으로는 실존하지 않는다는 점이다. 다시 말하면, 우리가 데이터에서 관찰하는 것은 행동 변수이지, '인격체'가 아니다. 우리의 '아바타'는 사후적으로 관찰된 행동 변수를 통해 귀납적으로 추리(推理)한다. 동시에 연역적으로 행동 변수를 조종하는 우리 '인격체'의 의지, 목적, 동기 등을 추론해 '인격체'의 행동 강령 '제1원칙(the first principle)'을 수립한 후, '선험적'으로 우리 '인격체'의 존재를 (논리적으로) 증험한다. 귀납적으로 추리하고, 동시에 연역적으로 논증할 때, 우리의 '아바타', 우리의 '그림자', 우리의 '일그러지고, 이름 없는 피아니스트'의 실체를 어느 정도 가늠할 수 있을 뿐이다. 그러나 이렇게 파악한 '인격체'의 실존조차도 철저히 '맥락적(contextual)'이다. 다시 말하면, 가공의 '인격체'는 세상 이치의 맥락에 따라, 다양한 몸짓을 보여 줄 뿐이지, 전체의 모습을 절대 보여주지 않는다. 이 가공의 '인격체'가 맥락적일 수밖에 없는 이유는 바로 세상의 회색을 재현하는 것이 아니라, 세상의 흑색만을 재현하는 '모형적 사고' 안에서 또는 모형의 '사고 실험'에서만 존재할 수밖에 없기 때문이다. 모형의 용어로 좀 더 '정확히' 표현한다면, 각 모형은 일정하게 변하지 않는 상수항(常數項)을 가지고 있고 또한 관찰 기간 동안은 변하지 않는 '안정성'을 갖는 매개 변수를 포함하고 있다. 이러한 불변적인 성질의 '안정성'은 우리가 관심 있는 필수 불가결하다고 믿는 요소(행동 변수)를 어떻게 잡을 것인가, 다시 말하면 필수 불가결한 요소를 흑색으로 볼 것인가, 또는 백색으로 볼 것인가, 또한 관찰

기간을 어떻게 잡을 것인가에 따라 정해지기 때문에, 모형 세계는 맥락적일 수밖에 없고, 그 모형 세계에 살고 있는 우리의 '아바타' 또한 맥락적인 존재이다.

서두에서 지적했듯이, '빅데이터 분석'에서 부지불식간에 드러나는 관찰자인 '보편적' 인간과 관찰 대상인 '보편적' 인간 사이의 '모호한' 거리는 바로 모형 세계에 사전적으로, 또는 선험적으로 존재할 수밖에 없는 '아바타'의 불완전한 실존성, 그리고 맥락적인 관념성에 기인하는 것이라 하겠다. 그리고 이 '아바타'의 불완전하고 맥락적인 몸짓은 오로지 '모형적 사고'의 진지한 고찰에서만이 파악될 수 있다 하겠다.

이제 다시 구글 독감 예측 프로그램 사례로 돌아가자. 구글이 고안한 '행동 변수'는 "사람들은 독감 증상이 있는 경우, '기침', '고열', '해열제' 등과 같은 독감 증상 관련 용어를 검색 엔진을 통해 검색할 것이다."라는 것이었다. 그런 '모형적 사고'의 관점에서 보면, 이 행동 변수는 좀 더 세심하게 다루어질 수 있다. 다시 말하면, '모형적 사고'를 반영하여 이 '행동 변수'를 세심하게 재구성하면 아래와 같다.

> "우리의 '아바타'는 독감 증상이 있는 경우, '기침', '고열', '해열제' 등과 같은 독감 증상 관련 용어 50개를 검색 엔진을 통해 '즉시' 검색할 것이다."

또한 이 행동 변수와 더불어 구글은 '암묵적'으로 아래의 '매개 변수' 또는 '상수항'을 고안했다.

> "우리의 '아바타'는 독감 증상이 없다면, 독감 증상 관련 용어 50개를 결코 검색하지 않을 것이다."

다시 말하면, 구글의 예측 프로그램은 적어도 관찰 기간 동안 "우리의 '아바타'는 독감 증상이 없다면, 독감 증상 관련 용어 50개를 결코 검색하지 않을 것이다."라는 '아바타'의 행동 강령이 준수된다는 '매개 변수'를 부지불식

간에 고안했다. 실제 2008년에 이 '매개 변수'는 매우 '안정적'이었다. 다시 말하면, 우리의 '아바타'가 그러했듯이 실제 독감 증상이 있는 사람들만 검색을 했기에, 구글의 예측 프로그램은 정확했다. 그러나 2012년은 전혀 다른 양상이 펼쳐진다. "우리의 '아바타'는 독감 증상이 없다면, 독감 증상 관련 용어 50개를 결코 검색하지 않을 것이다."라는 이 '매개 변수'가 전혀 '안정적'이지 않았다. 다시 말하면, 2012년에는 독감 증상이 없는 사람들도 독감 관련 단어를 검색했다! 그렇다면 2012년에는 왜 독감 증상이 없는 사람도 독감 관련 단어를 검색했을까? 그것은 바로 뉴스, 신문과 같은 언론 매체가 구글의 독감 예측 프로그램에서 내놓은 예측치를 보도하기 시작했고, 이 예측치가 보도됨에 따라, 독감에 걸리지 않은 사람들도 다가올 독감을 미리 예방하기 위해 독감 관련 검색어를 검색하기 시작했기 때문이다. 모형으로서의 구글 독감 예방 프로그램은 바로 "우리의 '아바타'는 독감 증상이 없다면, 독감 증상 관련 용어 50개를 결코 검색하지 않을 것이다."라는 매개 변수를 잘못 다루었다고 말할 수 있다. 결론적으로 구글 독감 예측 프로그램은 명백히 '모형'이다. 그러나 이 독감 예측 프로그램의 설계자는 이 프로그램이 '모형적 세계'를 재현하는, 세상의 축소판이라는 것을 간과한 듯하다. 또한 '모형'은 '맥락적'이라는 것도 간과한 듯하다. 2008년의 매개 변수와 2012년의 매개 변수가 다를 수 있다는 것을, 그래서 '모형'은 맥락적이라는 것을 간과한 것이다. 마지막으로 구글 예측 프로그램은 "프로그램 설계자(관찰자) = 모형 세계의 '아바타' = 실존하는 사람들"이라는 중대한 과오를 범한 듯이 보인다. 이 절의 서두에서 밝혔듯이 관찰자인 '보편적' 인간과 관찰 대상인 '보편적' 인간 사이에는 시어 '저만치' 정도의 모호한 거리가 태생적으로 존재한다. 그리고 그 모호한 거리는 바로 '모형적 사고'의 이면일 뿐이다.

결어

한강의 연작 소설 ≪채식주의자≫는 한 편의 음악과도 같은 소설이다. 주인공이 '비폭력성'의 상징인 '나무'로 변용되어 가건만, 놀랍게도 그 '비폭력성'을 성취하도록 돕는 것은 다면 형태의 '폭력성'이다. '폭력성'은 우리의 살점을 먹어 대는, 천진한 어린 악마가 우리 귀에 속삭이는 방식으로, 또는 바로크 음악의 푸가처럼 '폭력'이라는 일관성 있는 주제를, 여러 성부에 있는 화자의 입을 통해 노래하는 방식으로 전달된다. '폭력성'은 때로는 무지로, 때로는 무관심으로, 때로는 '꽃'과의 '모호한' 거리 '저만치'를 '명백하게' 파괴하는 피에 버무린 살점을 꽃잎처럼 산화하는 사내의 모습으로 등장한다. 주인공 또한 '폭력성'의 주제를 또 다른 성부에서, 자기 학대, 소통 부재의 '폭력성'의 마법에 걸린 것처럼 노래한다. 그러나 이 다양한 성부에 기거하는 '폭력성'은 주인공을 '십자가' 형태의 나무가 있는 골고다 언덕으로, 마치 예정된 길을 인도하듯이 인도한다. 그럼에도 '진흙 속의 연꽃'이라는 '상투적인' 노래, 즉 '명백히' 한 가지 으뜸음을 가진, 조성(調聲, tonality)을 가진 음악을 허락하지 않는다. 오히려 1721년에 출간된 바흐(Johann Sebastian Bach)의 ≪평균율≫ 1권 중 마지막 푸가에서, 인류 역사상 최초의 12음 기법으로 작성된, 무조(無調, atonality) 음악의 가능성을 목도한 쇤베르크(Arnold Schönberg)의 환희를, ≪채식주의자≫는 재현한다. ≪채식주의자≫는 독자로 하여금 으뜸음을 허용하지 않는, 무조성(無調聲, atonality)의 소리를 발견하는 '쇤베르크의 그 환희'를 만끽하게 해 준다.

'빅데이터 분석' 또한 우리의 삶을 한 편의 음악처럼 들려준다. '빅데이터'가 들려주는 음악은 그냥 지나치기는 아쉬운 거리, 그러나 내 손길을 뻗친다 한들, 인연의 끈을 잡을 수 없는 거리, 아련히 다가오는, 그러나 이내 시나브로 멀어져가는 '복학생 오빠'와의 거리, 그래서 가슴 설레는, 한(恨)이 이슬처럼 망울지는, '마음'의 '모호함'을 드러내는 '저만치'를 으뜸음으로 하는 음악이

아니다. 소설 ≪채식주의자≫처럼, '빅데이터'를 움직이는 동인(動因)은 다양한 형태의 '행동 변수'가 보여 주는 '폭력성'이다. '명확히' 대칭적이고 균등한 가치를 가진 12개의 '폭력적인' 음은 '저만치'라는 으뜸음을 허락하지 않는다. 12개의 음은 으뜸음 '저만치'를 분석(分析)의 이름으로 갈기갈기 찢어버리고, 까발린다. 바로 쇤베르크의 12음 기법에 의한 무조 음악의 이상을 삶이라는 악보에 기록하는 것, 이것이 '빅데이터'의 기본 성향이다. 20세기 최고의 바흐 건반 음악 연주자인 글렌 굴드(Glenn Gould)의 연주에서도 이러한 '빅데이터'의 '시체 해부자' 성향을 볼 수 있다. 글렌 굴드의 주법은 고적한 산사의 목탁 소리와도 같은 '명확한' 언어를 전달한다. 굴드는 특히 그냥 지나치기는 아쉬운 거리, 그러나 내 손길을 뻗친다 한들, 인연의 끈을 잡을 수 없는 거리, 아련히 다가오는, 그러나 이내 시나브로 멀어져가는 '복학생 오빠'와의 거리, 그래서 가슴 설레는, 한(恨)이 이슬처럼 망울지는, '마음'의 '모호함'을 드러내는 '저만치'를 으뜸음으로 하는 낭만적인 피아노곡의 얼굴을 목탁 소리와 같은 점묘법을 통해 가차 없이 지워 버린다. 곡의 '민낯'을 여과 없이 까발린다. 남겨지는 것은 회색빛 불교의 승려복 그리고 머리카락 하나 없는 두상뿐이다. 그럼에도 굴드는 멈추지 않는다. 굴드는 '명확한' 목탁 소리로 언제나 해체하고 싶은 시체 해부자의 본능에 시달렸다. 실제 '바흐' 연주자로 유명하지만, 12음 기법으로 작성된 쇤베르크의 곡 또한 너무나 잘 연주하였다. 그러나 끊임없는 해부의 본능으로 탄생한 쇤베르크의 곡을 듣고 있노라면, '모호함'을 제거한 후에 샘솟는 '명확한' 환희가 아니라, 해체하고 나니, 아무것도 없다는 '공허함'만이 몰려온다. 오직 시체 해부자의 본능에 시달리는 '빅데이터 분석'은 '모형적 사고'를 진지하게 고찰해 본 적이 없는 서툰 '빅데이터 분석가'의 모습이지, '빅데이터'의 진면모와는 거리가 멀다. '분명히' '모형적 사고'는 또 다른 차원의 '저만치'라는 '모호한' 거리를 만들어 낸다. ≪채식주의자≫는 폭력적이지만, 삶의 색채를 지워 버리지는 않는다. 진짜 '빅데이터 분석'은 폭력적이지만, 우리 삶의 색채를 더욱 다채롭게 만들어 줄 것이다. 삶에서 무조성의 소리를

발견하는 쇤베르크의 환희는 우리의 모호한 삶을 더욱 모호하면서도 분명하게 이해하게 해 줄 것이다.

마지막으로 '모형적 사고'를 가장 잘 이해했던 사람은 시인 무리에 있다. 하나의 몸짓 외에는 명확히 서술할 수 없는 '모호한' 삶에 '명확'하고 '폭력적인' 언어의 이름을 붙임으로써 '저만치' 거리에 있는 존재를 '꽃'으로 확정해 버리는 것, 바로 회색에서 필수 불가결하다고 믿는 것에 흑색이라는 명확한 모형을 설정해 버리는 행위를 시인은 아래와 같이 읊조린다.[6]

내가 그의 이름을 불러주기 전에는
그는 다만
하나의 몸짓에 지나지 않았다.
내가 그의 이름을 불러주었을 때
그는 나에게로 와서
꽃이 되었다.

6) 김춘수의 시 <꽃> 중에서.

언어 자료로 세상 보기
— 산업 분야의 언어 처리와 세종 말뭉치 운용

전채남
더아이엠씨

1. 빅데이터의 시대, 쌓이는 언어 자료

빅데이터의 시대는 소셜 미디어의 일상화로부터 시작되었다. 몇 년 사이에 카카오스토리, 페이스북, 트위터, 인스타그램, 유튜브 등 다양한 소셜 네트워크 서비스(SNS)가 등장하고 이용자들이 급증하면서 엄청난 양의 데이터들이, 또 다양한 형태의 데이터들이 실시간으로 생산되고 있다. 소셜 네트워크 서비스는 우리가 살아가는 세상의 모든 흔적을 데이터로 남기고 있기 때문이다. 이와 함께 정보 통신 기술(ICT)의 발달로 우리가 이용할 수 있는 데이터들은 엄청나게 늘어나고 있다.

소셜 미디어 시대가 되면서 수십 억 페이지에 이를 만큼 늘어난 반면 띄어쓰기 오류, 구어체, 미등록어, 오용어 등으로 인해 문서의 질은 낮아졌다. 비공식적, 비격식적 문서의 양이 절대적으로 늘어나면서 불완전한 문장 또는 문법에 어긋나는 표현들도 함께 늘어나게 되었다. 이러다 보니 사용자들은 정보 과부하에 의한 피로감을 해소하기 위해 필요한 것만 요약하길 원한다.

실제로 최근 2년 사이에 세계는 이전 인류 역사의 전체 기간보다 더 많은 데이터를 생산하였다. 그리고 2020년경이 되면 1초당 약 1.7메가바이트의 새로운 데이터가 생산될 것으로 예상된다. 이런 데이터는 페이스북, 카카오

톡, 메일 등을 통해 보내지는 메시지와 메일들로부터 생산되고 있을 뿐만 아니라 디지털 사진들과 점차 증가하는 비디오 데이터로부터도 생산되고 있다.

2020년경에는 전 세계에서 60억 개 이상의 데이터를 온전히 수집하는 스마트폰이 사용될 것으로 예상된다. 전화기가 스마트해지고 있을 뿐만 아니라 스마트 텔레비전, 스마트 워치, 스마트 미터, 스마트 홈, 스마트 테니스 라켓, 스마트 전구 등 우리 주변의 대부분 기기가 스마트해지고 있다. 2020년경에는 500억 개 이상의 인터넷 연결 기기(IoT)가 운영될 것이다. 이것은 엄청난 양의 다양한 데이터(텍스트와 비디오 데이터로부터 센스 데이터까지)가 상상할 수 없는 수준까지 증가할 것이라는 의미한다.

세상이 스마트해지는 만큼 빅데이터가 갈수록 중요해지고 있다. 각종 언론에서 연일 빅데이터와 관련된 보도를 하고 사람들은 빅데이터를 대명사처럼 사용하고 있다. 많은 정부 기관과 기업은 업무를 효율적으로 수행하기 위해서, 성과를 개선하기 위해서, 가치를 창출하기 위해서 빅데이터를 점점 더 활용하는 추세이다.

빅데이터는 몇 년 전까지는 불가능하였지만 현재는 가능해진 기술, 즉 데이터를 수집하고 분석할 수 있는 기술의 발달과 관련이 있다. 새로운 기술로 인해 향상된 능력을 가질 수 있게 되어 더 많은 데이터를 수집하고 저장, 분석할 수 있어 빅데이터 이용이 가능하게 되었다. 위키피디아(Wikipedia)는 “빅데이터를 기존 데이터베이스 관리 도구로 데이터를 수집, 저장, 관리, 분석할 수 있는 역량을 넘어서는 대량의 정형 또는 비정형 데이터 집합”이라고 정의하고 있다. 그리고 여기에는 이런 데이터로부터 가치를 추출하고 결과를 분석하는 기술까지 포함하고 있다. 쇤버거와 쿠키어(2013)는 “큰 규모를 활용해 더 작은 규모에서는 불가능했던 새로운 통찰이나 새로운 형태의 가치를 추출해 내는 일”로 빅데이터를 정의하고 있다. 그들은 이유를 아는 것보다 결과를 아는 것이 중요하기에 인과성(causality)에서 상관성

(correlation)으로 분석의 초점을 옮겨야 한다고 주장한다. 아이디시(IDC, 2011)는 대규모의 다양한 데이터로부터 수집, 검색, 분석을 신속하게 처리하여 경제적인 가치를 발굴하도록 설계된 차세대 기술 및 아키텍처로 4브이(Volume, Variety, Velocity, Value)를 특성으로 제시하고 있다. 비정형 데이터의 활용에 주목하여 '생각을 만드는 기술'이라고 빅데이터를 정의하기도 하는데, 이는 사람들의 자연스러운 디지털 대화를 수집하여 그 속에 내포된 인식, 이해, 의견, 반응 등을 읽어 내는 기술이라는 의미를 가지고 있다(김정선, 2015).

지금까지 연구자들의 빅데이터 정의가 다소 차이는 있지만 전반적으로 데이터 수집, 저장, 정제, 분석 등과 관련된 기술뿐만 아니라 이를 활용하는 해석 능력, 이를 통해 가치를 창출 할 수 있는 통찰력을 포함하고 있다.

빅데이터에 대한 장점을 쇤버거와 쿠키어(2013)는 다음의 세 가지로 정리하였다. 첫째, 빅데이터로 인해 훨씬 더 많은 데이터를 분석할 수 있고 어떤 때는 특정 현상과 관련된 모든 데이터를 분석할 수 있다. 둘째, 빅데이터와 같은 방대한 데이터를 들여다볼 때는 정밀성에 대한 욕구가 다소 느슨해져서 샘플링 오류가 줄어들어 측정 오류에 대해서는 좀 더 관대해질 수 있다. 셋째, 거대 규모의 데이터를 취함으로써 인과관계 추구라는 오래된 습관에서 멀어지는 대신 패턴이나 상관성을 찾아내어 새로운 이해와 귀중한 통찰을 얻을 수 있다.

2. 텍스트 마이닝(Text Mining), 언어 자료 처리

빅데이터가 있다고 해도 우리가 잘 활용하여 데이터를 통한 통찰력을 발휘할 수 없다면 빅데이터는 거의 가치가 없다. 데이터를 잘 활용하기 위해서는, 즉 통찰력으로 유용한 결과를 도출하기 위해서는 데이터를 분석할

수 있도록 데이터의 정제와 처리가 필요하다. 그런데 일반적으로 텍스트 데이터는 비정형 데이터로서 복잡한 구조를 갖기 때문에, 이를 정제하고 처리하는 일은 쉽지 않다.

데이터 정제의 방법에는 수동적인 방법과 자동적인 방법이 있다. 수동적인 방법은 데이터를 수집한 이후에 연구자가 워드나 엑셀 프로그램을 사용하여 일일이 특수문자, 조사, 띄어쓰기 등을 확인한 후 분석 가능한 형태로 수정하고 정리하는 것이다. 때때로 텍스트 마이닝의 과정에서 수집된 단어를 정제하는 단계에서 컴퓨터 프로그래밍을 통해 자동으로 실시하기도 한다.

텍스트 데이터를 자동으로 정제하는 방법은 언어 정보 처리의 한 분야이다. 언어 정보 처리는 컴퓨터와 인간 사이의 언어 소통을 강화하는 일로, 컴퓨터를 비롯한 기기가 인간의 언어를 잘 알아듣도록 하는 일과 인터넷이나 문서로 축적되어 있는 언어 자료를 인간이 잘 이용할 수 있도록 하는 일이다. 정보 통신 기술의 발달로 사람들의 스마트 기기 사용이 늘어나면서 점점 언어 정보 처리가 중요해지고 있다.

언어 정보 처리는 컴퓨터가 사람의 일상 언어를 이해하고 생성할 수 있도록 함으로써, 컴퓨터를 인간의 지적 활동의 보조자 및 지원 도구로 활용하도록 한다. 다시 말해 언어 정보 처리의 한 목적은 컴퓨터가 인간의 언어를 자동 번역, 요약, 인식하도록 하는 것이다.

우리나라에서는 국어 정보 자료를 구축하여 정보를 쉽고 편리하게 활용할 수 있는 국어 정보화 기반을 조성하기 위해 1998년부터 '21세기 세종계획: 국어 정보화 추진 중장기 사업'을 실시하였다. 이에 따라 본격적으로 한국어 말뭉치(Corpus), 즉 '세종 말뭉치'가 구축되었다. 일정 규모 이상의 크기를 갖추고 그 시대의 언어 현실을 골고루 반영한 기계 가독형 국어 자료의 집합체인 말뭉치를 만들기 위해 현대 국어, 역사 자료, 구어 자료 등 다양한 분야의 자료를 망라하였다. 구축된 자료는 사전 편찬을 위한 용례 추출, 어휘의 빈도 조사, 언어 교육, 철자 교정기 및 번역 프로그램을 만들기 위한

분석 대상 자료 등으로 활용되고 있다.

언어 정보 처리 기술은 기술적인 특성에 따라 언어 처리 기반 기술과 응용 기술로 구분할 수 있다. 언어 처리 기반 기술은 입력된 텍스트의 형태, 구문, 의미, 구조 등을 자동 추출하거나 문장을 생성하는 기술을 말하며 언어 정보 처리 응용 기술은 정보 검색, 자동 번역, 텍스트 마이닝 기술 등을 포함한다.

텍스트 마이닝은 텍스트 데이터에서 형태소 분석과 자연 언어 처리(Natural Language Processing) 기술을 활용하여 일정한 의미 단위로 구획한 뒤 유용한 정보를 추출하는 기법이다. 텍스트 마이닝 기술을 통해 방대한 텍스트 뭉치에서 의미 있는 정보를 추출해 내고, 다른 정보와의 연계성을 파악하며, 텍스트의 빈도를 계산하거나 카테고리를 찾아내는 등 단순한 정보 검색 그 이상의 결과를 얻어낼 수 있다. 자연 언어 처리(NLP)는 인간이 발화하는 언어 현상을 기계적으로 분석해서 컴퓨터가 이해할 수 있는 형태로 만드는 자연 언어 이해, 혹은 컴퓨터의 언어를 다시 인간이 이해할 수 있는 언어로 표현하는 제반 기술을 의미한다. 컴퓨터가 인간이 사용하는 언어, 즉 자연 언어를 분석하고 그 안에 숨겨진 정보를 발굴해 내기 위해 대용량 언어 자원과 통계적, 규칙적 알고리즘을 활용한 형태소 분석이 사용되고 있다.

형태소 분석은 자연 언어 처리의 가장 기본적인 단계로, 어절을 의미를 갖는 가장 작은 단위인 형태소로 분리하고 품사를 찾아내는 것이다(곽수정 외, 2013). 빠른 속도의 형태소 분석을 위해 세종 형태 분석 말뭉치와 같은 기분석 말뭉치를 활용하기도 한다. 세종 형태 분석 말뭉치는 장기간에 걸친 형태 분석 및 검토 과정을 거쳐 형태 분석의 신뢰도가 상대적으로 높기 때문에, 규칙이나 통계를 기반으로 한 형태소 분석기와 결합하여 사용된다.

규칙에 의한 접근은 정해진 규칙에 따른 형태소 분석으로 특수한 부분에 대한 조건까지 제시해 줄 수 있으므로 중의성을 해결하는 성능이 우수한 반면 구축하고 유지 관리하는 데 부담이 많다. 시간과 비용이 많이 들어 많은 규칙을 만들기가 어렵고 규칙 관리자의 능력이 중요하므로 전문가가

필요하다. 띄어쓰기의 규칙으로 휴리스틱(heuristic), 바이어블 프리픽스(viable prefix)를 이용한 최장 일치 기법, 접두 명사 및 접두사와 이웃하는 명사 간의 조합 규칙 이용법, 형태소 분석 결과 이용법 등이 있다.

통계에 의한 접근은 단순하고 계산적으로도 부담이 적으며, 언어의 생산성에 잘 대처할 수 있고 영역 지식의 쉬운 활용이 가능하다. 그러나 이 접근은 통계를 낼 수 있는 정도의 자료(대용량 말뭉치 필요)를 확보하여야 하고 통계를 추출한 해당 말뭉치 분야나 유사 분야에 대해 우수한 성능을 보이지만 다른 분야에는 적용이 힘들다.

텍스트 마이닝은 텍스트로부터 고품질의 정보를 도출하는 과정과 관련이 있다. 고품질의 정보는 일반적으로 통계적 패턴 학습과 같은 수단을 가지고 패턴과 트렌드의 장치(devising)를 통해 도출된다. 텍스트 마이닝은 보통 입력 텍스트의 구조화－보통 파싱(parsing), 몇 가지 파생된 언어적 특징의 추가와 제거, 그리고 데이터베이스에 후속적 추가－, 구조화된 데이터 안에서의 패턴 도출, 그리고 마지막으로 출력의 평가와 이해 등의 과정을 포함한다.

텍스트 마이닝에 있어 고품질은 통상 관련성, 참신함, 그리고 관심도 등의 몇 가지 조합을 의미한다. 전형적인 텍스트 마이닝 기술은 텍스트 범주화(categorization), 텍스트 군집(clustering), 콘셉트/개체 도출, 알갱이(granular) 분류의 생산, 감성 분석, 문장 요약, 개체 관계 모형화(예: 개체 인식 사이의 관계 학습) 등을 포함한다.

텍스트 분석은 정보 검색, 단어 빈도 분포를 연구하는 어휘 분석(lexical analysis), 패턴 재인(pattern recognition), 태깅/주석(tagging/annotation), 정보 추출, 링크와 연상 분석을 포함하는 데이터 마이닝 기법, 시각화(visualization), 그리고 예측 분석(predictive analytics)을 포함한다. 핵심적으로 무엇보다도 중요한 목표는 분석을 위해 자연 언어 처리(NLP)와 분석 방법을 활용하여 텍스트를 데이터로 바꾸는 것이다.

텍스트 마이닝의 전형적인 활용은 자연 언어로 쓰인 일련의 문서를 스캔하

거나, 예측 분류 목적을 위해 문서군(document set)을 모형화하고, 혹은 추출된 정보를 가지고 데이터베이스 또는 검색 지수에 덧붙이는 것이다.

3. 빅데이터 분석, 언어 자료의 활용

빅데이터를 잘 활용하기 위해서는 우선 데이터를 잘 분석할 필요가 있다. 빅데이터 시대에 맞추어 데이터 분석의 방법도 놀랄 만한 진전이 이루어지고 있어 웹과 소셜 네트워크 서비스상의 다양한 데이터를 수집하고 분석할 수 있게 되었다. 웹과 소셜 네트워크 서비스에는 이용자들의 자발적 참여에 의한 비정형 데이터들이 축적되어 있어 특정 현상에 대한 색다른 분석을 해 볼 수 있다.

한편 디프 러닝(Deep Learning)과 같은 기계 학습(Machine Learning)의 알고리즘이 빅데이터 분석에 활용되기 시작하였다. 알고리즘은 현재 발성된 단어들을 이해할 수 있고 음성 단어를 텍스트로 바꾸어 기술할 수 있다. 그리고 내용, 의미, 감성을 알기 위해 이런 텍스트를 분석할 수 있다. 예를 들면 우리가 어떤 사람이나 사물에 대하여 좋게 이야기하는지 아닌지를 텍스트의 감성 분석을 통해 알 수 있다. 사람들이 세상을 이해하고 미래를 예측할 수 있도록 매일 점점 더 향상된 알고리즘들이 나타나고 있다. 이런 알고리즘은 기계 학습과 인공지능(Artificial Intelligence)－독자적으로 학습하고 의사결정하는 알고리즘의 능력－이 짝을 이루어 괄목할 만한 성과를 내고 있다. 이세돌 9단과 세기의 바둑 대결을 벌여 널리 알려진 알파고(AlphaGo)가 대표적인 사례이다.

빅데이터 분석은 통계학과 전산학의 분석 방법을 주로 사용한다. 정형 빅데이터의 분석에는 데이터 마이닝과 기계 학습의 알고리즘을 대규모 데이터 처리에 맞도록 개선하여 활용하고 있다. 인터넷과 소셜 미디어의 생활화로

비정형 빅데이터가 폭발적으로 증가하면서 비정형 빅데이터의 분석에는 주로 시맨틱 네트워크 분석, 감성 분석, 군집 분석 등을 많이 활용하고 있다.

시맨틱 네트워크 분석(Semantic Network Analysis)은 소셜 네트워크 분석의 한 종류이다. 소셜 네트워크 분석(Social Network Analytics)은 수학의 그래프 이론(Graph Theory)에 뿌리를 두고 있는데 사람이나 사물의 관계를 노드와 링크의 구조로 파악하는 기법이다. 소셜 네트워크의 연결 구조, 연결 중심, 연결 강도 등을 바탕으로 사용자의 명성 및 영향력을 측정하여, 소셜 네트워크상에서 입소문의 중심이나 허브 역할을 하는 노드를 찾는데 주로 활용된다.

시맨틱 네트워크 분석은 소셜 네트워크 분석을 텍스트 데이터에 응용하여 특정 현상의 인식이나 개념의 해석에 있어서 의미의 관계를 중심으로 분석하는 방법이다. 소셜 네트워크 분석이 사람들 사이의 특정 네트워크 특성으로 네트워크에 포함된 사람들의 사회적 행위를 설명하는 시도라면, 시맨틱 네트워크 분석은 소셜 네트워크를 기반으로 개념을 노드로 나타내고 개념 간의 관계를 연결로 나타낸 그래프이다. 개념은 단어나 구로 표현되는 정보 단위이며 의미는 다른 개념들과의 관계 속에 내재되어 있는 것이고 관계는 개념들 간의 연결을 나타내는 개념의 특정 범주를 의미한다. 시맨틱 네트워크는 다양한 개념들을 연결하고 의미가 주요하다는 관점에서 개념(concept)은 관련된 단어들의 합성체로서 사회 네트워크에서의 노드와 같고 개념 간 연결은 서술(statement)이며 네트워크 분석의 선이다.

시맨틱 네트워크 분석은 행위자 간 연결성을 중시하는 소셜 네트워크 분석과 달리 단어들의 공유된 의미를 토대로 체계적 구조를 분석하는 데 주안점을 두고 있다. 시맨틱 네트워크 분석은 핵심 단어 사이의 의미론적 연관이 중요한 요소이고, 핵심 단어의 동시 발생 빈도는 소셜 네트워크 관점의 중요한 요소이다. 시맨틱 네트워크 분석의 장점은 표준화되지 않은 텍스트 자료로부터 구조화된 형태의 정보를 추출함으로써 커뮤니케이션 과정의

양상을 시각화할 수 있다는 점이다. 검색된 결과를 추출하여 핵심 단어의 빈도와 매트릭스 자료를 만들어 핵심 단어 간 관계를 알아봄으로써 전체 데이터에 대한 구조화된 자료를 시각적으로 나타낼 수 있다. 시맨틱 네트워크 분석 방법은 핵심 어휘 및 단어 간의 의미론적 관련성을 규명하는 데 있어서 객관성을 확보하기 위해 유시넷(UCINET), 노드엑셀(NodeXL), 게피(Gephi), 파옉(Pajek) 등 자동화 도구인 소프트웨어가 개발되어 사용되고 있다.

시맨틱 네트워크 분석은 자동화된 도구인 소프트웨어를 활용하여 수행하는데, 가장 먼저 분석의 대상이 되는 데이터의 수집으로부터 시작된다. 데이터 수집의 원천으로부터 획득된 개별 객체들은 메시지 내 주요 단어의 빈도 분석을 통해서 주요 키워드를 도출하는 단계로 이어지고 이를 기반으로 네트워크 다이어그램의 설계를 통한 의미론적 분석이 진행된다.

네트워크 분석을 이용하여 연결 구조의 특성을 파악하는 것은 여러 지표를 통해 이루어진다. 분석은 네트워크를 구성하는 단위들을 노드로 단위들의 관계를 링크로 정의하여 이루어지는데 링크의 연결 정도(degree), 밀도(density) 등을 통해 네트워크가 어떻게 얼마나 결속되어 있는지 그 형태를 알아볼 수 있다. 시맨틱 네트워크 분석에서 활용되는 여러 지표 중에서 가장 중요한 개념이자 많이 쓰이는 측정 방법 중 하나는 중심성(centrality)이다. 중심성은 노드가 전체 네트워크에서 중심에 위치하는 정도를 표현하는 지표를 의미하는데, 이는 연결 중심성(degree centrality), 근접 중심성(closeness centrality), 매개 중심성(betweenness centrality) 등으로 세분화할 수 있다(freeman, 1978).

시맨틱 네트워크 분석을 실시한 후에 특정 대상에 대한 에고 네트워크 분석(Ego Network Analysis)을 실시하여 감성 분석(Sentiment Analysis)을 실시할 수 있다. 감성 분석은 소셜 미디어와 인터넷 데이터, 문서 등의 시맨틱 텍스트를 긍정, 부정, 중립으로 판별하여 선호도를 측정하는 기법이

다. 감성 분석은 특정 서비스 및 상품에 대한 시장 규모 예측, 소비자의 반응, 입소문 분석 등에 활용되고 있다. 정확한 감성 분석을 위해서는 전문가와 데이터에 의한 선호도를 나타내는 표현이나 단어 자원의 축적이 필요하다.

또 시맨틱 네트워크 분석의 한 분야로 군집 분석(Cluster Analysis)을 실시할 수도 있다. 군집 분석은 비슷한 특성을 가진 개체를 합쳐 가면서 최종적으로 유사 특성의 집단을 발굴하는 데 사용된다. 예를 들어 페이스북상에 주로 여행에 대해 이야기하는 사용자군이 있고, 자동차에 관심 있는 사용자군이 있을 때 이러한 관심사나 취미에 따른 사용자군을 군집 분석을 통해 분류할 수 있다. 군집 분석은 시맨틱 네트워크 분석에 활용하여 비슷한 의미를 표현하기 위해 사용하는 단어들을 묶어 가면서 담론(discourse)을 발굴하는 데 사용한다.

인터넷의 방대한 언어 자료를 활용하기 위해 거의 자동에 가까운 텍스트 마이닝을 실시하는 다양한 분석 솔루션들이 개발되고 있다. 다음소프트의 '소셜 매트릭스'는 온라인과 소셜 네트워크 서비스에서 공개된 데이터를 실시간으로 수집하고 자연어 처리와 텍스트 마이닝을 통해 언급 횟수, 연관 긍·부정어, 인플루언서(영향력자) 도출 등의 결과를 제공한다. 아르스프락시아(Ars Praxia)의 '심플'은 수집된 텍스트 데이터의 언어 정보 처리를 통해 온라인 위기 관리 모니터링을 온타임에 가까운 실시간성으로 단어의 양뿐만 아니라 양태를 함께 보여 주면서 단어의 파급력과 파급 효과의 확산 속도, 예측 상황 등을 고유 지표를 통해 다각도로 보여 준다. 유저스토리랩(Userstorylab)의 '트렌드믹스'는 주요 포털 사이트의 데이터를 수집한 후 내부의 언어 정보 처리 로직에 따라 분석하여 하나의 주요 단어에 대해 제일 파급력이 높은 사용자가 누구인지, 중요 이슈는 무엇인지 등의 결과를 제공한다. 더아이엠씨(The IMC)의 '텍스톰'은 텍스트 데이터 일관 처리 솔루션으로 웹과 소셜 네트워크 서비스에 수집된 데이터뿐만 아니라 연구자의 내부 데이터도

원시 데이터(Raw data)로 활용하여 전처리, 형태소 분석 등 텍스트 마이닝을 자동적으로 실시하여 단어 빈도, 트렌드 등의 결과를 제공하고 추가 분석을 위한 매트릭스 데이터도 제공한다.

4. 언어 자료로 세상 보기: '알파고' 열풍 분석

소셜 미디어의 일상화로 생활 속의 다양한 장면에 설명과 대상에 대한 인지 및 특정 현상에 대한 의견과 태도를 표현하는 비정형 빅데이터가 실시간으로 인터넷에 축적되고 있다. 소셜 미디어 대부분의 데이터는 언어 자료이고 비개입적 상황에서 자연스럽게 생산되었기에 이를 수집하여 발달된 알고리즘으로 분석할 수 있다면 정확하게 세상을 볼 수 있다. 즉 사회 여론, 기업과 브랜드에 대한 소비자 인식, 브랜드 태도 등을 세종 말뭉치와 언어 정보 처리를 기반으로 하는 텍스트 마이닝을 통한 언어 자료로써 다양하게 조사할 수 있다. 특히 응답자들의 소극적 협조에 의해 조사의 정확도가 떨어지고 점점 예측도 빗나가고 있는 여론조사를 대체해 나가고 있다. 텍스트 마이닝을 통해 정제된 데이터는 연구자가 주요 단어들을 선택한 후 매트릭스 데이터를 만들어 추가적으로 시맨틱 네트워크 분석(semantic network analysis)을 할 수 있다. 주요 의미 단어를 찾을 수 있고 인플루언서(영향력자)와 특정 의미 단어에 대한 인지, 연상(association)을 파악할 수 있다.

지난 3월 우리나라에 바둑과 인공지능(AI) 열풍을 몰고 온 '알파고'를 통해 산업 분야의 언어 처리와 세종 말뭉치 운영의 실제를 살펴보았다. 알파고 빅데이터 분석은 이세돌 9단과 인공지능 알파고의 바둑 대국이 온 나라의 뜨거운 화제가 되자, 알파고의 주요 의미 단어를 통해 연상을 살펴보고 시민들의 인식과 의견을 알아보려는 목적으로 실시하였다.

데이터의 수집은 빅데이터 일관 처리 솔루션인 '텍스톰(TextoM)'을 사용

하여 2016년 3월 3일에서 21일까지 우리나라 대표 포털인 N과 D 사이트의 뉴스, 블로그, 카페 등에서 '알파고'를 수집 단어로 하여 실시하였다. 수집한 데이터의 양은 약 17만 2천여 개였다.

표 1 채널별 상위 빈도 의미 단어

전체	뉴스	블로그	카페
이세돌	이세돌	이세돌	이세돌
인공지능	인공지능	인공지능	인공지능
바둑	바둑	바둑	바둑
구글	구글	구글	인간
인간	인간	인간	구글
승리	딥마인드	승리	사람
딥마인드	승리	딥마인드	컴퓨터
프로그램	프로그램	컴퓨터	프로그램
컴퓨터	개발	프로그램	관련주
개발	쇼크	충격	실수

수집한 알파고 빅데이터의 전처리와 형태소 분석은, 세종 말뭉치를 기반으로 통계적 형태소 분석을 하는 텍스톰을 사용하여 실시하였고, 그 결과는 [표 1]과 같다.

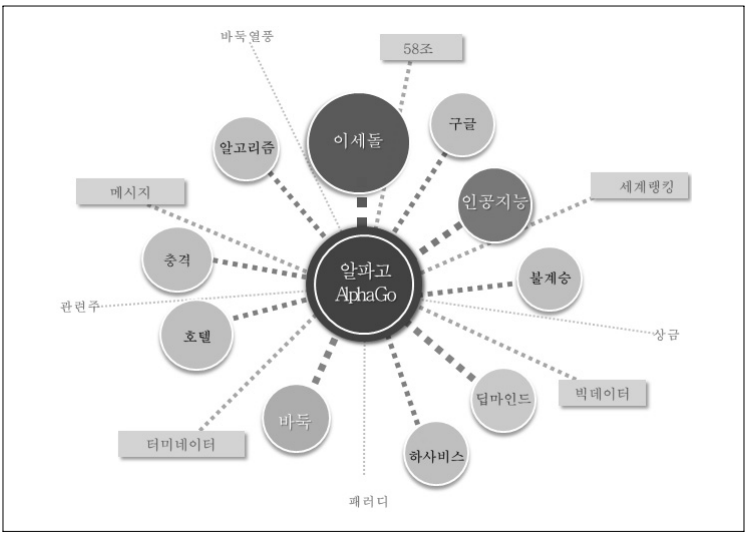
알파고 빅데이터 전체의 상위 빈도 의미 단어를 살펴보면 '이세돌'이 첫 번째에 위치하고 있었다. 알파고와 세기의 바둑 대국을 펼친 이세돌이 알파고 빅데이터에서 가장 많이 등장하고 있어, 알파고가 이세돌과 밀접한 관련이 있음을 보여 주었다. 다음으로 '인공지능'과 '바둑'이 높은 빈도수를 보여 알파고는 인공지능 바둑 프로그램이라는 정체성을 보여 주고 있었다. 그 밖에 '구글', '인간', '승리', '딥마인드', '프로그램', '컴퓨터', '개발' 등의 알파고 개발 주체와 바둑 대국과 관련된 의미 단어들이 나타나고 있었다.

채널별로 보면, 뉴스의 경우 인공지능의 성능에 대한 충격과 이와 연관된

개발의 필요성에 대해 좀 더 의미를 가지는 단어가 상위에 올라와 있었으며 블로그의 경우는 알파고 4승과 이세돌 1승의 바둑 대결의 승패에 대한 의미 단어가 상위에 있었다. 카페의 경우는 알파고와 관련한 주식에 대한 관심도를 보여 주는 의미 단어의 순위가 높았고 알파고와 이세돌의 바둑 대국에서 승패를 좌우했던 바둑 수에 대한 평가와 관련된 의미 단어들이 상위에 있었다. 이와 같은 텍스트 마이닝을 통해 알파고의 화제성과 유명세를 알게 해 주는 의미 단어들이 많이 출현하고 있음을 알 수 있었다.

주요 의미 단어의 영향력과 연상을 파악하기 위해 알파고에 대한 시맨틱 네트워크 분석을 실시한 결과는 [그림 1]과 같다. 알파고 전체 데이터에서 유의미한 단어 100개를 선정하여 네트워크를 그린 결과, 알파고 의미망에서도 바둑 대국을 펼친 이세돌과 가장 강한 연결을 보이고 있었다. 인간과 인공지능의 대국으로 관심이 높아지면서 이세돌과 알파고의 연결성이 매우 높게 나타났다.

그림 1 알파고 의미망



다음으로 '구글', '딥마인드', '하사비스'와 같이 알파고의 개발자에 대한 관심도가 매우 높게 나타났다. 알파고를 개발한 하사비스(Demis Hassabis)는 컴퓨터공학과를 졸업하여 게임 개발자로 활동하면서 인공지능 알고리즘에 대해 관심을 갖고 인간의 뇌구조에 대해 공부했으며, 마침내 인공지능 알파고를 탄생시켰다. 이렇게 알파고를 만든 개발자의 이력이나 경력, 구글과의 관계 등 알파고 프로그램 개발자와 구글 회사에 대한 관심도가 매우 높았다는 것을 의미망을 통해 알 수 있었다.

또한 '인공지능', '구글', '딥마인드', '빅데이터', '알고리즘'과 같은 단어들과도 큰 연결을 보이고 있었다. 알파고가 어떻게 바둑의 수를 스스로 계산하여 둘 수 있느냐에 대한 궁금증으로 이와 연관된 텍스트가 많았으며 이와 관련된 다양한 의견을 보여 주었다. 인공지능 바둑 프로그램인 알파고는 엄청난 경우의 수를 가진 바둑에서 통계적으로 이길 수 있는 확률에 의해 수를 둔다. 사람들은 이러한 알파고의 바둑 두는 방식에 대해 많은 관심을 가지고 있었으며, 이세돌이라는 한 사람과 1,200여 명의 바둑 기사들이 동시에 바둑 대결을 벌이는 상황에 비유하며 대국 자체가 불공정 대국이라는 의견도 있었다.

또한 '불계승', '호텔', '58조', '상금', '바둑 열풍', '바둑 랭킹' 등의 단어는 이번 바둑 대결로 인해 생겨난 유행 현상과 홍보 효과을 보여 주고 있다. 구글의 자산 가치 상승과 대국이 치러진 호텔, 바둑 용어와 바둑에 대한 관심 등 알파고 전체 네트워크에서 관련 기업의 홍보와 마케팅 측면에서의 효과를 보여 주는 단어들과 바둑 열풍과 연관된 단어와의 연결이 높았다.

한편 '충격', '터미네이터'와 같은 단어들과도 연결되어 많은 사람들이 이번 알파고를 영화 속 인공지능과 겹쳐 생각하여 무섭다거나, 섬뜩하다거나 하는 감정을 드러내기도 하는 것을 알 수 있었다. '패러디'는 텔레비전 프로그램에서 알파고의 모습을 패러디하거나 알파고와 관련한 유머러스한 내용들이 있었음을 보여 준다. 알파고가 자사고, 특목고와 같이 특수 고등학교를 일컫는 말이라는 유머를 통해 사회를 풍자하고 뿐만 아니라, 알파고의 천적은

두꺼비집, 바이러스로 컴퓨터 프로그램을 다운시키는 방법들로 바둑 대결을 이길 수 있다는 의미 단어도 있었다. 이처럼 여러 분야에서 알파고를 패러디한 의미 단어가 많았으며 그만큼 알파고가 큰 화제였다는 것을 보여 주고 있다.

5. 언어 자료의 미래

실제로 언어 자료를 통해 세상 보기가 가능한지를 알아보기 위해 '알파고' 빅데이터를 수집하여 분석해 본 결과 알파고의 대결자, 개발자, 인공지능으로 바둑을 두는 방식 등 바둑 대국과 알파고 프로그램 자체에 대한 관심도가 매우 높게 형성되어 있다는 것을 알 수 있었다. 알파고 전체 네트워크에서도 알파고의 주요 의미 단어, 연상, 의견, 태도 등을 통해 세상을 볼 수 있었고 알파고의 화제성을 확인할 수 있었다.

앞으로 정보 통신 기술이 점점 더 발달하면 인간과 디바이스의 커뮤니케이션은 더 일상화될 것이다. 그러면 데이터 생성 속도는 더 빨라지고 데이터양은 기하급수적으로 늘어나며 데이터의 형태도 다양해질 것이다. 이와 같은 정보 통신 환경의 변화에 적응하면서 데이터를 적극 활용하기 위해서는 향상된 언어 정보 처리 기술이 필요하다. 또 데이터 기반의 인공지능이 발달하면 할수록 기계가 데이터를 잘 읽고 말할 수 있도록 하는 언어 정보 처리도 점점 더 중요해질 것이다. 언어 정보 처리가 발달하면서 세종 말뭉치의 활용은 더 늘어나고 중요성이 높아질 것이다. 언어 정보 처리와 세종 말뭉치는 사회간접자본(Social Overhead Capital)과 같은 기반 기술이기에 연구개발을 더 강화해야 한다.

언어 자료가 빅데이터가 되고 분석 알고리즘이 발달하면 더 정확하게 세상을 보고 더 나아가서 미래를 예측할 수도 있을 것이다. 언어 자료의 가치를 높이기 위해서는 시대와 상황의 변화에 적응하는 언어 생활에 맞도록 세종 말뭉치를 지속적으로 구축하고 더 정교화할 필요가 있다.

참고 문헌

권혁철(2004), 인터넷 환경에서 언어 정보 처리 기술의 응용과 발전 방향, 한국어정보처리연구실, http://klpl.re.pusan.ac.kr/.../20040720104458_충북대-인터넷환경에서언어정보처리기술의응용과발전방향.ppt/.

곽수정·김보겸·이재성(2013), 한국어 형태소 분석을 위한 효율적 기분석 사전의 구성 방법, 정보처리학회논문지, ≪소프트웨어 및 데이터 공학≫ 제2권 제12호, 881~888.

김유경(2002), 전자정보통신 연구계를 움직이는 사람들: 언어정보처리, ≪전자신문≫, 2002년 8월 7일 자.

김윤정·이조은·이유리(2016), 네트워크 분석 기법을 통한 패션 상권의 특성 분석, ≪한국의류학회지≫, 제40권 제2호, 203~221.

김정선(2015), 혁신기술로서의 빅데이터 국내 기술수용 초기 특성 연구, 박사학위 논문, 이화여자대학교 대학원.

문화관광부(2003), ≪(21세기 세종계획) 말뭉치 활용 방안 연구≫, 문화관광부.

빅토르 마이어 쇤버거·케네스 쿠키어(2013), ≪빅데이터가 만드는 세상≫, 21세기북스.

서일원·전채남·이덕희(2013), 시맨틱 네트워크 분석을 이용한 원천기술 분야의 잠재적 기술 수요 발굴기법에 관한 연구, ≪기술혁신연구≫, 제21권 제1호, 279~301.

IDC(2011), The 2011 IDC Digital Universe Study.

Marr, B.(2016), *Bigdata in Practice,* Wiley, West Sussex.

Wikipedia. http://ko.wikipedia.org/wiki/자연_언어_처리/(검색일: 2016. 5. 31.).

_____. http://en.wikipedia.org/wiki/Big_data/(검색일: 2016. 5. 31.).

_____. http://en.wikipedia.org/wiki/Text_mining/(검색일: 2016. 5. 31.).

21세기 세종 말뭉치 제대로 살펴보기
— 언어정보나눔터 활용하기

황용주·최정도
국립국어원

1. 서론

잘 알려져 있다시피 '21세기 세종계획'은 1997년에 그 계획이 수립되었고 이듬해인 1998년부터 2007년까지 10년 동안 시행된 한국의 국어 정보화 중장기 발전 계획이다(홍윤표, 2009). 이 사업을 통해 한국어 연구를 목적으로 하는 기초 자료에 대한 갈증을 상당히 풀게 된 것이 사실이다. 최초 이 글은 '21세기 세종계획'이 끝난 지 약 10년이 되는 지금 시점에 다시금 '21세기 세종계획'과 그 성과를 소개하고 세종 말뭉치에 대해 전망하는 것이 목적이었다. 그러나 2009년 ≪새국어생활≫ 봄 호에 이와 관련된 특집이 실려 상세히 전달되고 있는 바 '21세기 세종계획'의 성과와 전망은 모두 홍윤표(2009), 서상규(2009), 홍종선·남경완(2009)에 기대기로 하고, 여기서는 '21세기 세종계획'의 결과물을 간단히 알아보고 실용적인 차원에서 세종 말뭉치의 활용과 관련된 몇 가지를 소개하고자 한다. 그리고 '21세기 세종계획'의 결과물과 관련된 사항 중 그간 알려지지 않았던 몇 가지 도움말을 전하고자 한다.

2. ‘21세기 세종계획’ 결과물의 구성

‘21세기 세종계획’의 성과물은 최초 공개 성과물을 조금씩 수정해 가면서 총 네 가지 판으로 배포되었다. 이들을 간단히 소개하자면 아래와 같다.

- ‘21세기 세종계획’의 성과물

- 첫 번째 배포판: 2007년, DVD 4장
  - DVD 1: 세종 말뭉치
  - DVD 2: 전자 사전
  - DVD 3: 한민족 언어 정보화
  - DVD 4: 전문 용어/문자코드 표준화/글꼴 개발/정보화 인력 양성
- 두 번째 배포판: 2009년, DVD 2장(2007년의 수정판)
  - DVD 1: 세종 말뭉치
  - DVD 2: 전자 사전/한민족 언어 정보화/전문 용어/문자코드 표준화/글꼴 개발/정보화 인력 양성
- 세 번째 배포판: 2010년, DVD 2장(2009년의 수정판)
  - DVD 1: 세종 말뭉치
  - DVD 2: 전자 사전/한민족 언어 정보화/전문 용어/문자코드 표준화/글꼴 개발/정보화 인력 양성
- 네 번째 배포판[1]: 2011년, DVD 2장(2009년의 수정판, ‘한마루2.0’ 포함)
  - DVD 1: 세종 말뭉치
  - DVD 2: 전자 사전/한민족 언어 정보화/전문 용어/문자코드 표준화/글꼴 개발/정보화 인력 양성

1) 네 번째 배포판이 처음 공개되었을 때에는 DVD 케이스가 노란색이었으나, 수량 부족으로 좀 더 제작하면서 다시 공개된 것은 연두색으로 배포되었다. 그 내용은 서로 같다.

다양한 판이 존재하기에 사용자가 혼란을 겪을 수 있겠으나, 처음 배포된 것보다는 최신 배포판을 사용하는 것이 무난할 것으로 보인다. 현재는 DVD 형태로는 더 이상 배포하지 않고, 2013년 개통한 '언어정보나눔터(http://ithub.korean.go.kr/)'를 통해 가장 마지막에 배포된 DVD의 내용을 모두 공개하고 있다.[2)]

2) '언어정보나눔터'는 사용자가 직접 가입을 해야 이용할 수 있다. 기존 DVD에 들어 있는 내용물은 모두 '말뭉치' 또는 '통합자료실' 탭의 '기타 참고자료'에 유형별로 압축되어 등록되어 있다. 따라서 필요한 것을 직접 내려받아 사용할 수 있다.

2.1. 세종 말뭉치[3]

'세종 말뭉치'는 크게 '문어, 구어, 병렬(한영, 한일), 역사' 말뭉치로 구성되어 있다.[4] 사용자의 목적에 맞는 말뭉치를 선택하여 연구할 수 있다.

- 세종 말뭉치의 구성

· 문어 말뭉치

· 구어 말뭉치

· 병렬 말뭉치(한영, 한일)[5]

· 역사 말뭉치

2.2. 전자 사전

세종 전자 사전은 해당 어휘가 가질 수 있는 언어적 정보를 모두 '코드화'하여 기술하고 있는 사전이다. 여기에는 기본적으로 해당 어휘의 의미와 의미 분류, 예문이 포함됨과 동시에 종이 사전에서는 제시하기 힘들었던 결합 정보, 격틀 등등을 갈래뜻(sense) 단위로 부가한 것이기에 사용자에게 큰 도움이 될 것이다.

- 전자 사전(전체 약 5만 항목)

· 체언 25,458항목

· 동사 15,180항목

· 형용사 4,398항목

· 부사 4,320항목

---

3) 각 말뭉치에 대한 자세한 설명은 3장에서 기술한다.

4) '북한 및 해외 말뭉치'는 저작권 등의 문제로 배포되지 못하고 있다.

5) '21세기 세종계획'에서는 '한영 병렬 말뭉치'와 '한일 병렬 말뭉치'가 주로 구축되었다. 이 외에 '한러, 한불, 한중' 말뭉치가 시범적으로 구축되었으나 양이 적은 편이기에 연구 결과의 신뢰성을 담보하기는 어려울 듯하다. DVD를 통하여 배포되지는 않았으나 적은 양이나마 사용하고자 하는 연구자는 '언어정보나눔터'를 통하여 담당자에게 연락하면 자료를 접할 수 있을 것이다.

사용자는 낱낱의 파일을 각종 텍스트에디터에서 열어 확인할 수 있으며, 함께 제공되는 검색 도구인 '~단순 검색기/상세 검색기'를 통하여 연구자가 원하는 어휘 정보를 추출할 수 있다. 참고로 '단순 검색기'보다 '상세 검색기'로 검색하면 검색하는 자료 자체가 상세 사전이기 때문에 더욱 풍부한 내용을 확인할 수 있다.

2.3. 한민족 언어 정보화

'한민족 언어 정보화'의 결과물은 다음과 같이 구성되어 있다.

- **한민족 언어 정보화** [6]
  - · 국어 어문 규정(검색 프로그램)
  - · 국어 어휘의 역사(검색 프로그램)
  - · 한국 방언(검색 프로그램)
    - 남한 방언(검색 프로그램)
    - 북한 방언(검색 프로그램)
    - 중국 및 기타 지역(검색 프로그램)
  - · 문학 작품에 사용된 방언(검색 프로그램)
  - · 남북한 언어 비교 사전(검색 프로그램)

위 자료들은 각각이 '언어정보나눔터'에 파일의 형태로 등록되어 있는데, 사용자는 '통합자료실 > 프로그램'에서 '한민족 언어 정보화 통합 검색 프로그램'을 내려받아서 설치한 다음, 단어 단위의 검색어를 입력함으로써 자신이 원하는 정보를 얻을 수 있다. 예를 들어 검색창에 '부추'를 입력하면

6) 여기서 '국어 어휘의 역사'는 국립국어원 누리집의 '자료 찾기 > 국어 어휘 역사'에서도 확인할 수 있고, '한국 방언'과 '문학 작품에 사용된 방언'은 국립국어원 누리집의 '자료 찾기 > 지역어 자료 > 지역어 찾기'에서도 확인할 수 있다.

‘부추’와 관련된 어문 규정 사항을 알 수 있고, 국어 어휘의 역사에서는 ‘부추’가 어느 시기부터 사용되었는지, 한국 방언에서는 ‘부추’가 어느 지역에서 사용되고 있는지 등을 확인할 수 있어 사용자에게 유용한 정보를 제공한다.

2.4. 정보화 인력 양성[7]

국어 정보화 인력 양성의 일환으로 해마다 ‘국어정보화아카데미’가 개최되었는데, 국어 정보화에 몸담고 있는 연구자들의 강의를 통해서 국어 정보화의 실태를 직접 체득할 수 있는 기회가 되었다. 여기 ‘정보화 인력 양성’에는 그때 사용되었던 강의 자료가 담겨 있어 참고할 만하다. 현재는 ‘언어정보나눔터’의 ‘통합자료실 > 기타 참고자료’에서 당시 ‘국어정보화아카데미’에서 열었던 강의의 자료를 내려받아, 사용자의 연구에 참고할 수 있다.

3. 세종 말뭉치

아무래도 일반인(연구자)가 가장 많이 사용하는 ‘21세기 세종계획’의 결과물은 ‘말뭉치’가 아닌가 싶다. 말뭉치는 머릿속으로 쉽게 만들어질 것이라 예측되지만 실제로는 구축에 시간과 노력이 많이 투여되기에 만들기가 쉽지 않다. 그리고 저작권 등의 문제로 말미암아 쉽게 구해서 쓰기도 힘들다. 이러한 현실에서 ‘21세기 세종계획’의 결과로 배포되고 있는 ‘세종 말뭉치’는 한국어 연구에 지대한 역할을 하고 있는 것으로 보인다. 하지만 아직까지도 말뭉치가 무엇인지 모른다거나 어디서 어떻게 구할 수 있는지를 몰라 활용하지 못하는 사람들이 존재하는 바, 이 자리를 통해 몇 가지 유용한 사항을 전달하고자 한다.

---

7) ‘전문 용어/문자코드 표준화/글꼴 개발/정보화 인력 양성’ 중 일반인이 사용할 수 있는 것은 ‘정보화 인력 양성’이기 때문에 이에 대해서만 언급한다.

3.1. 세종 말뭉치의 내용

우선 세종 말뭉치는 2015년까지 배포되었던 DVD에 포함되어 있기 때문에, DVD를 가지고 있는 사용자는 네 가지 판 중 어떠한 DVD를 사용하여도 같은 내용의 말뭉치를 이용할 수 있다. 하지만 그 이후로는 DVD가 배포되지 않고 그 내용이 모두 '언어정보나눔터'를 통하여 배포되고 있기 때문에, 간단한 절차로 '언어정보나눔터'에 가입하고 로그인을 하면, DVD를 받았을 때와 같이 '21세기 세종계획'의 결과물을 모두 만나 볼 수 있다.[8]

앞서 세종 말뭉치에는 '문어, 구어, 병렬, 역사' 말뭉치가 있다고 했는데 이들은 말뭉치의 종류(유형)에 따라 다시 여러 가지로 구분된다. 말뭉치의 종류를 부가되는 정보의 수준에 따라 구분하면 아래와 같은 말뭉치들이 존재한다고 할 수 있다.

(1) 말뭉치의 종류

· 원시 말뭉치: 원문을 입력해 놓은 말뭉치
· 형태 분석 말뭉치: '원시 말뭉치'에 형태(어휘) 단위의 정보를 부가한 말뭉치
· 의미 분석 말뭉치[9]:
  ① '형태 분석 말뭉치'에 동형어 구분 표지를 부가한 말뭉치
  ② '형태 분석 말뭉치'에 갈래뜻(sense) 구분 표지를 부가한 말뭉치

8) 꼭 로그인을 해야 하는 이유는 자료를 이용하는 데 필요한 '사용 각서'의 작성을 대신하기 위한 것이다. DVD를 신청하여 받을 때에도 이와 같은 절차를 거쳤는데, 사용자가 서면으로 작성하고 다시 우편으로 제출해야 하는 불편함을 줄일 수 있는 것으로 결국 사용자의 편의를 위한 것이다.

9) 세종 말뭉치의 '의미 분석 말뭉치'는 ①에 해당한다. 만약 사용자가 ②와 같은 효과를 얻고 싶다면 ①의 검색 결과에서 자신이 원하는 방식으로 갈래뜻(sense)을 구분해 주는 작업이 병행되어야 한다.

· 구문 분석 말뭉치: '의미 분석 말뭉치'에 통사 구조/기능 표지를 부가한 말뭉치

(2) 말뭉치의 층위

· 원시 말뭉치 ⇒ 형태 분석 말뭉치 ⇒ 의미 분석 말뭉치 ⇒ 구문 분석 말뭉치

원시 말뭉치에서 구문 분석 말뭉치로 갈수록 언어학적 정보가 더 부가되기 때문에, 사용자는 자신이 원하는 언어 정보가 무엇인지를 고민한 다음 대상 말뭉치를 선택하면 된다. 사용자가 검색어로 입력하는 것(검색어)이 보통 음절이 길다거나 동형어가 없는 어휘(나 형태)의 경우는 '원시 말뭉치'나 '형태 분석 말뭉치'를 사용할 것을 권하고, 음절이 짧다거나 동형어가 존재하는 어휘(나 형태)의 경우는 '의미 분석 말뭉치'를 사용할 것을 권한다.

다음으로 현재 배포되고 있는 세종 말뭉치의 양에 대해서 알아보자.

(3) 세종 말뭉치의 양[10]

- 문어 말뭉치

· 원시 말뭉치: 약 3,700만(36,879,143어절)

· 형태 분석 말뭉치: 약 1,000만(10,066,722어절)

· 의미 분석 말뭉치: 약 1,000만(9,071,054어절)

· 구문 분석 말뭉치: 약 45만(433,839어절, 43,828문장)

---

10) 여기서는 '21세기 세종계획'의 전체 보고서에 나와 있는 구축량이 아니라, 실제 DVD 혹은 언어정보나눔터를 통하여 배포되고 있는 말뭉치의 양을 제시한다. 연구자들이 사용할 수 있는 자료는 현재 배포되고 있는 자료가 전부이기 때문에 혼동을 주지 않고자 함이다. 세종 보고서와 결과물의 양이 다른 이유는 저작권과 같은 문제 때문인 것으로 볼 수 있다.

- **구어 말뭉치**
  - · 원시 말뭉치: 약 80만(805,646어절)
  - · 형태 분석 말뭉치: 약 80만(805,646어절)

(2)에서 확인한 바와 같이 언어 정보가 많이 부가될수록, 말뭉치의 구축이 어렵다. 따라서 '원시 말뭉치'에서 '구문 분석 말뭉치'로 갈수록 양이 적어지는 것을 확인할 수 있다. 일반적으로 문어보다는 구어 말뭉치의 구축이 상당히 힘들기 때문에 (3)에서 보듯 그 양에서도 차이가 있으며, 문어에서 구축된 의미 분석 말뭉치가 구어에서는 구축되지 못하여 아쉬운 점이 있다.

한편 2000년대 초반에 세종 말뭉치가 한 번 배포된 적이 있다는 것을 소개한다. 이는 '21세기 세종계획 균형말뭉치'라는 이름으로 배포된 것인데 그 내용은 아래와 같다.

- **21세기 세종계획 균형말뭉치**
  - · 1,000만 말뭉치
  - · 문어 대 구어 비율: 90%:10%(BNC 참조)

'21세기 세종계획'이 진행되고 있을 때 배포된 말뭉치로 영국의 국가 말뭉치인 BNC(1억 어절)의 구성을 참조하여 전체 1,000만 어절로 만들어진 말뭉치이다. 이 말뭉치는 장르 구성이 이미 짜여져 있어 사용자가 고민하지 않아도 되는 균형 말뭉치이며, 자료가 모두 글잡이Ⅱ(색인)용으로 인덱싱이 되어 있어 바로 형태 중심의 검색과 빈도 산출이 가능하다.[11] 이 자료는 언어정보나눔터('말뭉치 > 기타 참고자료')에서 내려받을 수 있다.

11) 원시 말뭉치를 사용하는 글잡이Ⅱ(직접)에서는 '인덱싱' 과정이 필요 없지만, 형태 분석 말뭉치를 사용하는 글잡이Ⅱ(색인)에서는 '인덱싱'이라는 과정이 꼭 필요하다.

3.2. 세종 말뭉치의 활용

여기서는 세종 말뭉치를 어떻게 활용할 수 있는지를 살펴보자. 즉 어떠한 연구에 활용할 수 있느냐를 소개하는 것이 아니라, 어떠한 말뭉치를 사용할 경우 어떠한 도구를 사용해야 한다는 것 등과 같은 실용적인 방법을 소개한다. 아래는 배포되고 있는 각종 도구이다.

- 말뭉치 구축 및 검색 도구
  · 지능형형태소분석기
  · 글잡이Ⅱ(직접, 색인)
  · 한마루, 한마루2.0
  · 한영 병렬 말뭉치 용례 검색기(hepman)

'지능형형태소분석기'는 사용자가 자신의 자료를 만들고자 할 때 이용할 수 있는 형태 분석기이다. 원본이 입력된 자료(원시 말뭉치)만 있으면 언제든지 쉽게 분석이 가능하다. 이 도구를 이용한 분석 결과는 바로 '글잡이Ⅱ(직접, 색인)'에서 검색과 빈도 산출이 가능하다[12].

'글잡이Ⅱ(직접, 색인)'는 검색기이자 빈도 산출 도구인데, 사용자가 직접 구축한 말뭉치를 이용하고자 할 때 사용하기에 적합한 도구이다. 다만 아쉬운 점으로는 2000년대에 개발되었기 때문에 유니코드(UTF)를 지원하지 못하는 것과 닫힌 프로그램이기에 앞으로 수정이 될 가능성이 낮은 것을 꼽을 수 있다. 따라서 사용자는 '글잡이Ⅱ'가 지원하는 형식의 자료로 변환해서 사용하여야 한다. 즉 이 말은 현재 배포되고 있는 세종 말뭉치는 전체가 다 인코딩이 유니코드(UTF-16)로 구축되어 배포되고 있기 때문에 '글잡이

12) '지능형형태소분석기'는 기본적으로 유니코드를 분석하지 못한다(UTF-8의 자료를 분석해 주기는 하지만, 결과물에서 코드가 깨어질 수 있다.). 그래서 세종 말뭉치를 모두 일반 텍스트 코드로 변환하여 분석해야 한다.

Ⅱ'에서는 사용할 수 없다는 말이다. 하지만 앞서 언급했듯이 유니코드(UTF-16)로 저장되어 있는 말뭉치를 '글잡이Ⅱ'의 입력 형식인 '일반 텍스트 코드(CP949, ANSI, ASCII, KS-5601, Euc-kr 등)'로 변환하면 세종 말뭉치를 '글잡이Ⅱ'에서 사용할 수 있다[13].

그렇다면 유니코드(UTF-16)로 되어 있다는 세종 말뭉치는 어떤 도구에서 사용할 수 있을까? 정답은 바로 '한마루'라는 프로그램이다. '한마루'는 세종 말뭉치에 최적화된 검색 도구이다. 다만 2011년 배포판 이전까지 공개되었던 '한마루'는 일정 수의 용례 추출과 구어 검색이 문제가 있었는데, 2011년판부터 공개된 '한마루2.0'은 그러한 문제를 모두 해소한 훌륭한 검색 도구이다. 따라서 앞으로 세종 말뭉치를 활용하고자 할 때에는 무조건 이 말뭉치에 최적화된 도구인 '한마루2.0'을 사용하자. 이 도구는 기본 검색은 빈도 산출(장르별 빈도 산출도 가능), 연어 검색 등등의 고급 기능을 제공하기 때문에 사용자에게 아주 유용한 프로그램이라 할 수 있다. 다만, 이 도구를 사용할 때에는 오직 세종 말뭉치만을 사용하거나, 사용자가 직접 자신의 말뭉치를 세종 말뭉치의 형식으로 구축해서 사용해야 한다는 것만 언급해 두기로 한다.

그리고 한영 병렬 말뭉치를 다룰 수 있는 도구로는 '한영 병렬 말뭉치 용례 검색기(hepman)'라는 프로그램이 있다.[14] 일반 사용자들은 프로그래밍을 하지 않는 이상 병렬 말뭉치를 그대로 이용하여 검색할 수 없다. 따라서 말뭉치를 이용한 한국어와 영어의 대조 연구를 진행하기 위해서는 이 도구를 이용해야 하고 또한 이 도구에 적합한 검색용 한영 병렬 말뭉치를 입력 말뭉치로 사용해야 한다. 한영 병렬 말뭉치 용례 검색기용 한영 병렬 말뭉치

13) 코드 문제는 도구마다 성격이 조금씩 달라 발생하는 것이지 그 자체가 문제점은 아니다. 한편 검색 대상이 되는 자료를 불러올 때(loading) 파일 단위가 아닌 폴더 단위로 불러와야 하는 점을 언급할 수 있겠다(즉, 한 개의 파일을 불러올 때에도 그 파일을 폴더에 넣어서 폴더 단위로 불러와야 한다).

14) 아쉽게도 다른 병렬 말뭉치를 검색할 수 있는 프로그램을 개발되지 않았다.

는 언어정보나눔터에서 배포하고 있다.

마지막으로 세종 말뭉치에는 (한국어) 역사 말뭉치가 있는데 이 말뭉치를 사용하고자 할 때에는 역사 말뭉치 전용 검색기를 사용해야 한다. 그러한 도구로는 '깜짝새'나 '유니콩크'라는 프로그램이 있는데 홍윤표(2012)의 부록 CD에 첨부되어 있어 활용할 수 있다. 이때 주의할 것은 '깜짝새'라는 프로그램은 말뭉치의 입력 형식이 '.2b'라는 형식이어야 하고, '유니콩크'라는 프로그램은 말뭉치의 입력 형식이 유니코드(UTF-16) 형식이어야 한다는 것이다[15].

간혹 사용하고자 하는 말뭉치가 낱낱의 파일로 되어 있어 한데 묶어 사용할 필요성이 있다거나 '한글' 워드프로세서로 입력되어 있는 여러 파일을 일괄적으로 텍스트 파일로 바꾸어 사용하고자 할 때에는 '디지털 한글 박물관'에서 배포하고 있는 '한글 자료 처리 프로그램'을 사용해 보자.[16]

4. 결론

여기서는 '21세기 세종계획'의 성과와 전망에 대한 것은 특집으로 마련되었던 2009년 ≪새국어생활≫ 봄 호에 기대고, 구체적인 '21세기 세종계획'의 결과물을 간단히 소개함으로써, 세종 말뭉치의 활용 방법에 대한 도움말을 제시하였다. 아직도 세종 말뭉치의 존재를 모른다거나 혹은 세종 말뭉치를 손에 쥐고 있어도 제대로 활용하지 못하는 이들에게 작은 도움이나마 되고자 한 것이 이 글의 목적이다. 각각의 결과물들에 대한 사용법은 '21세기 세종계획'의 보고서나 각 도구의 사용 설명서로 밀어 두고, 주로 실제 말뭉치

15) 이 두 프로그램의 차이에 대해서는 이주현(2013)에 상세히 설명되어 있다.
16) '한글 자료 처리 프로그램'은 디지털 한글 박물관(http://archives.hangeul.go.kr/)에서 내려받을 수 있다.

나 도구를 사용하면서 경험적으로 터득할 수밖에 없는 도움말을 전달하고자 하였다. 그 외 자세한 모든 사항은 국립국어원 누리집의 '자료 찾기 > 연구 결과'에 탑재되어 있는 '21세기 세종계획' 관련 보고서를 통해 확인할 수 있다. 각 보고서에는 사업이 어떻게 시작되고 진행되었는지, 그 결과물은 무엇인지에 대한 내용이 담겨 있고, 자료의 구축 방법과 도구 사용법 등이 소개되어 있어 유용하게 활용할 수 있다.

한편 현재 세종 말뭉치는 조금씩 나이를 먹어 가고 있다. '21세기 세종계획' 이후로 한국어의 전반적인 언어 사실을 담고 있는 자료가 구축되지 않아, 현재 배포되고 있는 세종 말뭉치가 조금씩 오래된 말뭉치가 되어 가고 있는 것이다. 일부 민간 단체에서 말뭉치가 구축되고 있기는 하지만 특정 영역의 말뭉치이기나 검색 결과만 확인할 수 있는 말뭉치가 대부분이다. 한시라도 바삐 한국 사람들이 시대별로, 영역별로 한국어를 어떻게 사용하고 있는지를 확인할 수 있는 자료가 구축되고, 그 자료에서 분석된 언어 정보가 말의 저장고인 사전에 오롯이 담겼으면 하는 바람이다.

참고 문헌

서상규·한영균(1998), ≪국어 정보학 입문≫, 태학사.
서상규(2009), 국어 특수 자료 구축의 성과와 전망, ≪새국어생활≫, 19-1(봄), 국립국어원.
이주현(2013), 17세기 국어의 명사형 어미 연구, 연세대학교 석사학위 논문.
최정도(2011), 말뭉치를 이용한 사전 편찬에서의 몇 문제에 대하여, ≪언어 사실과 관점≫, 27, 언어정보연구원.
홍윤표(2002), ≪한국어와 정보화≫, 태학사.
_____(2009), 21세기 세종계획 사업 성과 및 과제, ≪새국어생활≫, 19-1(봄), 국립국어원.
_____(2012), ≪국어 정보학≫, 태학사.
홍종선·남경완(2009), 국어 정보화 사업의 미래와 전망, ≪새국어생활≫, 19-1(봄), 국립국어원.

국어 정보화와 전문용어 표준화의 선구자
— 최기선 한국과학기술원 교수를 만나다

최기선(한국과학기술원 전산학과 교수)

질문자 이경우(서울신문 어문팀장)
때 2016. 6. 8.(수) 곳 교수 연구실

1970년대 중반 대학생이 된 청년은 수학을 택했다. 그런데 그에게 그만 컴퓨터가 눈에 쏙 들어오고 말았다. '컴퓨터가 언어를 이해할 수 있다면? 사고를 한다면?'이라는 생각에까지 미쳤다. 도서관에 살다시피 하면서 관련 서적을 탐독하기 시작했고, 학술대회를 쫓아다녔다. 그러던 어느 날 꼭 읽어야 할 책이 생겼는데, 거금 3만 원이었다. 어머니는 그런 책이 있는지 믿기지 않아 아들과 함께 서점에 갔다. 대학가 한 달 독방 하숙비가 3만 원, 80㎏들이 쌀 한 가마니가 2만 원 정도 할 때였다.

전산학과가 없던 시절, 대학원에서는 응용수학을 할 수밖에 없었지만, 그는 본격적으로 언어공학을 파고들기 시작했다. 언어와 관련한 지식이 필요하다는 생각에 국어학 수업도 열심히 들었다.

국어 정보화, 그는 컴퓨터가 국어를 분석하고 이해하도록 하는 일을 개척해 나갔다. 이것은 국가의 힘을 키우고 지식정보화 시대, 정보에서 소외되는 사람이 없게 하는 것이기도 했다.

주시경 선생은 일찍이 "말이 오르면 나라가 오른다."라고 말했다. 그는 전문용어가 정비돼야 이 분야가, 과학이 발전한다고 여겼다. 또한 사회 수준

이 전반적으로 같이 올라간다고 믿었다. 그래서 전문용어를 표준화하는 일에도 발 벗고 나서게 됐다. 그는 전문용어가 제대로 정비되는 생태계를 만들어야 한다고 말했다.

2015년 한글날 그동안의 공로를 인정받아 옥조근정훈장을 받았다.

이경우 선생님, 반갑습니다. 지난해 한글날상을 받으셨는데 늦었지만 축하드립니다.

최기선 고맙습니다.

이경우 축하를 많이 받으셨을 텐데요. 어떤 소감을 전하셨는지요.

최기선 한글날 수상자들이 대개 어문 계열에 있는 분입니다. 공학계이고 전산학에서 문화 관련 상을 받은 게 뜻깊었습니다. 한글은 휴대전화에서도 쓰고 우리 생활에 밀착돼 있습니다. 여기에다 한국어 음성으로 검색하는 게 공학이니까 거기서 평가받았다는 데 큰 보람을 느낍니다.

이경우 선생님처럼 전산학 분야에서 국어와 관련해 상을 받으신 분이 없는 것으로 알고 있습니다. 국어 정보 처리 분야에서 오랫동안 연구를 해 오셨고 만만치 않은 족적을 남기셨다고 들었습니다. 전산학을 하시면서 국어와 관련해 연구하시는 게 특별해 보이는데, 어떤 계기가 있었는지요.

최기선 대학교 1학년 때 컴퓨터가 막 쓰이기 시작했죠. 학교 전체에 컴퓨터가 한 대 있을 때였죠. 컴퓨터가 막 쓰이기 시작할 무렵인데, 막연하게 컴퓨터가 말을 하거나 생물학적인 구조가 있다면 대단하겠다는 생각을 했죠. 확실치 않지만 어떤 글을 읽었던 것도 같습니다. 이런 생각을 나도 모르게 믿었죠. 그래서 이쪽으로 많이 생각하고 있었습니다. 그러다가 한국과학기술원(이하 '카이스트')

석사 과정에 입학하게 됐습니다. 카이스트가 홍릉에 있을 때였죠. 한국외국어대도 홍릉에 있었는데요, 박사 과정 1학년 어느 날 우연히 거기를 가다가 게시판에서 서울언어학국제학술대회(Seoul International Conference on Linguistics)가 열린다는 안내문을 보게 됐습니다. 그 프로그램을 보니까 컴퓨터를 가지고 언어를 분석하는 게 있더라고요. 잘 몰랐지만 가 봤습니다. 가서 보니까 제가 배우는 오토마타(auto-mata) 이론이 이용되더라고요. 관련된 책을 하나 알게 됐는데, 그 책을 한번 꼭 보자는 마음이 생겼습니다. 그래서 서울 종로 종각에 있는 종로서적센터 외서부에 가니까 그 책이 있더라고요. 너무 비싸더군요. 당시 3만 원이었습니다.

이경우 그 책을 샀습니까.

최기선 네, 그런데 학교에 사 달라고 할 생각을 못 하고 어머니께 사 달라고 했죠. 어머니께서 무슨 그런 책이 있느냐고 하시더군요. 그러면서 직접 같이 가 보자시더니 두말없이 그 책을 사 주시더라고요.

이경우 어머님이 같이 가셨어요?

최기선 너무 믿기지 않는 큰돈이니까 어머니가 가서 직접 보자시더군요. 정말 읽고 싶다고 했죠. 읽고 싶더라고요.

이경우 정말 호기심이 많으셨나 봅니다.

최기선 그랬던 거 같아요. 그 책을 한 열 번쯤 읽으니까 알겠더라고요. 그게 시작이었습니다. 그다음부터는 하루 반은 도서관에 살았습니다. 좀 막무가내로 했죠.

이경우 선생님께서 대학 들어갈 당시만 해도 전산학이라는 것이 낯선 학문이었잖아요. 우리나라에 거의 갓 들어왔을 때였나요.

최기선 그렇죠. 저는 대학에서는 수학과를 나왔습니다.

이경우 아, 대학에서는 원래 수학을 공부하셨군요.

최기선 네, 학부 때는 수학을 공부했습니다. 그런데 옆에 계산통계학과가

생겼어요. 계산통계학의 김현택 교수님께서 컴퓨터를 가르치셨지요. 카이스트 들어올 때도 응용수학과였습니다. 들어와서 1년 지날 때 전산학과라는 게 생겼죠.

이경우 그러니까 처음부터 전산학을 하신 게 아니었군요.

최기선 뭐, 없었으니까요.

이경우 연구실 오기 전에 선생님 공적이 담긴 보도자료를 봤는데요. 봐도 잘 모르는 내용이 있더군요. '국어 자연언어 정보 처리 분석기'를 만드셨다고 돼 있습니다. 얼른 머릿속에 들어오지 않는데 어떤 일을 하신 건지요.

최기선 현재의 정보 검색, 인터넷 검색이 가능하게 한 것입니다. 제가 1992년께 '정보 검색'이라는 강의를 열었습니다. 지금 네이버에서 검색하는 프로그램을 만든 사람들이 그때 학생들이었지요. 지금 컴퓨터에서 한국어를 다루는 데 기본적인 일들을 맨 처음 다 했습니다. 예를 들어 화학공학에는 화학물질을 분석하는 기술이 있습니다. 새로운 화학물질을 만들어 내기도 하는데, 그때 실험 장비가 필요하잖아요.

이경우 그렇겠죠.

최기선 그런 실험 장비라는 게 결국은 말뭉치 또는 코퍼스(Corpus)라는 것이고요. 그것을 만드는 일을 우리나라 최초로 한 거죠. 대량으로 그것을 많이 만들고 공개했습니다. 그걸 바탕으로 '21세기 세종계획'이라든지 이런 것들이 가능하리라 보게 됐습니다. 스펠링을 고치는 프로그램들도 최초로 제가 다 만들었죠. 이런 것을 보고 사람들이 '되는 구나' 하고 따라서 한 거고요.

이경우 좀 더 구체적으로 자연언어 처리 분석기에 대해 설명을 해 주시죠.

최기선 구글이나 네이버에서 검색하는 데 들어가는 프로그램이나 이론들이죠. 그런데 그걸 한국어가 되게, 한글 처리가 되게 기본적인 일

들을 한 거죠. 이것을 최초로 만들어서 이쪽 분야의 바탕이 되게 했습니다.

이경우 이제는 컴퓨터가 한국어를 인식하는 능력이 훨씬 발전하지 않았습니까.

최기선 그렇죠. 옛날에는 키워드를 쳐서 검색을 했지만 지금은 문장을 쳐도 이해할 수 있습니다. 또 요약까지 해 주기도 합니다. 구글에서 '버락 오바마'를 치면 오바마에 대한 특성들이 간추려져 일목요연하게 나오기도 하잖아요. 거기에 필요한 기본적인 일들을 지금도 하고 있습니다.

이경우 그럼 거기에 있는 엄청나게 많은 정보, 즉 빅데이터를 가지고 언어 변화, 언어 분석을 쉽게 해 줄 수 있는 거죠?

최기선 그렇습니다. 빅데이터는 여러 가지 측면이 있는데요, 언어 분석이라고 하는 것을 제가 처음 했을 때는 문법을 직접 썼어요. 규칙을 다 쓰고 문장의 구조를 언어학의 이론대로 만들었습니다. 지금은 확률 통계로 해서 빅데이터에서 반복되는 패턴을 봅니다. 이런 조건일 때는 이렇다든가 하는 경우의 수가 많아지면 예측을 더 정확하게 할 수 있는 거죠. 지금 상업적인 분야에서도 많이 이뤄지고 있습니다.

이경우 지금은 지식정보화 시대이고 정보가 넘쳐 나고 있습니다.

최기선 수많은 웹에서 만들어진 지식정보들을 추출해 내야 하는데요. 텍스트도 굉장히 많죠. 이에 비례해서 지식정보에 대한 것도 점점 커지고 있거든요. '시맨틱 웹'이라는 분야에서도 커지고 있는데, 이것 역시 또 하나의 빅데이터입니다. 지식이라고 하는 것과 웹 텍스트가 커지면 커질수록 상응하는 패턴들이 굉장히 많아져서 앞으로는 텍스트를 지식화할 수 있습니다. 무슨 일을 하는 건지 이해할 수 있는 방법을 기계에 부탁할 수 있는 거지요.

이경우 최근 서울 강남역과 구의역에서 사망한 이를 추모하는 포스트잇을 많이 붙이는 일이 일어났습니다. 이 포스트잇에 들어 있는 글들을 분석해 어떤 현상들을 읽을 수도 있겠습니다.

최기선 그렇죠.

이경우 감정까지 읽을 수 있을까요.

최기선 감정도 일종의 분류 체계라고 보면요. 우울한 것도 있고, 화난 것도, 기쁜 것도 있습니다. 여러 가지 세분화된 잣대들이 있는 거죠. 그래서 나타난 것과 감정의 대응이 결국 하나의 지식 체계를 만들게 되는 것이지요. 이것도 빅데이터화해서 감성 분류를 할 수 있는 거죠.

이경우 요즘 인공지능 얘기를 많이 접하게 됩니다. 지난번 알파고와 이세돌의 바둑을 통해서 더 그렇게 됐지만요. 컴퓨터가 인간의 언어를 배우는 게 아니라 '습득'한다는 말도 나올 수 있는 건가요.

최기선 그렇게 볼 수 있습니다. 습득한다고 하는 것은 머신 리딩인데, 리딩을 한다는 것은 글을 읽고 컴퓨터가 이해를 한 뒤에 어떤 것이 되게 하는 것입니다. 거꾸로 사람한테 자신이 무엇을 알고 있는데 부족한 게 뭐냐, 또는 어떻게 하느냐 하는 것을 얘기할 수도 있습니다. 이런 것을 머신 리딩이라고 하는데, 기계가 독해를 한다는 것이죠.

이경우 이게 지금도 충분히 가능한 건가요.

최기선 가장 뜨거운 분야죠. 지금 학계에서는 열심히 노력하고 있는 분야죠. 이게 언어학적인 이론도 있지만 머신 러닝이나 수학적인 모델에 의해서 어떻게 분석하느냐도 중요한 부분입니다. 이것이 빅데이터이기 때문에 너무 양이 많아서 컴퓨터가 빨리 할 수 있는 방법을 찾아야 하는 것이 큰 문제입니다.

이경우 좀 전에 '시맨틱 웹'이라는 용어를 사용하셨습니다. 선생님의 최고

관심 분야라고 들었습니다.

최기선 자연언어 처리는 문장을 어떻게 이해하고 해석할 것인지에 대한 분석이나, 컴퓨터가 문장을 생성하는 것과 언어 자원, 실험 장치들을 얘기하는 체계입니다. 시맨틱 웹은 그것으로 인해 만들어지는 웹의 다음 세상을 얘기하고 있는 거죠. 지금 웹에서는 웹 페이지를 클릭하면 링크가 있고, '위키피디아'라는 것이 있습니다. 그리고 위키피디아를 정제해 구축한 '디비피디아'라고 하는 데이터베이스들도 있습니다. 시맨틱 웹은 그걸 좀 더 개념화하고 체계화해서 추론도, 이해도 되게 하는 하나의 차세대 웹이죠.

이경우 지금보다 정교해진 다음 버전이라고 생각하면 되나요.

최기선 다음 버전이죠. 그러니까 사람이 이해하는 것만큼 기계가 똑같이 이해할 수 있는 상태인 거죠.

이경우 그럼 그다음은 없는 거겠네요. 사람만큼 이해하게 되는 거니까요.

최기선 그렇죠. 궁극적인 목표는 다 이해를 하게 하는 것이 되겠지요.

이경우 그러면 시맨틱 웹이 궁극적으로 추구하는 목표라고 해야 할까요. 그건 무엇인지요.

최기선 우선 사람이 할 수 있는 것 가운데 이해를 하고 추론을 해서 증명할 수 있는 것들을 기계가 다 할 수 있게 하는 거죠. 예를 들어 우리가 주식 투자를 할 때 현재 상황이 어떤데 어떤 결론에 의해서 하게 되잖아요. 이런 정보를 전부 기계가 이해해서 투자를 해야 할지 말지 의사 결정을 내릴 수 있는 단계가 시맨틱 웹입니다. 그렇지만 기계가 할 수 있는 것과 사람이 할 수 있는 것은 분류가 된다고 봐야죠.

이경우 그렇군요. 저는 이 말씀을 듣고 그럼 사고하는 것조차도 기계가 다 해 주는 것 아닌가 하는 우려가 들기도 했습니다. 그래서 인간이 더 퇴보하는 것 아닌가 하는 의심도 해 보게 되고요.

최기선 그러면 인간은 더 좋은 일을 할 수 있겠죠. 예를 들어 어떤 연구를 할 때 나와 비슷한 연구를 하는 사람이 어디까지 했는지 아는 게 엄청난 일이거든요. 기계가 그것을 어느 정도 알려 주면 저는 그다음 일을 하면 되는 거죠. 연구 속도가 더 빨라지는 겁니다.

이경우 선생님, 그런데 컴퓨터는 어떤 방식으로 대상을 인식하게 되는 건지요.

최기선 지금 시맨틱 웹 세상에서는 위키피디아를 보면 200개 언어가 링크돼 있습니다. 우리나라의 박근혜 대통령은 한글로 쓰인 한국어로 돼 있지만 영어, 독일어, 중국어로도 돼 있지요. 각각의 언어로 된 페이지가 다 있습니다. 그러면 시맨틱 웹은 개념의 세계이니까 이 세상의 모든 사물에 대해 일련번호, 즉 주민등록번호 같은 것을 주게 됩니다. 유아르아이(URI: Uniform Resource Identifier)라고 해서요. 만일 박근혜 대통령이다 하면 웹에서 볼 수 있는 유아르아이가 있습니다. 한글로 어떻게 쓰고 영문자로 어떻게 쓰는지 미국에서는 어떤 방식으로 기술하고 있는지 다 다르겠죠. 하지만 사람에 대해서는 어떤 말을 써도 하나인데, 여러 가지 말로 번역돼 있거든요. 이런 것들이 시맨틱 웹까지는 안 가지만 다국어 데이터 웹이라는 것으로 하나의 간접적인 번역이 되는 거죠.

이경우 언젠가는 컴퓨터가 인간이 하는 모든 말을 이해할 수 있겠네요.

최기선 궁극적으로 그렇게 되는 거죠.

이경우 외국어를 배우는 게 아주 어려운 일인데, 컴퓨터는 쉽게 배워서 사람보다 통역을 잘하겠습니다.

최기선 그걸 지향하고 있죠. 학술적으로나 상업적인 분야에서도 그렇고요. 빅데이터 문제니까 데이터가 쌓이면 쌓일수록 더 잘하게 되는 거죠. 한국어가 다른 나라 언어에 비해 얼마나 데이터화가 많이 돼 있느냐가 중요한 문제입니다. 위키피디아나 웹에 있는 한국어

백과사전들 크기가 영어의 10분의 1도 안 되지요.

이경우 그 정도인가요.

최기선 네, 영어는 약 500만 페이지인데 한국어는 35만 페이지 정도밖에 없으니까요. 그러니까 우리나라 말이 더 많아질수록 더 쉽게 다른 나라말하고도 소통할 수 있게 됩니다. 그 상태가 되면 우리나라의 문화적인 힘도 커지는 거지요.

이경우 구글이 한국어를 번역하는 데 오류를 많이 낸다고 하는데 이것은 한국어와 관련한 데이터가 적어서 그렇다고 봐야 하나요.

최기선 적은 거죠. 직접적인 텍스트 양도 적고 웹에 떠 있는 백과사전의 양도 엄청나게 적습니다. 그리고 아까 말씀드렸듯이 화학공학에서 실험을 하려고 하면 실험 장치가 있어야 하잖아요. 거기에 해당하는 사전 같은 것들도 아주 적죠.

이경우 일부에서는 한국어가 다른 언어에 비해 구조적으로 더 복잡하고 어려운 면이 있어서 그런 것이라고 말씀하시는 분도 있습니다.

최기선 제 생각에는 그렇지는 않은 것 같습니다. 영어는 수요가 많고 실험 장치를 만드는 사람도 많고 실험 장치도 굉장히 다양합니다. 한국어는 그렇지 못하다는 것뿐이죠. 영어와 관련해서는 미국에서 공용 데이터베이스를 구축해 놨어요. 펜실베이니아 대학 등에서 해 놓은 것도 있고 굉장히 많습니다. ≪월스트리트저널≫ 같은 곳에서도 연구하라고 자기네 신문 데이터베이스를 통째로 다 줬습니다. 라이선스 문제 같은 게 하나도 없습니다. 사람들이 단어에 대한 분류도 해 주고, 의미도 데이터베이스에 다 올려놓습니다. 아까 말씀드린 유아르아이가 뭐다 등 온갖 것을 뱅크에 다 올려놓습니다. 많으니까 '뱅크'라고 부르죠. 그것으로 연구를 합니다. 우리나라에는 없거든요. 우리는 거기의 밑바닥에 좀 깔린 상태라고 봐야죠.

이경우 우리나라는 그동안 별 관심이 없었던 건가요.

최기선 관심은 있죠. 영어의 경우에는 그것을 쓰려는 사람이 많고 그 체계에다 모든 것을 공개하는 원칙이 있습니다. 돈을 받긴 받아요. 돈을 안 받는 데도 물론 많고요. 돈을 주면 장치들을 사서 연구를 할 수 있는 모든 장치가 한꺼번에 있는 거죠. 우리나라는 그것을 좀 하긴 했는데 여러 가지 저작권법도 있고 돈을 주고도 살 수 없는 것들이 많은 거죠. 숨겨진 자료들이라고 할 수 있습니다. 우리나라의 연구 평가 제도가 기술료를 받아야지 평가를 받는 체계여서 그렇죠. 이것을 남한테 공개해서 얼마나 쓰였는가에 따라 평가를 하면 안 그럴 텐데, 기술료를 얼마나 많이 받아서 수입이 생겼느냐에 따라 평가를 하거든요. 그러다 보니까 공유에 대한 미덕을 잘 모르는 거예요. 그것 때문에 발전을 못 하는 거죠. 그래서 저변 확대가 안 되는 것이고요.

이경우 정확하게는 언어공학이라는 말을 사용하는데 이 분야와 국어의 발전은 어떤 관계가 있는 건지요.

최기선 서로 시너지 효과가 분명히 있죠. 제가 박사 과정 때 연세대에 가서 국립국어원장을 지내신 남기심 교수님과 영어영문학과 이익환 교수님의 강의를 한 학기 들었습니다. 제가 이쪽 분야에서 한 로직 프로그램으로 규칙을 어떻게 쓰는가에 대해 그쪽에서 받아서 하시는 분들도 계시죠. 지금은 언어공학의 방법으로 국어사전에 용례를 많이 찾아서 싣고 있습니다. 그리고 사람들이 많이 쓰는 정의에 대한 태그도 만들어 줄 수 있습니다. 그러면 국어사전들을 좀 더 생활 밀착형으로 만들 수 있는 거죠.

이경우 사전 말씀을 하시니까 아까 잊었던 게 생각나는군요. 선생님 공적 사항 보니까 전산 쪽 사전이 하나 있더라고요.

최기선 저기 보이는 '다국어 어휘 의미망'인데요. 중국어, 일본어, 한국어,

영어가 있고, 숙명여대 교수께서 만든 독일어도 있습니다. 이게 우리나라에서는 최초인데요. 영어 워드넷(WordNet)과 똑같이 출발했습니다. 결국은 영어 워드넷이 중심이 돼서 거기 개념 번호와 우리 개념 번호를 일치시켜 놨어요.

이경우 컴퓨터 사전을 개발했다는 표현을 썼던데요.

최기선 네, 컴퓨터용 사전 개발이죠. 그게 사람이 읽을 수 있는 형태가 아닙니다. 온갖 기호가 들어가 있어요.

이경우 컴퓨터가 보는 사전인가요. 이해가 안 가더라고요. 도대체 컴퓨터가 보는 사전이 뭐죠.

최기선 컴퓨터가 이해할 수 있는 사전입니다.

이경우 사람만 사전을 보는 게 아니군요.

최기선 그 사전을 펴 봐도 사람은 읽을 수 없죠.

이경우 그러니까 한국어를 이해할 수 있는 컴퓨터용 사전인가요.

최기선 네, 그렇죠.

이경우 공문서들을 보면 문장이 매끄럽지 못한 것들이 많은데요. 이런 것에 대한 컴퓨터의 지도도 가능하겠네요.

최기선 미국이나 영국에서는 언어라는 것이 공공성을 추구하는 면이 있는데 복합명사여서 어려우면 다 풀어 써야 하고, 문장도 쉽고 간결하게 잘 만들어야 하는 원칙들이 있습니다. 그 원칙들을 컴퓨터로 만들어서 쉽게 쓸 수 있게 하는 부분과 제약 언어와 관련된 부분에 대한 워크숍을 지난주에 주최했죠. 제가 국제표준화기구(ISO) 일도 하는데요, 여기서 '언어자원운영'도 제가 창립했습니다.

이경우 표준화라는 것이 무엇을 표준화하는 건지요.

최기선 표준에는 무엇을 만들기 위한 하나의 절차, 가이드라인에 대한 표준이 있고요. 또 명세에 대한 표준이 있습니다.

이경우 명세요.

최기선 네, 시맨틱 웹은 하나의 어떤 방식으로 하잖아요. 예를 들어 에이치티엠엘(HTML)을 어떻게 쓴다고 하는 방식이죠. 이것을 어떻게 쓰는지에 대한 표준이 있고, 글을 어떻게 쓰면 된다든지, 전문용어 같으면 전문용어 개발 원칙에 대한 표준이 있는 거죠. 그러니까 두 가지입니다. 직접적인 대상에 대한 표준이 있고, 그것을 만들기 위한 절차에 대한 표준이 있습니다. 그리고 용어나 심벌에 대한 표준도 있고요. 제가 하는 '언어자원운영'에서는 절차 표준이 많고요. 원칙이 있고 언어 분석을 했을 때 분석의 결과는 어떤 형태가 돼야 한다는 기준이 있는 거죠.

이경우 그곳에서 하는 일이 어떤 용어에 대한 표준보다 컴퓨터 언어의 표준이라고 봐야 하는군요.

최기선 그렇죠. 컴퓨터가 처리했을 때의 분석 결과 형태는 어떠해야 하고 분석하는 절차는 무엇 무엇이 있는 것을 추천한다고 돼 있는 것이지요.

이경우 여기 플러그가 있는데 이것의 규격을 모두 맞춰야 하는 것과 같은 건가요.

최기선 소켓으로 치면 우리나라는 2구로 돼 있고 일본 같으면 일자로 돼 있습니다. 이런 게 형태 자체에 대한 표준이고요, 플러그를 만들 때 절연체가 얼마만큼 들어가야 하는지 추천하는 것은 절차 표준이죠. 저는 언어에 대한 것을 하니까 해석을 하면 뒤에 동사라고 쓸 거냐, 어떻게 하고 동사라고 할 거냐 등 결과에 대한 명세를 하는 것이지요.

이경우 아까 유아르아이를 말씀하셨는데 유아르엘(URL: Uniform Resource Locator)은 많이 들어 봤어도 유아르아이는 생소합니다.

최기선 유아르엘은 어드레스, 그러니까 로케이터(locator)이고 유아르아이는 아이덴티파이어(Identifier)입니다. 최기선 하면 최기선에 대

한 유아르아이를 줘요. 또 다른 최기선도 있을 수 있는데, 그러면 '최기선 1'이다 하면 저고, '최기선 0'이다 하면 다른 사람이 되는 거죠. 웹에서 식별할 수 있는 주민등록번호 같은 거죠. 기계가 자동적으로 최기선에다 번호를 부여해 가는데 컴퓨터가 자동적으로 인덱스를 만들어 가는 거예요. 저에 대한 최기선도 있고 문장 속에서 어디선가 선생님이라는 말도 있고 했을 때 의미적인 색인을 만든다면 첫째 줄의 최기선과 25번째의 선생님은 똑같이 지식 베이스에서 '최기선 1'이라고 인덱스를 만들어 주게 되죠. 이렇게 되면 트위터나 신문에서 어떤 최기선을 써도 컴퓨터가 다 알아서 해주는 거죠.

이경우 누군가가 새로 태어나서 최기선이란 이름을 갖게 되면 어떻게 되죠.

최기선 새로 태어난 아기가 지금까지 나온 최기선과 다른 양상을 보이면, 새 개체로 인식해 새로운 엔티티(단위)를 넣어 주죠. 컴퓨터가 알아서.

이경우 쉽지 않습니다. 국어를 전공하지는 않으셨지만, 누구보다 국어학에 대한 지식도 대단하신 것 같습니다. 학교 다닐 때 국어 과목도 좋아하셨겠습니다.

최기선 고등학교 때 국어를 잘했습니다. 고문도 좋아했고요. 이 공부를 하면서 좀 하면 되겠더라고요.

이경우 전문용어와 관련해서도 많은 일을 하신 것으로 알고 있습니다. 전문용어에 대해 관심을 갖게 된 배경은 무엇인지요.

최기선 1990년 텍스트 분석을 하다가 텍스트 색인 만드는 방법을 독일에서 발표하게 됐어요. 그런데 일본 전문용어협회에서 발표를 보고 일본에 한번 오라고 하더라고요. 그래서 갔더니 한·중·일 전문용어가 링크되면 좋을 거 같다는 얘기를 하더군요. 처음 듣는 거였죠. 정말 중요하겠더라고요. 무슨 얘기를 하는지 서로 알아들으면

정말 좋겠다는 거였죠. 국내 배터리 가게에서도 일본 말로 '마후라'라고도 하고 '머플러'라고도 씁니다. 이렇게 용어 통일이 안 되던 상황이었는데 '21세기 세종계획'을 진행하다가 전문용어센터를 만들게 됐지요. 그리고 그걸 표준화하게 됐죠. 저는 자연어 처리를 하면서 용어가 중요하다는 생각을 하게 됐는데, 전문 분야에 들어가서 용어를 모르면 이해할 수가 없거든요. 이것 때문에 일단 시작을 하게 됐습니다. 기계 매뉴얼 같은 곳에 해설을 달아야 하는데 안 되는 거예요. 전문용어는 지금도 안 되는 게 많이 있어요. 사전이 없어서 그렇지요. 이게 첫째 이유고요, 둘째 이유는 학술용어와 산업용어, 표준용어가 서로 연결이 안 돼요. 예를 들어 '디프 러닝(Deep Learning)'이 나오니까 그냥 쓰잖아요. 디프 러닝을 심화학습이나 심층학습으로 바로 쓰지는 못하는 거죠. 그런데 디프 러닝이라고 했을 때 이 말을 우리나라 사람들이 과연 얼마나 이해할지 모르겠어요. 영어 쓰는 사람들은 너무 쉬운 거죠. 우리나라 사람들은 왜 심화학습이나 심층학습 같은 말을 사용하지 못할까요. 한국어로 돼 있지 않으면 용어가 우리에게 주는 효과는 없는 거나 마찬가지입니다.

이경우 일상생활에서도 전문용어가 미치는 영향이 상당할 텐데요.

최기선 상당히 많이 있습니다. 영어와 일본어에서 전문용어 통계를 보면 영어에서는 일상용어와 전문용어가 겹치는 비율이 약 70%나 됩니다. 사람들이 일상에서 쓰는 용어와 전문 분야에서 사용하는 용어가 상당수 같으니까 일반 국민의 수준이 높아지는 거지요. 일본도 이보다는 못하지만 40%는 똑같다고 합니다. 일본에서 낸 통계지요. 우리나라는 이런 게 체계화돼 있지 않은 거죠. 그런 생태계가 안 돼 있는 거예요. 디프 러닝이라고 하면 분명하게 알아듣는 사람이 얼마나 되겠어요.

이경우 전문용어가 더 쉬워지고 일상용어에 가깝게 된다면 국력이 커지고 사회 수준이 높아지는 거겠죠?

최기선 당연한 말씀이죠. 스스로 생각할 수 있는 힘이 커지는 거잖아요. 그러면 얼마나 파급 효과가 커지겠어요. 2002년 일본에서 노벨화학상을 받은 화학자는 잘 알려지지 않은 대학을 나왔고, 조그만 기업에서 실험 장비를 만들던 사람이잖아요. 일본에 가면 일본어로 다 교육하지 영어 교과서 안 쓰더라고요. 그러니까 그 친구는 자생적으로 자기 생각을 표현할 수 있는 거지요. 말이 이해되고 와 닿으니까 그런 일을 한 거죠.

이경우 전문용어들이 일본어로 돼 있으니까 일본어로 이해하고 자기 생각을 하게 돼서 상을 받은 거란 말씀이군요.

최기선 그렇죠. 노벨상도 영어 잘하면 되는 거 아니냐고 그러는데 한국 사람이 영어로 추론하는 것보다 한국어가 훨씬 더 낫죠.

이경우 언어 규범도 중요하지만, 전문용어 분야 또한 이에 못지않게 중요한 분야 같습니다.

최기선 우리가 한국어 자연어 처리를 할 때 전문 분야 특허 해설을 해야 하고, 교과서에서도 전문 분야 해설을 해야 하거든요. 그런데 물리학도 있고 화학도 있고 세계사도 있잖아요. 그럼 그걸 다 이해해야 되잖아요. 그런데 지금은 개념들에 대한 정리가 잘 안 돼 있기 때문에 힘듭니다.

이경우 전문용어 분야에서도 선생님께서 선구자적인 역할을 하신 거네요.

최기선 제가 처음 시작할 때는 하시는 분들이 별로 없었어요. 그래서 오스트리아 빈에 계시던 크리스티안 할렌스키라는 분을 찾아가서 배웠습니다. 이후에 교육부나 문화체육관광부에서도 이 분야의 중요성을 알고 여러 방면으로 시작했죠. 그렇지만 원칙이 없는 상태에서 일을 하게 됐죠. 각계에서 만들어 놓은 덩어리들이 있습니다.

이것들이 각급 학교 수준에 맞게 교과서에 들어가야 하는데 뒤죽박죽인 상태가 된다는 얘기를 많이 들었습니다. 결국 전문 분야가 정리돼야 컴퓨터가 텍스트를 다 이해할 수 있는 경지가 되는 거죠. 번역도 마찬가지고 검색도 마찬가지인 거고요, 지식정보도 마찬가지일 거고, 국민 수준 향상 이런 것도 그렇고요. 지금 당장 닥친 것은 학술적으로 막 등장하는 디프 러닝이나 산업화되는 특허에 쓰는 용어들도 있고요. 이게 정리가 안 됐기 때문에 한쪽에서는 파란색으로 보이고, 다른 쪽에서는 하얀색으로 보이게 됩니다. 그래서 혼동이 일어나게 되는 거죠.

이경우 신문이나 방송에 나올 정도면 일반 대중도 어느 정도 감을 잡을 수 있어야 하는데, 외국어가 대개 그대로 들어오는 거잖아요.

최기선 하나의 생태계를 만들고 이걸 다룰 수 있는 체계가 있어야 합니다. 학계에서 준비가 돼 있어야 하죠. 학계가 이걸 하기에는 너무 바쁘거든요.

이경우 1998년 전문용어 작업을 시작한 이후 많은 시간이 지났습니다. 그동안 여러 제안도 하셨을 거고요. 그런데 그 생태계가 아직 정리되지 않고 있는 거죠. 이것은 정부가 해야 하는 거잖아요. 국어원이 나서야 하나요.

최기선 과학기술 분야는 여러 연구재단도 있고 한국 과학기술한림원 같은 곳도 있습니다. 그런 곳에서도 일할 수 있는 부서가 커지고 정부도 지원을 하고, 저 같은 사람은 생태계를 만들고 그래서 그것이 작동되도록 해야 합니다.

이경우 이제 기대하시는 만큼의 인식 변화가 됐다고 할 수 있는지요.

최기선 여러 번 시행착오가 있었고요, 지금은 이러한 것에 대한 이해가 많이 되고 있습니다. 이러한 밑바탕에 대한 알맹이들이 제대로 정리되고 잘 쓰이도록 해야죠. 인공지능 같은 사업과 병행돼야 하는

데, 많이들 도와주시는 것 같습니다. 분위기는 조금 달라진 것 같아요.

이경우 아직 좀 형식적이지 않은가요.

최기선 이 분야가 융합적이잖아요. 문체부도 있고 미래창조과학부도 있습니다. 연결해서 해야 하는 일인 것 같습니다. 용어가 표준화된다면 공부하는 사람에게도 얼마나 좋겠어요.

이경우 결국 과학기술이 발전하려면 국어가 밑바탕이 돼야겠습니다.

최기선 뭔가 연결될 색인이 있어야 찾을 수 있는 거죠. 이걸 찾아 나서야 하는 수고가 필요한 상태입니다. 개념화되어 논리적으로 정리가 돼 있으면 기계가 이해하고 추론할 수 있는 능력이 커지기 때문에 웹 문화, 웹 과학 등 우리가 과학 전 분야에서 앞서게 되는 거죠.

이경우 선생님께서 말씀하신 것들이 이뤄지려면 과학계에 계신 분들이 더 적극적으로 말씀해 주셔야 할 것 같습니다. 전문용어 정리의 중요성에 대해서 말입니다.

최기선 앞에서 말씀드린 일본의 그 친구가 노벨화학상을 받았듯이 전문용어가 쉬워지고 표준화되면 할 수 있는 게 굉장히 많아집니다. 그런 사람들이 노벨상을 받으려면 용어가 더 낮은 수준까지 내려가야 합니다. 디프 러닝이라고 하지 말고 누구나 이해할 수 있는 수준까지 가야 하는 거죠. 말이 쉬워져야 하고 교육이 거기에 맞춰서 가야 합니다.

이경우 선생님께서는 남북한 전문용어 표준화 작업까지도 하셨다고 들었습니다. 잘 진행되고 있나요.

최기선 다른 방향에서 사전 만드는 것은 진행되는 것 같아요. 1995년도께 중국 연길시에서 '코리안언어학회'라는 것이 열렸죠. 거기서 전문용어에 대해서 하니까 북쪽에서도 그걸 하시는 분이 나오셨죠. 거기는 전투적인 용어를 많이 쓰더군요. '스택'이라고 있는데 한쪽

에서 다른 쪽으로 빠져나가는 거죠. 이것을 거기서는 탄창이라고 사용하더군요. 우리나라도 뭐 대책이 없는 거지만요.

이경우 일본이나 미국에서는 일찍이 중요성을 인식하고 있었네요.

최기선 그렇죠. 일본에서는 메이지유신 때 용어 통일을 한번 했고요, 1930년부터 학술용어는 표준화를 계속하는 거예요. 일본은 일을 벌일 때마다 용어 표준화를 먼저 하죠.

이경우 이외에 관심을 갖고 연구하시는 분야는 무엇인지요.

최기선 지금은 주로 한국어 기계 독해예요. 머신 리딩이라는 거죠. 디비피디아라는 것이 있어요. 위키피디아 테이블의 정보들을 디비(DB)화한 게 디비피디아예요. 예를 들어 위키피디아 한 페이지를 보면 여기에 버락 오바마가 있고 글이 있고 오른편에 테이블이 있어요. 오바마의 생일은 언제고 정당은 뭐고 이런 식으로 돼 있는데, 이를 떼어다가 디비화하는 것이 디비피디아예요. 한국어판은 우리가 하고 있어요. 이것을 지식 베이스라고 하죠. 1차적인 지식 베이스예요.

이경우 언어공학을 하시면서 가장 아쉬운 건 무엇인지요.

최기선 교수로서 논문을 발표하고 인정받는 것은 중요한 일입니다. 한데 한국어 데이터를 가지고 발표하면 잘 안 됩니다. 영어만큼 고도의 결과를 보이기가 어렵습니다. 왜냐하면 데이터가 워낙 적으니까요. 그렇다 보니까 학생들이 데이터로서 한국어를 안 쓰려고 합니다. 영어를 쓰려고 하죠. 한국어도 영어만큼 실험을 할 수 있는 데이터나 알고리즘 같은 것이 발전해야 합니다. 그래서 우리 학생들이 자신이 쓰는 말로 자신 있게 발표할 수 있는 세상이 됐으면 좋겠습니다. 이것이 교수로서 첫째로 해야 할 일이라고 생각합니다. 이것을 하기 위한 교과서나 기본적인 데이터나 실험 장치들은 제가 꼭 어느 정도까지는 만들어 놓고 싶습니다.

이경우　네, 오늘 좋은 말씀 잘 들었습니다. 고맙습니다.

최기선　감사합니다.

이집트의 사회언어학적 특징과 언어 정책

윤은경
한국외국어대학교 아랍어과 부교수

1. 머리말

이집트는 북동 아프리카에 위치하고 있는 아랍 국가로서 정식 명칭은 '이집트아랍공화국(Arab Republic of Egypt)'이다. 이집트 인구는 2013년 기준으로 약 8천6백만 명이며 인구의 절반 이상이 20세 이하이다. 이집트는 북동 아프리카의 약 100만㎢를 차지하고 있으며 이 중에, 오직 3만 5천㎢(나일계곡과 삼각주)만이 경작된다. 나머지 국토는 서부(리비아) 사막, 동부(아랍) 사막, 그리고 시나이 반도로 구성되어 있다. 이집트는 서쪽으로 리비아, 남쪽으로 수단, 동쪽으로 홍해, 북동쪽으로 남부 이스라엘, 그리고 북쪽으로 지중해와 경계를 맞대고 있다.

이집트는 인류 최초 문명의 발상지일 뿐 아니라 아프리카 대륙과 아시아 대륙을 잇는 전략적 요충지로서의 중요성 덕분에 오래전부터 아랍·이슬람 제국의 정치, 문화, 교육의 중심지 역할을 해 왔다.

이집트의 공용어는 아랍어[1]이다. 아랍어는 언어 분류상 셈어(Semitic

1) 언어 분류상 셈어족에 속하는 아랍어는 아라비아 반도를 중심으로 쓰이기 시작해 7세기경 이슬람이 북부 아프리카 및 레반트 지역에 전파됨에 따라 오늘날까지 공용어로 사용되고 있다. 아랍어는 현재 아랍연맹에 가입된 22개국 3억 명에 의해 모국어 또는 공용어로 사용되고 있으며, 전 세계 약 13억에 달하는 무슬림들의 종교어로서, 유엔이 지정한 세계 6대 공용어 중 하나이다.

language)족에 속하는 언어이며, 계통 분류상 남부 셈어 계열이다. 또한 셈어족 가운데 가장 역사가 짧은 언어이면서 원시 셈어와 가장 밀접한 관계를 가지고 있는 언어이다. 셈어는 성경에 나오는 노아의 아들 중 한 명의 이름을 딴 것으로, 아라비아 반도와 아프리카에서 시작되었다는 견해가 현재 가장 지지를 받고 있다. 오늘날 셈어는 중동과 북아프리카에서 주로 쓰이는데, 아라비아 반도와 북부 아프리카 등지에서 쓰이는 아랍어, 에티오피아의 에티오피아어, 이스라엘의 히브리어, 이란과 이라크 및 시리아의 마을룰라 지역 등에서 쓰이는 아람어가 현재 사용되는 셈어이다. 셈어의 공통적인 특징은 쓰기 체계, 단어의 어근, 명사의 성, 동사의 주어 표시, 인후음의 존재 등이다. 특히 발음에 있어서 아랍어는 셈어의 인후음을 모두 보존하고 있다는 특징이 있다.

그러나 엄밀하게 말해서 현재 이집트에서 사용되는 아랍어는 코란의 언어인 '고전 아랍어' 중심의 문어체 아랍어(al-Luġah al-Fuṣḥā)와 화자들이 일상생활에서 사용하는 구어체 방언(al-Lahajāt al-ʿāmmiyyah)들을 통칭하여 일컫는 말이다. 흔히 '표준 아랍어'로 알려진 문어체 아랍어는 전 아랍 세계의 공용어로서 학교 교육을 통해 학습되며, 행정 서류, 뉴스 보도, 코란 낭송 등 극히 제한된 상황에서만 사용된다. 그 외 비격식 상황, 즉 가족 간의 대화나, 시장 상인과의 대화 등에서 아랍인들은 구어체 아랍어를 사용한다.

이러한 아랍어의 사회언어학적 상황은 '양층언어현상(Diglossia)'으로 정의된다. 양층언어현상을 의미하는 '다이글로시아(Diglossia)'[2]는 독일의 동양학자 칼 크룸바허(Karl Krumbacher)가 그의 저서 ≪현대 그리스어 문학 언어 문제(Das Problem Der Modernen Griechischen Schriftsprache)≫(1902)에서 그리스와 아랍의 언어 상황을 다루면서 처음 사용하였다. 그

2) diglossia: 'di' = two, 'glossia' = language, tongue.

후 프랑스 언어학자 윌리엄 마르세이(William Marçais)가 ≪아랍 양층언어(La Diglossie Arabe)≫(1930)에서 아랍 세계의 언어 상황을 '문어체 언어와 구어체 언어의 경합'이라고 설명하면서 최초로 공식적인 용어로 사용하였으며, 이후 퍼거슨(C. Ferguson)이 1959년 논문 '양층언어현상(Diglossia)'을 발표하면서 이 용어가 학술 용어로 본격적으로 사용되기 시작하였다.

이집트는 아랍 국가들 중에서도 아랍어 양층언어현상이 매우 뚜렷이 나타나는 대표적 국가이다. 과거 영국의 식민 지배를 받았던 역사적 배경과 일부 상류층들의 프랑스 유학으로 인해 영어, 프랑스어 혼용 현상이 나타나는 등 독특한 사회언어학적 상황을 보여 주고 있다. 본 소고에서는 이집트의 사회언어학적 특징에 대해 살펴보고, 이집트인들의 언어 사용 유형과 언어 사용에 대한 인식, 언어 정책과 교육 등에 대해 살펴보고자 한다.

2. 이집트의 사회언어학적 특징

2.1. 아랍어 양층언어현상과 언어 변종

오늘날 이집트의 사회언어학적 상황은 '아랍어 양층언어현상(Arabic Diglossia)'으로 특징지어진다. 현대 표준 아랍어로 대표되는 문어체 아랍어가 이집트의 공용어이긴 하지만 이 변종은 마치 외국어처럼 학교에 입학한 후에 인위적으로 교육되는 언어이고 일상 대화에서는 거의 사용되지 않는다. 실생활에서 이집트인들은 자신들이 가정에서 태어나면서 자연스럽게 익힌 구어체 아랍어를 사용하여 의사소통한다.

아랍어의 양층언어현상은 코란의 언어인 고전 아랍어에 근거하는 문어체 아랍어와 다양한 구어체 방언 간의 대립적 관계를 형성했으며 이슬람 이전 시대부터 현재까지 아랍 사회에 존재하고 있다. 흔히 '표준 아랍어'로 간주되는 문어체 아랍어는 예전부터 지금까지 아랍어와 이슬람의 밀접한 관계를 유지

하는 매개체로 간주되고 있다. 반면 구어체 아랍어는 이슬람 전파 이후 정복 활동이 증가함에 따라 고전 아랍어와 정복지의 기층 언어들이 혼합되면서 아랍어의 새로운 변종으로 나타났다. 구어체 아랍어는 계속적인 분화와 변화 과정을 거치면서 문어체 아랍어와는 다른 음운, 어휘, 문법 규칙, 통사적 특징들을 갖게 되었다.

대부분이 무슬림인 아랍인 화자들은 이슬람과 코란의 언어인 문어체 아랍어를 이 세상에서 가장 순수하고 완전무결한 언어라고 신성시하는 경향이 있다. 즉 문어체 아랍어는 아랍인에게 있어 단순한 언어가 아니라 예술적이며 정확한 표현의 수단, 종교의 도구, 문화의 전달 수단, 아랍 민족주의의 기둥으로서의 역할을 하고 있다. 그러나 7세기 등장 이후 이슬람의 신성한 언어라는 명분으로 인해 언어 체계가 거의 변화되지 않았다. 그 때문에 14세기 동안 아랍 사회에 나타난 수많은 변화와 새로운 문물을 표현하기에는 한계가 있다. 따라서 대부분의 아랍인들은 자신들이 태어나면서부터 가정에서 자연스럽게 익히는 구어체 아랍어를 일상생활에서 사용한다.

이집트의 사회언어학적 상황도 마찬가지이다. 이집트 화자들의 발화 형태를 연구해 보면, 상황에 따라 화자들이 사용하는 다양한 언어 변종들이 존재한다. 공식적 상황에서는 주로 문어체 아랍어인 현대 표준 아랍어, 일상생활에서는 구어체 방언들이 사용된다. 퍼거슨의 '다이글로시아' 이론 이후 아랍어 변종들은 많은 사회언어학자들의 연구 대상이 되어 왔다. 그 연구들은 주로 실제 아랍어 화자들이 대화에서 사용하는 변종들을 세부적으로 고찰하거나, 아랍어의 구조를 분석하는 데 중점을 두었다. 그 결과 아랍어의 구조는 고전 아랍어(Classical Arabic) 또는 현대 표준 아랍어(Modern Standard Arabic), 구어체 아랍어 방언(Colloquial Arabic)을 양 극단에 두고 그 중간에 교양 구어체 아랍어(Formal Spoken Arabic)와 같은 다양한 변종들이 연속체를 이루는 구조라고 결론지어졌다.

이집트 지역의 사회언어적 상황을 다층언어현상으로 정의한 대표적 학자

들 중에는 바다위(Badawi, 1973)가 있다. 바다위는 이집트 화자들이 사용하는 아랍어 변종들을 연구하여, 이집트의 언어 상황을 '다이글로시아'로 부르는 대신, '구어체 방언에서부터 고전 아랍어까지의 연속체(Spectroglossia)'로 표현했다. 이 연속체에서 바다위는 이집트인들이 사용하는 아랍어 유형(변종)을 다섯 개의 층위(mustawā)로 분류했고, 이 변종들의 문법적 특징에 대해 설명했다. 그가 분류한 이집트 아랍어 변종의 종류는 다음과 같다.

① 고전 아랍어(Fuṣḥa al-Turāṯ)
② 현대 표준 아랍어(Fuṣḥa al-ʕaṣr)
③ 교양 구어체 아랍어(ʕāmmiyyat al-Muṯaqqafīn)
④ 표준 구어체 아랍어(ʕāmmiyyat al-Mutanawwirīn)
⑤ 문맹인 구어체 아랍어(ʕāmmiyyat al-ʔummiyyīn)

'고전 아랍어'는 코란과 고대 시의 언어이며 고대 아랍 문법학자들에 의해 규정되었고 기술되었다. 이집트에서 이 변종은 단지 종교 분야, 특히 코란 정음학이나 설교, 종교적인 공식 회의석상에서만 사용된다.

'현대 표준 아랍어'는 고전 아랍어보다 많은 기능을 가지고 있으며 라디오, 뉴스, 정치적 연설, 과학적이고 기술적인 문서 등 모든 활동 분야에 관련된다. 또한 고전 아랍어보다 단순화된 문법 형태를 사용한다. 현대 표준 아랍어는 코란의 언어인 고전 아랍어를 근간으로 하여 주로 어휘의 증가와 의미 추가 및 전환과 문체의 차이를 배경으로 이루어진 현대 상류층의 공통어로서 교육과 뉴스 보도 등에서 사용되는 형태이다. '교양 구어체 아랍어'에 대해 바다위는 교육받은 사람들이 사전에 준비된 원고 없이 중대한 토론이나 대학 강의, 라디오와 TV 인터뷰 등에서 사용하는 아랍어 변종이라고 설명하였다. '표준 구어체 아랍어'는 교육받은 사람들이 가족이나 친구와 함께 일상 대화에서 사용하는 것으로 규정된다. '문맹인 구어체 아랍어'는 제시된 이름

이 말해 주듯 라디오 방송이나, 종교 연설 등에서 습득한 것을 제외하고는 정식으로 고전 아랍어 교육을 받지 못한 사람들이 사용한다.

이처럼 바다위는 이집트인들이 사용하는 아랍어 유형을 다섯 층위로 구분하였지만, 이 층위들이 실제로 서로 분명히 구분되어 존재하는 것은 아니며 각 층위 간의 경계선 역시 분명하지 않다. 즉 이집트인들이 대화에서 사용하는 아랍어의 유형은 상황에 따라 문어체 아랍어의 특징들을 보이기도 하고 구어체 방언의 특징을 보이기도 한다. 또는 두 가지 변종이 혼합된 언어 변종의 특징을 보인다.

이 언어 변종들 중 문어체 아랍어의 특징을 보면 다음과 같다.

① 14세기 전에 갖추어진 문법 체계를 거의 그대로 간직하고 있다.
② 단어들은 격, 법, 수, 성, 기타 문법적 기능에 따라 '격변화(al-ʔiʕrāb)'한다.
③ 형태론적으로 명사의 수를 단수, 쌍수, 복수형으로 구별한다. 단수형과 복수형은 다양한 형태소의 추가와 모음의 변화에 의해 변화하고, 여성과 남성의 구별이 있다. 형용사는 수식하는 명사와 성, 수가 일치해야 한다.
④ 문어체 아랍어의 어휘는 매우 풍부하다. 하나의 어근에서 파생되는 많은 명사, 형용사, 동사들이 있기 때문이다.
⑤ 문어체 아랍어는 자연스럽게 익히는 것이 아니라 학교 교육을 통해 교육된다.

문어체 아랍어가 문법적으로 복잡하고 난해한 반면, 이집트인들이 대화에서 사용하는 혼합된 언어 변종의 특징은 다음과 같다.

① 문장 구조에서 구어체 아랍어의 '주어+동사+목적어(SVO)' 어순을 가진다.[3]

3) 문어체 아랍어의 문장 구조는 '동사+주어+목적어'이다.

② 구어체 아랍어의 영향으로 명사의 격 변화는 하지 않는다.
③ 어휘상으로는 문어체 아랍어에 의존하지만 외래어와 구어체 어휘들의 사용에 관대하다.
④ 아랍어와 외국어들, 주로 영어와 프랑스어 사이에 말씨 바꿈 현상(Code-switching)이 일어난다.
⑤ 음운론의 층위에서 자음은 구어체 아랍어에 가깝다.

이처럼 이집트인들은 상황에 따라 다양한 아랍어 변종을 사용하는데, 변종의 선택은 토론의 주제나 화자의 교육 수준, 사회적 지위, 화자와 청자의 성별, 대화 환경, 상대방에 대한 화자의 태도, 심리 상태 등 다양한 요인에 따라 달라질 수 있다. 따라서 외국인으로서 이집트인과의 능숙한 의사소통을 원한다면 문어체 아랍어뿐 아니라 이집트 구어체 아랍어의 특징에 대한 이해가 필요하다. 또한 대중 예술 장르인 영화 속 대사나 대중음악의 가사 역시 대부분 구어체 아랍어로 이루어져 있어 이집트 구어체 아랍어에 대한 지식은 필수적이라 할 수 있다.

2.2. 이집트 구어체 아랍어의 특징

이집트 방언은 사용 지역에 따라 샤르끼야와 서부 델타 방언을 포함하는 '델타 방언', '카이로 방언', 기자와 아스유뜨 지역의 '중부 방언'과 '상 이집트 방언'으로 구분된다.[4] 칼라팔라(A. Khalafallah, 1969)의 경우 이집트 구어체 방언들을 크게 수도 카이로를 중심으로 하여 나일 삼각주에 널리 분포되어 있는 '하 이집트 방언', 그 밖의 지역에서 사용되는 '상 이집트 방언'으로 나누었다. 하 이집트 방언을 중심으로 한 이집트 구어체 아랍어의 대표적 특징을 살펴보면 다음과 같다.

4) '샤르끼야'와 '아스유뜨'는 아랍어 원어 발음에 가장 가까운 한글 표기이다.

1) 음운적 특징

가. 자음

현대 표준 아랍어로 대변되는 문어체 아랍어의 자음은 28자로 구성되어 있다. 이집트 구어체 아랍어의 경우는 외래어의 영향으로 /p/, /v/, /ž/를 포함한 31개의 자음을 갖는다. 이를 압둘 마시흐(1975)에 따라 분류해 보면 다음과 같다.

표 1 문어체 아랍어 자음 체계

	양순음	순치음	치간음	치음	구개음	연구개음	구개수음	인두음	성문음
폐쇄음 /무성	b			t ṭ		k	q		ʔ
폐쇄음 \유성				ḍ	j				
마찰음 /무성		f	ṯ	s ṣ	š		x	h	h
마찰음 \유성			ẓ ḏ	z			ġ	ʕ	
비음	m			n					
전음				r					
공명음	w			l	y				

표 2 이집트 구어체 아랍어 자음 체계

	양순음	순치음	치음	치조음	치조구개음	구개음	연구개음	후연구개음	구개수음	인두음	성문음
폐쇄음 /무성	p		t	ṭ			k		q		ʔ
폐쇄음 \유성	b		d	ḍ			g				
마찰음 /무성		f	s	ṣ	š			x		ḥ	h
마찰음 \유성		v	z	ẓ	ž			ġ		ʕ	
비음\유성	m		n								
측음\유성			l	ḷ							
전동음\유성				r ṛ							
반모음\유성	w					y					

자음 /j/는 아랍 국가들 내에서 그 조음 방법이 상당한 차이를 보인다. 코란 낭송 등의 고전적 낭독에서는 /j/로 발음되지만 지역에 따라 /ž/, /g/ 등으로 실현되어 발음되는 방식이 다르다. 이집트 구어체 아랍어에서 /j/는 대부분 /g/로 발음된다.

예 1

문어체 아랍어	이집트 구어체 아랍어	의미
/jadīd/	/gedīd/	'새로운'
/jabal/	/gabal/	'산'

또한 /q/ 역시 /al-qurʔān/ '코란', /al-qāhirah/ '카이로' 등의 단어를 제외하고 거의 /ʔ/로 발음된다. 예를 들어 /qāla/ '그는 말했다'는 /ʔāl/로 발음된다.

예 2

문어체 아랍어	이집트 구어체 아랍어	의미
/sūq/	/sūʔ/	'시장'
/qalam/	/ʔalam/	'연필'

문어체 아랍어에서의 무성 치간 마찰음 /t̲/와 유성 치간 마찰음 /d̲/는 구어체 아랍어에서는 각각 /t/, /s/ 및 /d/, /z/로 발음된다.

예 3

문어체 아랍어	이집트 구어체 아랍어	의미
/t̲alāt̲ah/	/talāta/ 또는 /salāsa/	'3'
/d̲ahab/	/dahab/ 또는 /zahab/	'금'

나. 모음

문어체 아랍어의 모음 체계를 보면 3개의 단모음 /a/, /i/, /u/와 3개의 장모음 /ā/, /ī/, /ū/, 2개의 이중모음 /ay/, /aw/이 있다. 이집트 구어체 아랍어의 모음 음소는 문어체 아랍어의 모음 음소 외에 4개의 다른 모음 음소인 /o/, /e/, /ō/, /ē/가 있다.

예 4

문어체 아랍어	이집트 구어체 아랍어	의미
/bayt/	/bēt/	'집'
/yawm/	/yōm/	'날' 또는 '하루'

2) 형태, 통사, 어휘적 특징

이 외에도 이집트 구어체 아랍어는 문어체 아랍어와 비교할 때 음운, 형태, 통사, 어휘 층위에서 많은 차이점을 보이는데 주요 예를 보면 다음과 같다.

예 5

문어체 아랍어	이집트 구어체 아랍어	의미
/ṣabāḥ al-xayr/	/ṣabāḥ il-xēr/	'좋은 아침입니다'
/kayfa ḥāluka?/	/ʔizzayak?/	'어떻게 지내세요?'
/bi-xayr/	/kuweys/	'잘 지냅니다'
/ʔurīdu ʔan ʔaʃraba al-qahwah/	/ʔana ʕāwiz ʃrab ʔahwe/	'나는 커피를 마시고 싶어요'
/ʔanā lastu ṭālib /	/ʔana miʃ ṭālib/	'나는 학생이 아닙니다'

동사 부정에서도 많은 차이를 볼 수 있다. 완료동사 부정의 경우, 문어체 아랍어는 '/lam+미완료 단축법 동사/' 구조로 표현되는 반면, 이집트 구어체 아랍어에서는 동사의 기본 형태 앞에 /ma/를, 뒤에 /š/를 붙여 부정 형태를 만든다.

예 6

문어체 아랍어	이집트 구어체 아랍어	의미
/lam yaktub/	/maktabš/	'그는 쓰지 않았다'

미완료동사 부정의 경우 문어체 아랍어는 '/lā+미완료 직설법 동사/' 구조로 표현되는 반면, 이집트 구어체 아랍어에서는 완료동사와 마찬가지로 /ma ~ š/를 붙여 부정 형태를 만든다.

예 7

문어체 아랍어	이집트 구어체 아랍어	의미
/huwa lā yašrabu/	/huwwa mayišrabš/	'그는 마시지 않는다'

2.3. 이집트인들의 언어 사용에 대한 인식

필자는 수년 전 이집트 카이로 대학교를 방문하여 재학생과 학교 관계자들을 대상으로 설문 조사[5]를 실시한 바 있다. 설문 문항은 아랍어 양층언어현상에 대한 이집트인들의 인식과 방언 선호도, 향후 전망 등에 대한 10개의 항목으로 구성되었다. 이 설문 결과에 따르면 전체 응답자 중 80% 이상이 이집트 사회 내 아랍어 양층언어현상을 잘 인지하고 있다. 또한 응답자 대부분이 일상생활에서 문어체 아랍어가 아닌 구어체 아랍어를 사용한다고 대답했다. 그러나 약 63%가 문어체 아랍어의 사용이 교육 수준이나 사회적 지위의 기준이 된다고 답했다. 일반적으로 화자들은 문법도 단순하고 보다 쉽게 발음할 수 있는 구어체 방언을 선호하는 언어적 습관이 있지만, 교육 수준이나 사회적 지위와 연관된 부분에서는 의식적으로 문어체 아랍어를 선호하는 경향을 보인다. 또한 이집트의 경우 여성보다는 남성들이 문어체 아랍어를

5) 설문에 참여한 인원은 모두 178명으로 성별 분포는 남성 96명, 여성 82명이다. 응답자의 연령 분포는 18세~25세가 65%, 25세 이상이 35%를 차지했다.

사회적 지위의 척도로 여기는 경향이 있다. 문어체 아랍어를 사회적 지위의 척도로 간주하고 있긴 하지만 실제 생활에서는 구어체 아랍어 방언의 사용을 선호하는 아이러니를 보인다. 일상 대화에서 문어체 아랍어인 표준 아랍어를 사용하는 화자는 거의 없고 구어체 방언이나 두 변종의 특징이 혼합된 언어 변종이 사용된다. 이집트 언어사회의 개방 수준에 대한 화자들의 견해를 묻는 항목에서는 전체 응답자들의 96.6%가 "매우 개방적이다."라고 대답했다. 이 결과는 이집트가 아랍 세계에서 정치, 경제, 문화적으로 중심지 역할을 하면서 서구 문명과의 각종 교류를 통해 다른 아랍 국가들에 비해 상당히 개방적인 사회 분위기를 유지하고 있는 것과 관계 깊은 것으로 분석할 수 있다.

또한 최근 인터넷과 소셜 네트워크 서비스(SNS) 같은 뉴미디어의 출현은 이집트 젊은이들의 언어 사용 유형과 인식에도 영향을 미치고 있다. 많은 이집트 젊은이들은 소셜 네트워크 서비스상의 빠른 의사소통과 문자 전송을 위해 새로운 형태의 아랍어를 사용하고 있다. 이것은 라틴 문자와 숫자로 구성되어 있어 심지어 같은 아랍어 화자에게도 완전히 새로운 언어처럼 보인다. 바로 '아라비시(Arabish)', '아라비지(Arabizi)', '프랑코 아라빅(Franco-Arabiac)' 등으로 알려진 언어 변종이다.[6] 아라비시는 아랍어(Arabic)와 영어(English)의 합성어로 이 두 언어 사이에서 만들어진 새로운 언어를 뜻하는 신조어이다. 기존의 아랍인들은 서식에서 문어체 아랍어, 일상 대화에서 구어체 아랍어를 사용하였던 반면 최근 일부 젊은이들은 인터넷과 소셜 네트워크 서비스에서 라틴 문자를 이용한 구어체 아랍어로 빠르게 의사소통할 수 있는 아라비시를 사용하고 있는 것이다.

아라비시는 현대 이집트 사회의 새로운 언어 현상이다. 아라비시가 확산된 시기는 새로운 기술이 보급된 1999년대 중반으로 거슬러 올라간다. 정보

6) 이하 '아라비시'라고 한다.

통신 기술력은 세계를 하나의 지구촌으로 만들었고, 연결 고리가 없던 전혀 다른 부류나 국적의 사람들을 하나의 공론장으로 모이도록 했다. 특히 휴대전화, 스마트폰은 아랍 세계에서 아라비시를 만들어 낸 가장 결정적인 요인으로 볼 수 있다. 휴대전화가 처음 아랍 세계에 소개됐을 때 자판은 영문자로만 되어 있었고, 이를 사용할 수 있는 계층도 영어로 의사소통할 수 있는 계층이었다. 이후 휴대전화나 스마트폰의 가격이 저렴해지고 점차 대중에게 보급되면서 아랍 대중들은 자연스럽게 영문 자판을 이용했고, 영어식 표기의 경제성을 체감하게 된다. 바로 이 지점에서 아라비시가 탄생했다고 볼 수 있다. 그 이후 아랍어 자판의 휴대전화나 이를 보조할 수 있는 아랍어 키보드가 보급되었지만, 일부 아랍인들은 대화에서 구어체 아랍어를 선호하는 것처럼, 스마트폰을 이용한 채팅에서 여전히 라틴 문자를 이용한 대화 방식을 선호하고 있다.

이집트의 경우 영어 교육이 확산되면서 이러한 경향은 점차 두드러진다. 영어 사용 여부를 고등 교육의 여부로 연결 짓는 세계적 분위기에 편승해 교육 현장에서 영어 과목 또는 아예 영어로 진행되는 수업을 대폭 늘린 것이다. 심지어 영어에 능통한 경우 취업에서 더 유리한 위치를 차지하기 때문에 외적인 세계의 흐름과 내부적인 동기가 만나 영어의 중요성은 점차 높아졌다. 이렇게 변화한 교육 환경에서 교육받은 젊은이들은 아라비시에 익숙해졌고, 이를 사용하는 데에 별다른 거부감을 느끼지 않는 것으로 보인다.

3. 이집트의 언어 정책과 교육

3.1. 이슬람 초기 시대 이집트의 언어 정책과 교육

이집트에서의 아랍어 교육은 아라비아 반도에서 이슬람 제국이 발흥된 후인 약 640년께, 이슬람 제국의 침공으로, 이집트 지역이 아랍 치하에 들어

서면서부터 시작되었다. 정통 칼리프 시대의 제2대 칼리프 오마르 집권기 때 장군 아므르는 카이로로 진군해 비잔틴의 바빌론 요새를 함락시키고, 같은 해 11월 당시 수도인 알렉산드리아를 함락시켜 이집트 지역을 이슬람 움마(al-ʔummah, 공동체)의 한 속주로 만들었다. 따라서 이집트인들은 이슬람 제국의 공식어였던 아랍어를 받아들이게 되었으며, 초기 아랍어 교육은 코란을 정확히 읽고 암송하고자 하는 이슬람 개종자들을 위해 주로 모스크에서 이루어졌다. 더불어 칼리프 오마르는 "아랍어가 마음을 정화하고, 품격을 향상시킨다."라는 명목으로 새로운 영토에 무슬림들을 대상으로 아랍어 교육을 지시하였으며, 아랍어 교사를 이집트로 보내기도 하였다. 이 당시의 아랍어 교육 내용은, 코란과 사도 무함마드 언행록인 ≪하디스(al-Hadith)≫, 일부 아랍 유목민의 '정형 장시(al-Qasīdah)' 등 상당히 제한적이었으며, 그 목적도 완전한 하나의 언어로서의 아랍어 습득이 아닌 단순히 종교적인 언어 학습으로 국한되었다.

3.2. 중세 이집트의 언어 정책과 교육

압바스 왕조의 제7대 칼리프 알 마으문은 아랍어 전문 기관인 '지혜의 집(Bayt al-Ḥikmah)'을 설립하였다. 이 기관은 도서관의 기능을 담당한 한편 번역 운동의 중심 기관이 되어 다양한 학문 분야의 많은 서적을 아랍어로 번역하였으며, 이슬람 제국의 황금기를 이끌었다. 이때 아랍어 교육과 관련하여 '쿠파 학파', '바스라 학파', '바그다드 학파'와 같은 아랍어 문법 학파들이 '지혜의 집'을 통해 국가 정책적으로 지원을 받으며, 본격적인 아랍어 연구를 시작하였다. 이러한 아랍어 연구의 취지는 비아랍인 이슬람 개종자들의 급증과 더불어 넓어진 영토의 다양한 외래어로부터 코란의 언어인 아랍어를 보호하고 코란을 정확하게 전승하며 이해시키기 위한 것이었다. 이후 지혜의 집은 아랍 사회에서 나타난 대부분의 아랍어 교육 및 연구

기관들의 기준과 본보기가 된다.

이집트 지역에서는 약 909년 시아파 무슬림들에 의해 파티마 왕조가 설립되고 공용어로 아랍어와 아랍어 문자가 지정된다. 하지만 파티마 왕조 역시 기층민들을 대상으로 유화 정책을 시행하였고 이에 따라 이집트 지역의 무슬림 수는 서서히 증가하였다. 아랍어는 무슬림들을 대상으로 한 코란 읽기와 암송을 위한 종교적인 언어의 학습으로 그 교육 범위가 국한되었다. 그러나 일부 이집트인들은 왕조 건국 전후로 바스라, 쿠파, 바그다드 등으로 유학을 다녀오거나 혹은 현지 학자들이 직접 이집트로 와서 이집트인들에게 아랍어 교육을 수행하였다.

초기 이집트 아랍어 학자들의 기반은 당시 파티마 왕조의 이슬람 교육이라는 명목하의 아랍어 지원 정책에 힘입어, 약 982년 알 아즈하르(al-Azhar) 대학교가 설립되는 결과물을 맺게 된다. '가장 빛나는 꽃들'이라는 의미의 알 아즈하르 대학교는 알 아즈하르 모스크(약 972년 설립)와 함께 이집트인들에게 전문적인 아랍어 교육을 실시하고 아랍어 교사 양성을 담당했다. 이후 알 아즈하르 대학교는 세계에서 가장 오래된 대학교 중 하나로, 오랜 기간 동안 아랍 사회에서 이슬람과 아랍어 교육의 중심지 역할을 하였으며 아직까지도 그 명성이 굳건하다.

12세기 이후 이집트는 아이유브 왕조를 거쳐 맘루크 왕조가 들어서며, 지중해 상권과 팔레스타인, 시리아 일대까지 장악한 실질적인 이슬람 종주국으로 발돋움하게 된다. 몇몇 왕조를 거치며 시아에서 순니로 이슬람의 내부 교리는 바뀌었지만 여전히 이슬람 국가로서 아랍어를 공용어로 사용하였다. 이후부터는 대부분의 기층민이 무슬림이 되며 아랍어 교육이 급격히 증가하였다. 일부 무슬림의 종교적인 언어 혹은 지배층의 지배 언어를 넘어, 대다수의 이집트인들을 위한 보편적인 아랍어 교육이 이루어졌고, 이때부터 이집트의 기층 언어에 아랍어가 큰 영향을 미치게 된다.

3.3. 근대 이집트의 언어 정책과 교육

이후 이집트는 화약 무기로 무장한 오스만 제국과 나폴레옹에 의해 차례로 정복당하며 쇠약해져, 제1차세계대전 이후 약 1882년부터 영국에 의해 식민 지배를 받게 된다. 영국은 영어를 공용어로 지정하고, 언어 교육 기관을 비롯한 공공 교육 기관의 문을 닫았으며, 아랍어 교육을 전면 금지했다. 그리고 아랍어 대신 특정 상위 계층을 위한 영어와 프랑스어 교육을 중점적으로 시행하였다. 이러한 식민 지배 상황에서의 아랍어 교육 및 사용 금지 정책은 당시 급격한 현대화와 더불어 새로운 인문학적 관념이 들어오는 시대적 상황 속에서 아랍어의 발전을 저해하였으며, 현대적 변화의 기회를 앗아갔다. 하지만 이러한 상황 속에서도 코란의 언어인 아랍어에 대한 이집트인들의 애정과 열정은 식지 않았다. 특히 알 아즈하르 대학교는 영국의 눈을 피해, 아랍어가 실제 살아 있는 언어이며 모든 이집트 대중들에게 교육되어야 한다고 제창하며, 아랍어 교육의 발전을 위한 선도적 역할을 담당했다. 1906년에는 이집트 시민들에 의해 '모든 이집트인에 대한 아랍어 교육'을 모토로 한 카이로 대학이 설립되었다. 영국의 아랍어 교육 탄압 정책 속에서, 이와 같이 아랍어를 유일한 공용어 및 모국어로서 지켜내려고 했던 이집트인들은, 다른 아랍 국가로 넘어가 아랍어를 배운 뒤 귀국하여 아랍어 교사가 되는 등 각고의 노력까지도 기울였다.

더불어, 아랍 사회가 당대 직면한 정치, 교육, 사회적 문제와 종교, 민족주의 등에 관심을 갖고 있었던 선구적 개혁가들에 의해, 아랍어 문제보다 더 심각한 것은 없다는 공동의 취지 아래 '아랍어 부흥 운동'이 일어났다. 이 운동은 아랍어 공부를 위해 카이로로 유학 온 예멘 등의 학자들로부터 시작되어 이후 이집트인들에 의해 이집트를 중심으로 펼쳐졌다.

이러한 이집트인들의 노력은 영국으로부터 부분적인 독립을 이루기 시작하는 1930년대부터 비로소 결실을 맺으면서 아랍어는 국가 공용어로서의

위치를 되찾게 된다. 1930년대부터는 국가의 행정과 법정에서 아랍어가 다시 사용되기 시작하였고, 1932년에는 푸아드 1세에 의해 왕립 아랍어 학술원(The Royal Academy of Arabic Language)이 설립되어 아랍어의 현대화, 문법의 개선, 언어의 단순화를 위한 작업이 수행된다. 이후 1940년대부터 초등학교에서 대학교까지 아랍어는 모든 교과 과정에서 상당한 비중의 과목이 되었으며, 나아가 외국 기관에서도 아랍어의 교육은 정책적 의무 사항이 된다.

3.4. 현대 이집트의 언어 정책과 교육

이집트는 앞서 언급한 알 아즈하르 대학교 및 아랍어 교사들의 개인적인 노력, 그리고 아랍어 부흥 운동 등에 의한 탄탄한 인프라를 통해 19세기 말부터 아랍 사회에서 아랍어 현대화를 주도하였다.

1923년 이집트 왕국 독립 시에 제정한 헌법으로 초등 교육의 무상 의무화가 규정되었다. 제2차세계대전 후 근대화 노력과 교육 개혁이 진행되었는데, 특히 1951년에는 법률로써 국가적 교육 체제의 정비를 갖추었으며, 1953년에는 교육법으로써 학교 제도를 고쳐 직업·기술 교육의 진흥에 기여했다. 1956년, 교육법은 현행 교육 제도의 기초로서 학교 제도의 근대화와 초등 교육 의무제를 규정하고 6세에 초등 교육을 시작하도록 하였다.

특히 1952년 혁명 이후, 이집트 대통령에 취임한 나세르는 모든 이집트인에 대한 무상 교육 정책을 실시하였고, 초등 교육을 의무화하였으며, 이때부터 거의 대부분의 이집트인들이 아랍어 교육을 접할 수 있게 되었다. 그는 1950년대와 1960년대를 거치는 동안 교육 과정에서 문어체 아랍어 교육(이슬람과 문학 등)의 비중을 극도로 증가시켜, 젊은 세대들에게 서구에 의해 훼손된 아랍어를 재교육하고, 이집트의 정체성과 국민성을 지켜 내고자 하였다. 더불어 그는 다른 아랍 국가들의 독립을 지원하며 아랍 국가들의 협력 증진을

도모하였는데, 이때 아랍 국가들의 구심점을 아랍어로 간주하기도 하였다. 또한 1960년 9월 28일 "우리는 우리가 하나의 아랍 공동체임을 믿습니다. 아랍 국가들은 항상 언어를 통해 연합되어 왔습니다. 언어적 통합이 사고의 통합입니다."라는 연설을 통해, 아랍 사회에 아랍어 교육 및 정책의 증진을 호소했다. 이러한 나세르 대통령의 정책적 지원에 힘입어 아랍어 학술원은 아랍어 현대화 작업의 일환으로, 인문학 및 과학적 분야에서 아랍어 전문 용어화를 이룩하였고, 1960년대부터는 대학교 전문 과학 분야에서 아랍어로 강의를 하는 것이 가능해졌다. 이에 1960년 이집트의 당시 교육부 장관 카말 알딘 후세인(Kamāl al-Dīn Ḥusayn)은 아랍어 학술원을 통해 대학교 교육에서 오직 문어체 아랍어만 사용하도록 지시하기도 하였다.

그러나 1970년대 이집트가 이스라엘과 평화 협정을 맺으면서, 이집트는 아랍 연맹국들로부터 정치적으로 배척을 당하게 된다. 당시 이집트인들의 아랍인으로서 정체성은 상당히 떨어지게 되었고, 이러한 정체성의 일시적 혼란은 이집트의 언어 정책에도 영향을 미쳤다. 1980년대부터는 국가 경제 발전과 지역 사회에 실효성이 높은 영어 등의 외국어 과목과 공학 등의 교육 비중이 더욱 높아졌다. 더불어 이집트의 구어체 아랍어에서 나타나는 'Qāf(ق)'의 함자 발음 현상 등의 구어체적 특성도 그들의 국민성에 수렴되어, 문어체 아랍어 교육은 위기에 직면하게 된다.

이집트의 교육 분야는 1990년대 초부터 다시 급격히 성장했다. 교육행정은 지방 분권화되어 이집트 문부성(文部省)은 실질적인 권한과 책임을 지방 교육 행정 당국에 위임하였다. 학교 체제는 6년제의 초등학교, 3년제의 하급 및 상급 중등학교, 고등 교육 기관, 대학교 등으로 되어 있다. 초급과 중등 교육 기관은 공립 학교와 사립 학교로 나뉜다.

공립 학교의 기본 교육 과정은 모두 문어체 아랍어로 이루어진다. 영어 교육은 4학년부터 시작되며, 프랑스어는 고등학교 과정에서 제2외국어로 추가된다. 일부 외국인을 위한 국제 학교에서는 교육 과정(과학, 수학, 컴퓨터)

대부분을 영어로 가르치고 예비 교육에서 제2외국어로 프랑스어를 교육하며, 2004년부터는 중국어가 제2외국어 과목으로 추가되었다.[7]

이집트의 2005년 문해율은 남성 83%, 여성 59%, 전체 평균 71%이다. 정부와 관련 기관들은 교육에서 남녀 차별을 줄이고, '전 국민을 대상으로 한 초등 교육 실시'라는 '2015년 밀레니엄개발목표(MDG)'를 달성하고자 노력하고 있다.

이처럼 이집트 정부는 공교육을 통한 문어체 아랍어의 교육을 위해 힘쓰고 있으며 구어체 아랍어와 외래어 사용을 줄이기 위해 노력하고 있다. 또한 아랍어를 헌법의 일부로 만들어서 예술, 문학, 과학 영역에서 사용하려고 노력하고 있으며, 이러한 언어 정책 입안과 실행에서 주된 역할을 담당하는 기관은 이집트 아랍어 학술원이다.

아랍어 학술원의 명칭은 원래 '왕립 아랍어 학술원'이었으나 1952년 나세르 혁명 이후 현재의 명칭인 '이집트 아랍어 학술원(Majmaʿ al-Luġah al-ʿarabiyyah bi Miṣr)'으로 바뀌었다. 아랍어 학술원은 원장, 부원장, 사무국장, 재무·행정 담당 부장, 사전 및 고대 문헌 복구 담당 부장, 기록 및 발행, 문화 사업 담당 부장, 총무 부장 등 아랍어 분야 저명 학자들로 구성된 임원들과 각 부서별 수십 명의 직원들로 이루어져 있다.

아랍어 학술원은 문어체 아랍어를 범국가적 언어로 유지하기 위한 최종 목표를 달성하기 위해 아랍 고유의 현대 과학, 문학, 예술을 제공하고 있다. 또한 아랍어의 완전성을 지원하며, 현대 사회와 미래에도 아랍어가 일상생활에서 가장 편리한 언어로 사용되도록 하는데 목적을 두고 있다.

1982년 대통령은 학술원의 연구를 체계화할 수 있는 새로운 법률을 공표하였는데, 이로써 아랍어 학술원은 아랍어와 관련된 모든 학문적 주제들

---

7) 2004년, 이집트 초·중등학교에서 중국어를 제2외국어로 채택했다고 이집트 교육부가 발표했다. 이집트와 중국이 아랍어와 중국어를 교환 교육하고 커뮤니케이션. 멀티미디어. 소프트웨어 기술 분야에서 상호 협력기로 협정을 체결한 데 따른 것이다.

이외에도 재정 및 행정 업무에 대해 전적으로 책임을 지니게 되었다. 이 법률에 따르면 이집트 정부와 아랍어 학술원은 다음과 같이 학술원의 설립 목표를 정하였다.

① 문어체 아랍어의 보존 및 현대 과학, 예술 분야 발달 변화에 부응하는 언어 발전
② 아랍어의 기원, 현대 변종 연구, 아랍어 문법, 형태론 규칙 용이화 및 교육의 보급, 아랍어 서식 용이화 연구
③ 현대 문명, 과학, 문학, 예술 분야 신조어 연구 및 외국 인명 연구, 관련 표기법 통일
④ 대학교에서의 아랍어 교육 진흥 및 교재의 아랍어화
⑤ 아랍어 발전 관련 제 분야 연구 및 보급을 위한 사업 추진
⑥ 상기 목적과 관련된 제반 연구

그리고 새로운 법률에서는 이러한 목표들을 달성하기 위한 아랍어 학술원의 세부 시행 사항을 다음과 같이 규정하고 있다.

① 현대화되고 교정 편집된 어휘집과 전문 분야, 일반 과학 어휘집 및 사전 편찬
② 언어학적으로 적용 가능한 용어들과 그 어원, 그리고 피해야 할 언어 구조들의 연구 및 분석
③ 언어, 문학, 예술 및 제반 지식 분야 아랍 유산의 부흥에 기여
④ 학문적 연구에 도움이 되는 고대 및 현대 아랍어 방언에 대해 학문적으로 연구
⑤ 문학 비평 및 이슈들에 관한 연구, 시상

⑥ 기타 관련 세미나와 경연 대회 개최를 통한 문어체 아랍어 및 아랍 문학 작업 장려
⑦ 학술원 활동이나 업무 등과 같은 학술원 결의안과 학술지, 간행물, 회보, 혹은 책자 등의 발간
⑧ 아랍어 용이화 및 용어의 표준화 등에 관련된 학술원의 배포물이 유익한 결과를 산출할 수 있도록 관할 단체에 필요한 조치 권고
⑨ 학술원 활동과 관련된 학술 대회와 세미나 개최 및 참석
⑩ 이집트 국내·외에 있는 아랍어 및 과학 학술원과 관계 당국 간의 연계 강화

이처럼 84년의 긴 역사를 자랑하는 이집트 아랍어 학술원은 정부의 지원하에 문어체 아랍어의 보존과 현대화에 힘쓰고, 급변하고 있는 과학, 지식 및 정보 기술(IT) 발전 등 시대 조류에 부응하여 아랍어를 개선하고 발전시키고자 노력하고 있다.

그러나 이러한 노력에도 불구하고 현재 이집트는 오랜 기간 정세 불안과 경제 불황이 이어지고 있고 경제, 정치적 불안정은 교육 분야에도 영향을 미치고 있다. 불안정한 국가 체제, 극심한 빈부 격차 등과 수준 미달의 공교육에 대한 국민들의 신뢰는 매우 저조한 상태이다. 이에 사교육의 비중이 늘어나고, 사립학교의 수도 꾸준히 증가하여 2010년 이집트 내의 사립학교 수는 5,000개를 넘어섰다. 사립학교에서는 영어와 제2외국어 교육의 비중이 높아지고 있다. 정부에서는 과외와 같은 사교육을 불법이라 규정하였음에도 불구하고, 현재 약 70% 이상의 학생들이 방과 후 사교육을 통한 보충 수업을 은연중에 실시하고 있다. 이는 이집트 국제 기업들의 외국어 대비 낮은 아랍어 선호 현상과 더불어 이집트의 아랍어 교육을 침체시키고 있는 요인이기도 하다.

4. 맺는 말

오늘날 이집트의 사회 언어학적 상황은 '아랍어 양층언어현상(Arabic Diglossia)'으로 특징지을 수 있다. 현대 표준 아랍어로 대표되는 문어체 아랍어가 이집트의 공용어이긴 하지만 이 변종은 7세기 이후 문법 체계가 거의 변하지 않은 언어로서 학교에 입학한 후에 인위적으로 교육되는 언어이고, 실생활에서 이집트인들은 자신들이 가정에서 태어나면서 자연스럽게 익힌 구어체 아랍어를 사용하여 의사소통한다.

이집트 사회에서 아랍어 양층언어현상은 앞으로도 계속 유지될 것이다. 문어체 아랍어는 이슬람과 코란의 언어이므로 언어 지도에서 사라지지 않을 것이다. 그러나 문어체 아랍어는 종교, 뉴스 보도, 서류 작성 등 아주 제한적인 상황에서만 사용될 것이고, 일상 대화에서는 구어체 아랍어가 계속 사용될 것으로 전망된다. 이러한 아랍어 양층언어현상의 간극을 좁히기 위해서는 이집트 정부와 아랍어 학술원, 교육 기관들의 아랍어 교육 및 보급, 대중들이 보다 더 쉽게 문어체 아랍어를 익히고 사용할 수 있도록 언어 체계의 용이화를 위한 적극적인 노력이 필요하다.

그러나 이집트를 비롯한 아랍 국가들의 아랍어 교육과 정책은 단순한 언어 교육적 차원을 넘어 이슬람과 결부된 국가 정체성과 정부의 정치 지향성 등 여러 가지 요인으로 결정되는 상당히 복합적인 문제이다. 또한 아랍어 정책 방향과는 별도로 글로벌 시대 국가 경쟁력 증대를 위한 외국어 학습 증가에 따라, 외국어 사용과 교육의 비중이 증대되고 있어, 이에 따른 언어 정책의 유연성이 필요한 것으로 보인다. 아울러 이집트의 경우 오랜 경제 불황과 '아랍의 봄(Arab Spring)' 이후 전반적인 과도기 상황으로 인해, 현재 효율적인 언어 정책에 대한 논의는 소강상태에 있는 것이 사실이다.

조속한 시간 내에 이집트의 상황이 안정되어 아랍어 교육과 언어 정책에 대한 심도 있는 논의가 재개되길 바란다. 이집트의 언어 정책은 아랍어 양층

언어현상이라는 사회언어학적 특징이 우선적으로 고려되어야 할 것이며, 또한 정부, 아랍어 학술원, 교육 기관 간의 소통을 통해 사회 변화에 부응하는 보다 더 합리적이고 효율적인 합의점 구축이 필요한 것으로 보인다.

참고 문헌

국립국어원(2010), ≪국제 언어 정책 비교 연구≫.

서정민(2009), Education and Development in Egypt, ≪중동연구≫, Vol.28(1).

오명근(1996), ≪아랍어 구문어체 비교론≫, 한국외국어대학교 출판부.

_____(2008), 현대 이집트 사회의 언어상황에 관한 연구 - 아랍어 사용 유형과 교육을 중심으로, ≪외국어교육연구≫, 22(1).

_____(2013), 구어체 아랍어 음운 특성과 교육에 관한 연구 - 이집트 구어체 아랍어를 중심으로, ≪외국어교육연구≫, 27(1).

윤은경(2002), 이집트 방송 아랍어의 사회언어학적 연구, 한국외국어대학교 박사 학위 논문.

_____(2003), 교양구어체 아랍어의 개념과 특징 연구, ≪아랍어와 아랍 문학≫, 제7집 1호, 한국 아랍어·아랍 문학회.

_____(2006), 이집트 지역의 사회언어학적 특징에 대한 연구, ≪아랍어와 아랍 문학≫, 제9집 1호, 한국 아랍어·아랍 문학회.

21세기 중동이슬람문명권 연구사업단(2006), ≪중동언어의 이해 3≫, 한울.

Abou-Seida, A. M.(1971), Diglossia in Egyptian Arabic: Prolegomena to a Pan-Arabic Sociolinguistic Study, *Doctoral Dissertation*, University of Texas, Austin.

Al-Toma, S. J.(1974), Language Education in Arab countries and the Role of the Academics in J. Fishman, ed., *Advances in Language Planning*. The Hague: Mouton.

Anwar, G. C.(1969), *The Arabic Language - Its Role in History,* University of Minnesota Press, Minnesota.

Badawi, S. M.(1973), *Mustawayāt al-ʿarabiyyah al-muʿāṣirah fi Miṣr*, Cairo: Dār al-Maʿārif bi Miṣr.

Ferguson, C. A.(1959), Diglossia, *Word,* 15.

_____(1991), Diglossia Revisited, *The Southwest Journal of Linguistics*. 10(1).

Shraybom, S. S.(1999), Language and political change in modern Egypt, *International Journal of the Sociology of Language* 137.

Khalafallah, A. A.(1969). *A Descriptive Grammar of Saeiːdi Egyptian*

Colloquial Arabic, Paris, Mouton.
Versteegh, K.(1997). *The Arabic language*, Edinburgh, Edinburgh University Press.
Zughoul, M. R.(1980), Diglossia in Arabic: Investigating Solutions, *Anthropological Linguistics,* 22(5).
http://en.wikipedia.org/wiki/Egypt/.
www.Arabicacademy.org.eg/.

선청어문(先淸語文)의 국어교육자, 난대(蘭臺) 이응백(李應百) 선생

민현식
서울대학교 국어교육과 교수

Ⅰ. 난대(蘭臺)의 생애

난대(蘭臺) 이응백(李應百) 선생은 국어교육계의 큰 스승으로서 국어교육계에 남기신 자취가 은은하고 거대하다. 제자들의 글에서는 '만인의 큰형(김상준)', '탁월한 천품을 타고나신 천생적 교육자요 사명적 인간인 스승(이명권)', '교사의 신언서판(身言書判)을 강조한 스승(유광렬)', '국어교육의 산 역사(윤의순)', '출천(出天)의 스승(김봉군)', '그립고 그리운 선생님(이주행)', '평생 정도를 지키셨던 분(김유중)' 등으로 추모된다.

1923년 4월 19일에 경기도 파주군 파평면 덕천리에서 상화(相和) 공과 양주(楊州) 윤씨(尹氏)의 2남 1녀 중 막내로 태어나 2010년 3월 29일 새벽 4시 40분에 향년 87세를 일기로 영면하였다. 어려서 서당을 다니며 한문과 일어를 익혔고 고소설, 신소설을 두루 읽었다. 파주 적성(積城) 공립소학교 졸업(1939), 경성사범학교 예과 5년 수료(1944) 후 본과 2년 때 광복을 맞았다. 경성제대, 경성공업전문과 함께 3대 명문인 경성사범은 100명 정원에 70명은 일본인, 30명은 조선인으로 그 30명에 전국 수재들이 뽑혔다. 경성사범이 국립서울대로 개편된 후 국어과를 1949년 7월에 졸업하였다.

졸업 후 서울중학교, 서울사대부고, 이화여대를 거쳐 1957년 4월부터

1988년 8월 정년까지 서울대 국어교육과에서 국어교육의 사표(師表)로서 학문 정립에 힘썼다. 일제하 교육과 미군정 교수요목기, 1~4차 교육과정기까지 국어교육의 산증인이요 역사이었고, 국어심의회 표준어 심의위원으로서 '표준어 규정'(1988) 해설 집필을 하는 등 국어 규범 확립에 기여하였다. 바른 국어교육이 바른 생각을 하는 국민을 길러낸다는 철학으로 먼저 어문생활부터 깨끗해야 한다는 '선청어문(先淸語文)'의 정신으로 살아온 삶은 오늘의 국어교육자들에게 일깨우는 바가 크다.

Ⅱ. 국어교육학의 학문적 확립

1955년 창립해 큰 숲으로 일군 '한국어교육학회'는 ≪국어교육≫ 132호(2010)를 난대 추모 특집호로 꾸며 보은하였는데 거기서 필자가 밝힌 난대의 학문적 특징은 다음과 같다.[1)]

(1) 규범문법과 어문규범 연구

초기에는 국어학 연구를 기반으로 규범문법 확립에 힘썼다. 40대 초에 낸 첫 교과서가 16종 교과서 중의 하나인 ≪중학 문법≫(남광우·이응백·유창돈 공저, 1966)이었고, 1979년에는 5종 문법 중 ≪고교 문법≫(이응백·안병희 공저, 1979)을 저술하였다. 그리고 150여 편의 학술논문은 어문규범 제정, 어휘교육, 한자교육 분야에 기여하였다.

규범에서는 표기법과 표준어 연구에 기여하여 ≪한글 맞춤법 사전≫(1961)을 냈고 그해부터 국어심의회 위원으로, 1970년부터는 '국어 조사 연구위원회'의 연구위원으로 참여하고 1988년에 공표된 '표준어 규정' 제정

1) "온고이지신(溫故而知新)의 국어교육학을 위하여: 국어교육 학문 정립의 선구자 난대 이응백 선생을 기리며"(≪국어교육≫ 132호, 2010)에 난대 선생의 업적을 정리한 바 있다.

에 관여하였다. 파주 출신의 표준어 구사자로서 언어생활에서도 말씨, 발음, 태도에서 표준어 사용의 모범을 보였다. 표준발음사전과 발음 교육 자료 발간, 교사와 방송인의 발음 교육, 낭독 교육에 앞장서고 학회장 시절에는 전국국어낭독대회를 주관하는 등 표준어 교육에 힘썼다.

첫 논문은 '언어와 도의'(서울 사대 교육회 2호)로 국어교육에서 언어예절 교육의 중요성을 강조하였다. 언어예절 연구는 표준화법 연구로 이어져 ≪말을 어떻게 할 것인가: 효과적인 화법의 비결≫(이주행 공저, 1992)로 나온다. 국어순화 영역에서도 여러 논문을 발표하고 퇴임 후인 1990년에도 문교부 국어심의회 국어순화분과위원장을 맡아 일생 국어 규범 확립에 힘썼다.

(2) 국어교육학과 국어교육사의 학문적 정립

1957년 서울대 교수로 부임한 이래 문교부 교수요목 제정심의회 위원, 국정교과용 도서 편찬심의회 위원 등 국어과 교육과정, 교과서 정책과 개발에 관여하였다. ≪고전, 작문, 문법, 한문≫ 등의 교과서 저자로도 학문적 깊이와 넓이를 보여 준다.

국어교육 개론서로 ≪국어교육≫(1963), ≪국어과교육론≫(1973), ≪국어과교육≫(1975)을 공저로 냈고, 현장 지침서 ≪국어과 교육실습의 계획과 실천자료≫(1981)를 냈으며, 제자들도 고희를 기려 ≪광복 후의 국어교육≫(1992)을 내어 국어교육사를 정리하였다.

국어교육의 전통을 중시해 한문교육을 국어교육에 계승함의 중요성을 강조하였다. 선인들의 한자 학습, 서사(書寫) 지도, 선인들의 독서와 작문 교육의 전통을 밝히는 논문을 다수 발표하고 ≪국어교육사연구≫(1975), ≪속국어교육사연구≫(1988)를 내놓았다. ≪국어학사≫(이기백 공저, 1987)도 내어 국어학사의 토대 위에 국어교육사 연구를 하였다.

(3) 어휘교육과 계량언어학적 어휘 연구

≪국어교육≫ 2호에 실린 '국어과 학습에 있어서의 단어 지도 문제'(1960)에서는 단순 낱말 풀이 지도식 교육을 비판하고 단어 뜻풀이를 넘어 문맥 속에서 문장 전체 의미를 파악하고 글 전체의 주제 파악에 이르도록 지도하는 문맥 기반 의미론적 어휘 지도를 제안하였다.

이 논문에서 '이해 어휘, 표현 어휘'의 개념을 제시하고 개인의 어휘 능력은 개인이 구사하는 어휘의 '양(量), 질(質), 영역(領域)'이라는 3차원에 걸쳐 나타난다고 한다. '양'은 어휘의 수효이고, '질'은 어휘의 심도와 어감이며, '영역'은 방면과 범위를 가리킨다.

작문 능력의 발달을 위해 이해 어휘가 표현 어휘로 전환되도록 때로는 어휘를 외우게 하고 짧은 문장 짓기를 통해 체득시켜야 하며, 국어 시간에 단어 뜻풀이에 매달려 시간을 보내는 것은 큰 폐단으로 단어 예습은 집에서 미리 해오고 교실에서는 난해어 중심으로 의미 파악을 도와야 한다고 주장한다.

'국민학교 국어 교과서 편찬을 위한 학습용 기본어휘 설정에 관한 연구'(1969)는 어휘교육의 기초 연구 보고서로 이의 미비점을 보완한 것이 ≪국어교육≫ 18호의 '국민학교 학습용 기본어휘'(1972)이다. 이 논문은 학습용 기본어휘를 통계언어학적으로 분석한 논문으로 문교부의 ≪우리말에 쓰인 글자 잦기 조사≫(1955), ≪우리말 말수 사용의 잦기 조사≫(1956) 이후로 대규모 어휘 조사를 한 것이다. '학습용 기본어휘'를 '기능적인 어휘, 가치 획득에 필요한, 효과적인 어휘'라야 한다고 정의한 후, 그 조건을 '①사용도가 높은 어휘, ②사용 범위가 넓은 어휘, ③조어력이 높은 어휘, ④기초적인 어휘' 등 네 가지로 들었다.

학습용 기본어휘는 각 교과 학습에 필요한 어휘로 이해 어휘와 표현 어휘로 나누어 성인 작품, 어린이 신문·잡지, 교과서, 어린이 작품에서 수집하였

고 어린이 대화도 녹취하였다. 총 329,123어[이어수(異語數) 17,335어(단, 고유명사 제외)]가 수집되었다.

어린이 대화는 도심, 변두리의 6개교 1~6학년 어린이들이 등·하교, 운동장에서 놀 때의 대화를 10분씩 녹취 및 전사하여 음성언어 조사의 효시를 보인다. 조사(助詞) 수효 77,638어[이어수(異語數) 231어]를 빼면 연어휘수(延語彙數) 총계는 251,485어[이어수(異語數) 17,104어]로 평균 빈도 10 이상의 어휘는 2,713어이며 최종적으로 제시한 이해 어휘는 147,733어(83.8%), 사용 어휘는 28,540어(16.2%)로 이해 어휘와 사용 어휘의 총계는 176,273개로 나타났다. 이에 따르면 이해 어휘는 사용 어휘의 5.2배이다.

이 연구에서 한자어 부분만 따로 분석, 연구한 것이 '초등학교 학습용 기본어휘의 한자 연구'(≪어문연구≫ 7·8호, 1975)이다. 이 논문에서는 기본 어휘 목록을 조사한 결과 799자의 한자를 추출하였고 이를 1~3급과 급외(級外)의 네 등급으로 했다.

'국민학교 입문기 학습용 기본어휘 조사 연구'(≪국어교육≫ 32호, 1978)는 1977년에 당시 초등학교 1학년 교과서와 서울·경기 지역 9개교 1학년(6~7세)과 미취학(5~6세) 어린이들의 대화를 녹취하여 초등학교 입문기 어린이들의 어휘 실태를 조사한 것으로 연어휘수(延語彙數)는 15,130어, 이어수(異語數)는 2,280어로 나왔다.

입문기 어휘 연구로는 '한글 구조에 따른 입문기 문자 지도'(≪어문연구≫ 15·16호, 1977)가 있다. 진주 지방에서 문맹자가 중학 신입생의 11.9%라는 보고가 있어서 문맹자를 예방하려면 입문기 아동에게 문장식(sentence method), 단어식(word method) 방법이나 시각어휘(sight word: 단어 전체를 한눈에 덩어리로 익히기) 방법보다는 한글 음절표를 활용한 문자 지도를 하라고 주장하였다. 한글은 음소문자 겸 음절문자로 운용되므로 ≪훈몽자회≫의 전통처럼 '가갸거겨…' 식의 기본음절표를 국어 교과서와 교실 벽에 붙여서 입문기 학생들이 익히도록 하고 낯선 단어도 음절 분해를 통해 지도하라는

것이다. 현대 국어교육에서 이것을 지도하지 않고 서구의 문장식, 단어식 낱말 지도를 하여 문맹자가 나온다는 것이다.

이인섭·김승렬 교수와의 공저인 '국민학교 아동의 어휘력 조사 연구: 저·중·고 학년별 표준 어휘 목록의 작성'(≪국어교육≫ 42호, 1982)은 '국민학교 학습용 기본어휘'(1972)와 '국민학교 입문기 학습용 기본어휘 조사 연구'(1978)의 어휘에서 외래어와 비속어를 제외하고 새로 1982년에 개편된 국민학교 1~3학년 교과서의 어휘를 종합하여 교사 8인과 함께 취학 전 수준과 저·중·고급 수준의 총 4단계 수준으로 조사하였다. 그 결과는 1수준(취학 전) 1,600어, 2수준(1·2학년) 4,389어, 3수준(3·4학년) 5,840어, 4수준(5·6학년) 3,176어, 총계 15,005어로 판정하였다.

이상의 계량언어학적 어휘 연구의 영향을 받아 제자 김광해 교수는 1980년대에 유의어·반의어 사전을 내놓는 등 어휘론 분야에서 창의적 연구를 보였고 그 유업은 '(주)낱말(natmal.com)'을 통해 결실을 맺고 있으니 '기사기제(其師其弟)', 곧 그 스승에 그 제자라 하겠다.

어휘력 확장법도 연구하여 '사전 속에 잠자는 가용(可用) 국어 어휘'(≪국어교육≫ 30호, 1977), '속(續) 사전 속에 잠자는 가용 국어 어휘'(≪국어교육≫ 67·68호, 1989)를 냈다. 전자는 한글학회의 ≪우리말큰사전≫에서 1,319어를 수집했는데, 이 중 '가두리(물건 가에로 돌린 언저리)'는 '가두리양식장'으로 잘 쓰인다. '까딱수(바둑, 장기 따위에 요행을 바라는 얕은 수), 갈매(짙은 초록빛), 나비잠(아기가 두 팔을 머리 위로 벌리고 자는 잠)' 등도 되살릴 만하다. 후자는 이희승 편 ≪국어대사전≫(민중서림)의 어휘 중에 1,192어를 소개해 '노루꼬리(몹시 짧은 것의 비유), 눈빔(남의 눈에 들게 겉으로 꾸미는 일), 도르리(음식을 돌려가며 제각기 내는 일)' 등도 쓰일 만하다.

고문헌에서 잠자는 옛말도 탐색해 '두시언해에 깃든 되살릴 말들'을 1984년에 신설된 '국어연구소'의 기관지 ≪국어생활≫에 다섯 번(2·3·4·5·11호, 1985~1987), ≪국어교육≫ 73호(1991)와 77호(1992)에 두 번 더

게재해 '잡들다(붙들다, 부추기다), 서의ᄒᆞ다(쓸쓸하다), 숫그리다(두려워하다)' 등을 소개하였다. 이 노력은 ≪아름다운 우리말을 찾아서≫(2001), ≪사전 속에 잠자는 보배로운 우리말≫(2007)로 나왔다.

(4) 한자교육 연구

우리나라에서는, 1965년 2월까지 국민학교 교과서에 괄호 한자를 병기하였고 1969년까지 4학년부터 200자씩 600자를 노출 혼용하였다가 1970년부터 모든 한자를 제거하였다. 그 후 반대 여론으로 1975년부터 국어, 국사만 중학교 교과서부터 괄호 한자를 넣었다. 결국 초등학교 국어과에서 한자교육을 정규교육으로 제공하지 않게 되자 남광우 박사와 함께 한국어문교육연구회를 만들고 초등학교 한자교육을 역설하였다.

한자 연구 논문과 논설은 ≪어문연구≫에 50여 편이나 실렸는데 첫 게재는 '현대 인명·지명에 쓰인 한자 연구: 기초 한자와 관련해서'(≪어문연구≫ 4호, 1974)이다. 고유명사의 경우 '노령, 재령'이라는 지명을 한글로만 쓰면 '노령(蘆嶺), 재령(載寧: 재녕>재령)'의 뜻이 구분되지 못한다. 이 연구는 1970년판 '전화번호부'의 인명과 1963년판 '지방행정구역편람'의 지명을 조사해 문교부 제정 교육용 한자 1,800자가 얼마나 쓰였는지를 분석하였다. 두 문헌에 쓰인 한자는 3,039자인데 교육용 한자는 1,348자(44.3%)에 불과해 교육용 한자로는 인명과 지명의 반도 못 읽는다. 인명과 지명에 공통으로 쓰인 한자가 888자이고 인명에만 쓰인 것이 212자, 지명에만 쓰인 것이 248자에 그쳤다. 또 교육용 한자 중 인명과 지명에 쓰이지 않은 한자가 452자로 교육용 한자의 1/4이다. 교육용 한자가 아니면서 인명과 지명에 쓰인 한자는 1,691자나 되는데 빈도 높은 90자는 최우선적으로 익혀야 할 한자들이라고 한다. 한자에 대한 계량언어학적 연구는 ≪자료를 통해 본 한자, 한자어의 실태와 그 교육≫(1989)에 집대성되어 있다.

한자교육의 장점은 다음과 같이 역설하였다. 첫째, 어휘 확장력에 유용하다. '農' 자를 통해 '농가, 농민, 농사, 빈농, 부농'을 확장하고, '家' 자를 통해 '가정, 가장, 친가, 외가, 초가, 농가' 등으로 확대할 수 있다. 둘째, 어휘 추리력과 철학적 사고에 좋다. '劇' 자를 통해 호랑이(虎)와 멧돼지(豕)가 송곳니를 칼(刀→刂)같이 세우고 싸우니 얼마나 극적인가. 또한 '學校'의 '學'에 '子' 자가 들어 있어 학생들은 아들과 같다는 것이고, '校'에 '父' 자가 들어 있어 스승은 아버지와 같다고 유추하여 결국 선생님은 아버지와 같고 우리는 선생님의 아들과 같다고 유추함은 놀라운 가르침을 준다. 셋째, 인격 형성에 유용하다. '孝' 자는 노인을 자식이 모신다는 뜻을 포함하고 있고, '親' 자는 나무에 서서 자식이 돌아오는 것을 멀리 바라본다는 뜻을 지니며, '信' 자는 두 사람 사이에 말의 신용이 있어야 하고, '仁' 자는 두 사람이 어질어야 한다는 것이며, '恕' 자는 내 마음(心)같이(如) 헤아려 용서해야 함을 뜻하고, '怒' 자는 내 마음이 노예처럼 묶인 상태일 때 성을 내게 되는 것이라는 교훈을 준다.

한자는 초등 1학년부터 가르쳐야 하며 중학부터 한문 과목으로 가르치면 학습 부담을 주므로 초등학교에서부터 익혀야 한다는 것이다. 한글 전용으로 지적 저하를 가져옴도 개탄하고 전문어를 과도하게 순화하여 오히려 독해를 방해하는 잘못된 국어순화도 비판한다.

'국민학교 각 과 교육과 한자'(≪어문연구≫ 53호, 1987)에서는 각 과목을 통한 한자교육의 방향도 제시하였다. '농도, 화농, 농아, 농로'를 한글로는 '농'이 구분되지 않으나 '濃度, 化膿, 聾啞, 農路'처럼 한자로 적으면 의미 구분이 명료하므로 수많은 한자 개념어들을 각 과에서 한글로만 가르치면 무수한 '무의미 철자'를 기계적으로 외워 고역을 치르고 어휘력 및 학력 저하는 필연적이라는 것이다.

'낭림산맥(狼林山脈)'도 '이리가 나오는 숲으로 이루어진 산줄기'로 이해하면 되는데 뜻도 모르고 '낭림산맥'이라는 소리만 외우듯이 인명, 지명 등의 고유명사를 외우면 문화적 교양도 쇠약해진다는 것이다. 한자가 배우기 어렵

다는 지적에 대해서도 어려서부터 한자를 익혀 온 전통이 있어서 아무 문제가 없으며 같은 또래의 중국, 일본 아동들이 한자를 잘 익혀 쓰고 있으므로 학습 부담은 성립되지 않는다고 하며, 우리는 하루아침이면 익힐 한글만 익히고 나서 한자 학습의 귀한 경험을 전 국민적으로 포기하고 있다는 것이다. 배워야 한다면 어려워도 배워야 하는 것이며, 한자는 국어와 별도로 배워서도 안 되며 같이 배워야 한다고 주장하여 국어와 한자교육의 통합을 주장하고 있다. 동아시아권에서 고립되지 않고 세계화하려면 한자교육이 국어과에서 이루어져야 한다는 것이다. 이상의 한자교육론은 ≪한자를 아는 것이 국력이다≫(2004)에 종합되어 있다.

퇴임 후에는 한·중·일(韓中日) 공통상용한자 제정에 깊은 관심을 가지고 '한자문화권내의 공통상용한자 검토'(≪어문연구≫ 112호, 2001)와 같은 계량적 연구 논문을 발표하였다. 별세 3년 전에도 '한자문화권 내 공통상용한자 선정 연구 경위와 향후 과제'라는 논문을 발표해 한중일 이체자(異體字) 문제를 다루고 공통상용한자 목록을 제시하였으니 80대의 원로로서 왕성한 학구열로 계량언어학적 연구를 제시한 열정은 후학들의 귀감이 된다.

고교 한문 교과서는 물론 대중 한문 교양서로 ≪한중한문연원(韓中漢文淵源)≫(2000)도 냈다. 월간 ≪한글+한자문화≫에 이응백 칼럼을 통해 주옥 같은 글을 실었으며, 투병에 들어간 2009년 10월부터 별세 한 달 전 2010년 3월호까지 ≪한글+한자문화≫에 '논어의 지혜'를 연재하였다.

우리의 어문정책은, 공공언어의 한글 전용하에서도 한자교육은 초등학교의 창의체험활동교육시간에, 중학교는 한문 선택과목을 통해 제공한다는 것인데 쟁점 사항은 초등학교 한자교육의 정규화 문제이다. 국민 여론조사는 초등학교에서 한자를 가르쳐 달라는 것이 70~80%로 나오므로 한글 전용 교육의 정신을 살리면서 초등 국어과에서 어휘 연습란을 통해 몇백자 수준으로 국어 어휘교육 보조 차원에서 제공하면 한글 중심의 정책도 살고 국민의 한자 소양도 높이게 될 것이다.

Ⅲ. 신심(信心)과 시심(詩心)의 수필가, 시조 시인

늘 단정하며 꼿꼿하고 자애로운 이 시대 마지막 선비는 수필가요, 시조 시인으로 허다한 작품을 남겼다. 독실한 불교 신자로서 많은 보시를 하였고 자애로운 미소는 자비로운 신심(信心)의 발로라 하겠으니 '수필문학진흥회' 회장 겸 잡지 ≪수필공원(隨筆公苑)≫ 대표로서 고전의 향기와 신심 가득한 수필을 남겼고, 시조 부흥 운동을 위해 '시조생활사' 대표로서 시심 어린 작품을 남기며 잡지 ≪시조생활≫을 펴냈다.

훌륭한 수필가로는 "깔끔한 서정 수필 작품으로 금아(琴兒) 피천득(皮千得) 선생의 수필이요, 소재의 어떠함을 가리지 않고 서정적이건 사회 비평적이건 간에 능란하게 처리해 내는 것은 김소운(金素雲)의 수필"이라고 두 분의 수필을 백미로 평가했다.

특히 수필에는 진한 가족애가 담겨 있으니 대학 4학년(26세) 때 소학교 은사의 중매로 은사 동창의 따님인 민영원(閔瑛媛) 여사를 만나 1949년 4월 창경궁 경춘전(景春殿)에서 혼례식을 올리고 9년 만에 외아드님을 얻고 두 손자를 얻었으니 가족 사랑을 그 무엇에 비교하랴. 결혼 30주년 기념으로 나온 ≪가족문집: 제비≫(1979)는 둘이 억센 비바람도 서로 견디고 의지하며 사이좋게 보금자리를 꾸미는 제비를 본받자는 뜻에서 지은 이름으로, 이후 ≪부부해외여행문집: 여적(旅滴)≫(1983)도 나왔다. 사모님께서 1993년 10월 65세에 먼저 하늘나라로 가시니 상배(喪配)의 아픔 속에서도 못다 한 사랑의 사부곡(思婦曲)을 1년 후 ≪아내 추모문집: 영원한 꽃의 향기≫(1994)와 ≪속(續) 영원한 꽃의 향기≫(1995)로 냈으며, 사모님의 아호 '혜순(慧橓)'을 딴 ≪세 번째 영원한 꽃의 향기: 혜순의 붓자취≫(1996)를 냈고, 홀로 맞은 결혼 50주년에는 ≪네 번째 영원한 꽃의 향기: 난향 죽정(蘭香竹情)≫(1998)을 냈다.

사부곡(思婦曲)

다시는 죽어도 님의 곁 안 떠나리
이제는 죽어도 님과 함께 있으리
그대 그림자 그대 따르듯
나도 그대 그림자 되어
죽어도 님의 곁 안 떠나리
차라리 그대 그림자 되었던들
이리도 그리운 마음 아니도 애달프련만

그 후에 ≪다섯 번째 영원한 꽃의 향기: 그리움의 사연들≫(2003)도 나왔다. 그 외에도 수필집 ≪기다림≫(1988), ≪고향길≫(1990), ≪우리가 사는 길≫(1999), ≪묵은 것과 새것≫(2008)이 나왔고, 퇴임 후에는 ≪제1 시조집: 인연≫(1992), ≪제2 시조집: 나들이≫(2002), ≪제3 시조집: 청(淸)과 음(陰)≫(2008)을 별세 2년 전까지 냈으니 왕성한 글쓰기로 시심을 밝힌 선청어문(先淸語文)의 삶이 수도자의 삶과 같았다. 댁에는 1944년부터 쓰신 일기가 63권이나 되니 그 기록 정신은 언젠가 생활사 차원에서 훌륭한 연구 자료로 공개될 때가 있으리라. 신심과 시심의 글쓰기야말로 17년을 홀로 외유내강의 풍모를 잃지 않으시고 지내는 원동력이 되었다.

Ⅳ. 학회 활동과 사회 활동

국어교육 최초의 학술단체인 '한국어교육학회(전 한국국어교육연구회)'를 1955년에 주도적으로 창립하고 회장 취임 이래 정년퇴임 5년 후인 1993년까지 38년간 학회장으로 학회를 이끌고 물심양면으로 후원하며 월례발표회를 지속해 왔다. 회장직의 장기 연임을 요즘에는 이해하기 어렵지만 당시는 한글학회 등에서도 원로 존중의 미덕으로 행해 왔다.

국어교육을 학문적으로 인정하지 않고 국어국문학의 말석쯤으로 아는 세태 속에서 국어교육의 가시밭길을 성실히 일구고 학회와 학술지 ≪국어교육≫을 최장수 학회지로 일구어 난향 가득한 큰 숲으로 만들었다. 그런 덕분에 국어교육학이 학문적 독자성을 인정받아 퇴임 직전에야 서울대와 교원대의 대학원에 국어교육 석사, 박사 과정이 생기게 되었다.

주시경 문하에서 '조선어학회'가 탄생하고 후학들에 의해 국어학이 성장했다면 국어교육은 난대 문하에서 '한국국어교육연구회'가 탄생하고 오늘날 '한국어교육학회'로 이어져 후학들에 의해 국어교육학이 학문적 정체성을 가지고 풍성하게 성장할 수 있게 되었다.

계량언어학적 연구를 가능케 한 꼼꼼하고 치밀한 성품의 난대는 행정 능력도 뛰어나 한국방송통신대학 초대 학장, '한국어문회' 이사장, '전통문화협의회' 회장, '한국수필문학진흥회' 회장, '중봉(重峯) 조헌(趙憲) 선생 기념사업회' 회장, 단군 숭모의 '현정회(顯正會)' 회장 등을 역임하여 국학계의 원로로서 헌신하였다. 본관(本貫)이 전주(全州)로 정종대왕(定宗大王) 제4남 선성군(宣城君)의 16대손이라 전주이씨 대동종약원(全州李氏大同宗約院) 이사 및 선성군파 종친회장도 역임하였다. 평소 나라 사랑을 기반으로 한 국어교육이 이루어지도록 강조하여 의로운 일을 현창(顯彰)하는 데 적극적이었다. 한자교육 주장도 한자교육이 국어교육이자 인간교육, 인격교육임을 강조하여 나라가 도의적, 인격적으로 발전하려면 초등학교에서부터 한자교육이 이루어져야 한다고 믿었기 때문이다.

국민훈장 동백장(1982), 대한교육연합회 교육특별공로상(1987), 국민훈장 모란장(1988), 수필과 비평문학 대상(1998), 제2회 동숭학술상(1998), 한국어문회 공로상(2009)을 수상하였고 2007년 10월에는 서울대 국어교육과 동문회가 첫 '자랑스러운 국어교육인'으로 선정했으며 2009년 5월에는 서울대학교 사범대학 동창회에서 국어교육을 이론적으로 체계화하고 실천한 공로를 인정해 제1회 청관대상(淸冠大賞) 학술상 수상자로 선정하였다.

Ⅴ. 고향 사랑, 제자 사랑

고향 파주에 대한 사랑도 각별하였다. 고향의 정서와 어릴 적 소년 시절의 꿈을 담아 적성종합고등학교(현 경기세무고등학교)의 교가를 작사하였고, 파주의 풍광을 담고 파주의 정신으로 통일을 염원하는 <파주의 노래>도 작사(김성태 작곡)하여 난대의 정신은 파주인의 가슴에 영원히 흐를 것이다. 가사에는 "내 고장 파주는 한국의 허리 우리 모두 힘을 모아 조국의 역군 되세."라는 구절이 나온다. 별세 한 달 전(2010. 2. 24.)에는 3,300여 권의 소장 도서를 파주 시립 중앙도서관에 기증하였다. 파주 지역신문인 ≪파주신문≫에는 250여 회 넘게 ≪논어≫를 중심으로 '어문수상(語文隨想)'을 연재하였다.

난대를 추모하는 제자들은 허다하다. 제자들에게는 교사로서 판서(板書)를 바르게 하라, 발음을 정확히 하라, 언어예절을 각별히 하라, 한글 어휘교육과 함께 초등학교 한자교육을 시행하여 어문교육을 정상화하라고 역설하며 본을 보였다. 제자들의 취업을 위해서는 동분서주하였다. 제자에게 사소한 일을 하나 시켜도 대가를 반드시 지불하여 제자를 함부로 부리는 일이 없었다. 흐트러짐 없이 온화하고 단아한 모습의 선비가 국어교육과에 계셨기에 그 제자들은 행복하였다. 퇴임 후 기탁 제정한 '난대 장학금'이 계속 유지되도록 기부해 왔고 별세 1년 전에도 한국어문회에서 수상한 공로상 수상금을 기탁해 난대의 학문 정신은 계속 후학들로 이어지고 있다.

난대의 삶은 전통 한문 국어교육의 튼튼한 기초 위에 국어학 연구로 시작하여 맞춤법, 표준어, 표준화법 규범 확립에 기여하고, 국어교육학의 초석을 놓았으며 특히 계량언어학적 방법으로 어휘교육, 한자교육, 국어교육사의 이론과 실제를 확립하고 교육과정, 교과서, 교사론 발전에 크게 기여하였다. 국어교육, 한자교육 관련 학회를 창립하여 큰 숲을 이루고 각종 사회·봉사 활동으로 사회적 책임을 다하였으며 신심과 시심의 가족 사랑, 나라 사랑,

제자 사랑을 담은 주옥같은 수필과 시조를 남긴 선비로서 선청어문의 수도자적 삶을 보여 국어교육의 큰 스승으로 기억되고 있다. 난초의 받침대만으로도 족하다는 아호 '난대(蘭臺)'의 뜻 그대로 우리 말글과 제자와 조국을 위해 겸손하게 희생한 선비의 향기로운 삶이었다.

국어 산책

‘오징어 윤, 쌍길 철’의 추억

홍성호
한국경제신문 기사심사부장

한자 혼용 시절, 기자들의 독법(讀法)

“오징어 윤(允), 쌍길 철(喆), 고무래 정(丁), 당나귀 정(鄭)….”

필자가 1980년대 후반 언론사 생활을 처음 시작할 때는 요즘 말로 소위 ‘아날로그 시대’였다. 지금같이 컴퓨터에 의한 전산 제작(CTS, Computerized Typesetting System)을 하는 게 아니라 납 활자로 글자를 하나하나 모아 기사 틀을 만들고 그것을 찍어 내던 시절이었다.

모든 기사는 종이 원고지에 손으로 직접 써야 했다. 공무국에는 글자 하나하나에 해당하는 납 활자가 갖춰져 있었는데 문선공(文選工)들이 원고지를 보며 활자를 뽑아냈다. 문선 작업이 끝나면 조판에 들어가기에 앞서 오자 없이 기사를 잘 뽑았는지 점검하기 위해 ‘게라(교정쇄)’가 편집국 교열부로 올라왔다.

당시는 교열을 볼 때 ‘允, 喆, 丁, 鄭’ 같은 한자를 정식 훈 대신 글자 형태를 따서 ‘오징어 윤, 쌍길 철’ 식으로 편하게 불렀다. 그중에서도 압권은 아마도 ‘탱크 설(卨)’이 아닐까 싶다. 성(姓)씨 중 하나인 이 글자는 정말 탱크처럼 생겼다.

신문에서 한자를 제법 쓰던 시절이라 당연히 한자 지식은 기자에게 필수였

다. 더구나 지금처럼 컴퓨터로 언제 어디서든 기사를 보낼 수 있는 환경이 아니었기에 마감 시간에 임박해 외부에서 급하게 전화로 기사를 불러 댈 때도 많았다. 그럴 때 서로 글자를 정확히 주고받기 위해 쓰던 '기법'이 바로 '오징어 윤, 쌍길 철' 방식이었다. 가령 같은 유 씨인데 '兪' 씨인지 '劉' 씨인지를 구별하기 위해 '인월도 유, 묘금도 유' 식으로 글자를 파자해 읽은 것이다. 이렇게 하면 서로 틀림이 없었다. 기사의 정확성을 높이기 위해 자구 방편으로 발전한, 기자들만의 독법이었던 셈이다.

그 시절 어문 담당 부서는 편집국에서 가장 시끄럽고 활기찬 부서였다. 십수 명이 앉은 공간에서 경쟁적으로 교정쇄에 박힌 '오징어 윤, 쌍길 철'을 읊어 대야 했기 때문이다. 이 작업은 통상 2인 1조로, 한 사람은 교정쇄를 읽고 다른 한 사람은 원고를 보면서 글자가 이상 없이 뽑혔는지 확인하는 과정이었다. 여기저기서 동시다발로 교정쇄를 읽어 대다 보니 목소리에서 밀리면 원고 잡은 선배에게 호된 꾸지람을 당하던 게 편집국 한 귀퉁이에서 벌어지던 풍경이었다. 그러니 어문 기자들에겐 업무에 들어가기 전에 목소리 상태가 필수 점검 사항이었다. 맑고 깨끗한 소리로 교정쇄를 읽어 내려가야 했기 때문이다. 혹여 전날 과도한 음주로 목소리가 탁해지거나 더듬기라도 한다면 그것이 곧 업무 생산성을 떨어뜨리는, 다소 황당한 시절이었다.

납 활자 시대의 언어 유산들

"누가 이렇게 많이 갈매기를 잡아 왔어?"

신문사 편집국 한 귀퉁이에서 갑자기 '갈매기' 얘기가 터져 나왔다. 무슨 일일까? 지금은 신문사에서 문선부라는 데가 없어졌지만, 납 활자 시절 공무국 문선부는 중요 부서였다. 특히 마감 시간이 임박해서는 그야말로 전쟁터였다. 다음 날 신문 수십 면에 들어가는 모든 글자를 문선공이 일일이 뽑아내야

했기 때문이다. 이때 어문 기자가 어법에 충실히 한다고 열심히 갈매기(띄어쓰기 표시인 '√'를 뜻하는 은어)를 그렸다가는 선배들에게 치도곤을 당했다. 띄어쓰기 하나를 바꾸려면 조판 과정 중 연쇄적인 글자 이동이 불가피하기 때문이었다.

차라리 오자를 잡는 것은 같은 자리에 글자 하나만 바꾸면 되니 쉬운 일이었다. 그러나 띄어쓰기를 하려다 보면 한 단에 들어가는 글자가 넘치거나 모자라게 되기 때문에 행 전체를 움직여야 할 때가 많았다. 경우에 따라선 다음 행, 그다음 행까지 연이어 글자를 조정해야 했기에 때로는 기사 전체를 움직여야만 했다. 그래서 문선부에서는 이를 제일 싫어했다. 자연스레 기자들이 띄어쓰기를 대수롭지 않게 여기게 됐고 웬만한 것은 붙여 쓰는 게 용납됐다.

당시 기준으로 띄어쓰기는 '단연코' 중요한 게 아니었다. 오히려 원활한 신문 제작을 방해하는 걸림돌 정도로 생각하는 경향이 강했다. 비록 맞춤법에는 틀리지만 웬만한 것은 붙여 쓰던 신문 기사의 관행은 그런 오랜 납 활자 시대의 제작 특성에서 나온 것이다. 지금도 그런 '고난의 시절'을 거쳐 온 신문 기자들 사이에는 띄어쓰기를 여전히 대충대충 하는 인식이 남아 있다.

아날로그 시절에 어문 기자들은 '조물대감'으로도 불렸다. 신문의 전반적인 표기·표현에 관해 실무 책임을 맡고 있는 어문 기자들은 당시 '끗발(?)'이 대단했다. ㅎ일보에서의 일이다. 교열 기자가 출근해 보니 편집국 한편에 '교열부는 造物大監'이란 글귀가 씌어 있었다. '해와 달을 바꾸는 권능은 조물주밖에 없느니라…'라는 주석과 함께. 한자를 많이 쓰던 시절 교열 기자의 실수로 전날 신문에 '日' 자가 '月' 자로 나간 것을 꼬집은 것이다. 이후 교열 기자에게 조물대감이란 별칭이 붙었다고 한다. 지금은 전설이 되다시피 한 1960~1970년대 납 활자 시대 얘기다.

당시엔 '淸凉里'나 '原麥 20萬 톤' 식으로 명사류는 물론 '犯하다, 加하다' 같은 서술어도 죄다 한자로 적었다. 또 기사 첫머리에 으레 따라붙던 '측문한

바에 의하면'과 같은 일본어 투 말을 '들리기론(지금은 이도 '~에 따르면' 꼴로 변했다)' 식으로 바꿔 쓴 것도, (원로 교열 기자이신 민기 선생의 회고에 따르면) 1960년대 들어서다.

'大統領'이 글자 하나가 빠져 '大領'으로 나가는가 하면, '共和黨'을 '共産黨'으로 개명하고, '社會正義'를 '社會主義'로 둔갑시키기도 했다. '장녀'를 순식간에 '창녀'로 만들고, '복지부'가 받침 없는 글자로 들어가 아찔했던 사건들도 비일비재했다. 한자와 문선(文選)의 시대에 볼 수 있었던 '언어의 마술'인 셈이다. 이 모두가 지금은 지난 시절 아련한 기억의 편린으로 남았다.

디지털 시대 언어의 '불편한 진실'

1990년대 들어 시티에스(CTS)가 도입되면서 신문 제작 환경이 급변했다. 컴퓨터로 작업하면서 당연히 원고지도, 교정쇄도 없어졌다. 어문 기자들이 매끄럽고 맑은 목소리로 교정쇄를 읽을 필요도 없어졌다. 지금 어문 담당 부서는 편집국 안에서 가장 조용한 부서가 됐다. 하루 종일 컴퓨터 자판을 두드리는 소리만 울려 퍼질 뿐 다른 소리가 나올 이유가 사라진 것이다.

신문 제작의 디지털화가 20여 년간 진행돼 오는 동안 기자들이 만들어 온 신문의 말도 시간의 축을 타고 명멸해 왔다. 그것은 우리말의 변천과 궤를 같이하는데, 좋게 말하면 시쳇말로 아날로그 언어에서 디지털 언어로 진화(?)하는 모습이라 할 수 있다.

우리나라 최초의 민간 신문인 ≪독립신문≫에서 표방한 한글 전용과 띄어쓰기가 제대로 구현된 것은 가장 큰 변화다. 1896년에 창간했으니 120년 전의 선구적 혜안이 지금 우리 언론을 통해 빛을 보는 것이다. 디지털 시대로 들어오면서 신문의 세로쓰기가 가로쓰기로 바뀌어 자연스레 기사를 한글 위주로 쓰게 됐다. 인터넷을 비롯해 모바일, 소셜 네트워크 서비스(SNS)

등 미디어 형태가 다양화하면서 매체 간 경쟁이 치열해지고 이에 따라 신문 기사도 독자가 좀 더 읽기 쉽게, 좀 더 이해하기 쉽게 바뀌었다. 정보기술(IT) 시대에 들어서면서 컴퓨터에 딱 맞는 한글의 우수성과 경쟁력이 빛을 본 것은 자연스러운 일이다. 하지만 얻은 것만큼 잃은 것도 있다. 그런 것들이 꼭 신문에서 한자가 사라진 부작용이라 단정할 수는 없을지언정 적어도 관련성은 있는 것 같다.

신문 제작이 아날로그에서 디지털로 바뀌면서 우선 기자들의 언어 의식에도 많은 변화가 뒤따랐다. 요즘 수습기자들의 우리말 능력은 납 활자 시대에 비해 뒤떨어지는 게 사실이다. 영어는 기본으로 하고 제2외국어, 제3외국어 능력까지 갖추고 있지만 정작 한자는 자기 이름도 쓰기 힘든 형편이다. 굳이 한자가 아니더라도 한자어를 비롯한 어휘 등 우리말 이해력이 많이 떨어지는 '불편한 진실'과 마주해야 한다.

SBS TV가 지난 5월 새로운 형식의 예능 프로그램으로 선보인 <스타꿀방 대첩 좋아요>도 제목이 생뚱맞아 불편하기는 마찬가지다. 제목만 봐선 이게 무슨 뜻인지, 프로그램 성격이 뭔지 알 수가 없다. 여기에 쓰인 '대첩'이 제대로 쓰인 것인지에 이르면 더욱 고개가 갸우뚱해진다. 대첩(大捷)은 '싸움에서 크게 이기는 것'으로, 이미 싸움이 끝나 승패가 갈린 상황에서 쓰는 말이다. '우리가 학창 시절에 배운 귀주대첩이니 한산도대첩이니 할 때의 그 대첩을 가져다 쓴 거 같은데…' 하는 생각을 하다 보면 제목이 전혀 어울리지 않고 따로 노는 느낌이다. 한자 의식이 약해지다 보니 말의 의미를 정확히 이해하지 못하고, 자연히 용법을 제대로 알지 못하게 되는 것 아닐까. 대첩이란 말을 아마도 '대전(大戰)' 정도로 알고 있는 것 아닐까. 그러니 이 말을 '큰 싸움', '한판 승부' 정도의 의미로 쓴 게 아닐까 하는 의구심도 든다.

‘면접관’은 아무 데서나 써도 되나

“면접관님, 드릴 말씀이 있습니다.”(민아)
“말해 봐요.”(면접관)
“종합 계획, 사업 등 좋은 우리말이 있는데 꼭 마스터플랜, 프로젝트라고 해야 합니까? 외국어를 너무 무분별하게 쓰는 것 아닌가요?”(민아)

걸스데이의 멤버 ‘민아’가 최근 합류해 열연 중인 KBS 1TV <안녕 우리말> 프로그램의 한 장면이다. 지난 4월에 방영된 ‘면접의 정석’ 편에서는 일반 회사에 취직하기 위해 면접을 본 민아가 외국어를 남용하는 세태를 날카롭게 꼬집었다. 이 프로그램은 일반인들의 언어생활에 파급력이 큰 공공언어와 청소년들의 언어 개선을 위해 제작·방송되고 있다. 일상에서 흔히 생기는 우리말 오류를 바로잡고 올바른 우리말 사용을 권장하는 내용으로 구성돼 있다. 그런데 직업의식의 발로일까, 여기에도 옥에 티가 있었다. ‘면접관’이란 표현이 그것이다.

면접관에서 ‘-관(官)’은 ‘공적인 직책을 맡은 사람’의 뜻을 더하는 접미사다. ‘감독관’을 비롯해 ‘경찰관, 법무관, 사령관, 소방관’ 같은 데 쓰는 말이다. 사전의 예에서도 보듯이 모두 정부 관리 따위를 말할 때 쓴다. 뒤집어 말하면 민간 기업의 직책에 쓰기에는 적절치 않는 말이다. 민간 회사의 사원 채용 기사를 보면 ‘면접관’이란 말이 많이 보인다. 민간인에게 ‘○○관’이란 표현을 쓰면 어색하게 느껴야 하는데 현실은 그렇지 않은 것 같다. ‘면접 담당자’나 ‘면접인’, ‘면접원’ 등 다른 적절한 말을 찾아 써야 한다.

하지만 한자 의식이 약해진 탓인지 요즘 젊은이들은 우리말의 이런 미묘한 용법 차이를 잘 이해하지 못한다. 그냥 대충 뜻만 통하면 되는 것으로 여기는 경우가 많다. 어휘력이 약해져 의미를 제대로 모르고 쓰는 것인데, 이는 우리말 육성과 진흥의 관점에서 당연히 역주행이다.

가뜩이나 우리말에는 개념어가 부족한데, 단어의 의미가 흐려지면서 용법

을 구별하지 못하고 아무 데나 말을 가져다 쓰는 사례가 점점 많아지고 있다.

신문 언어에서 범하는 흔한 오류 중 하나인 '역임'을 봐도 그렇다. 역임(歷任)은 '여러 직위를 두루 거침'을 뜻하는 말이다. 'A, B, C 등을 역임하다'처럼 나열할 때 써야 한다. 요즘 사람들은 이를 '대변인을 여덟 번 역임', '보건복지부 차관을 역임한 ○○○ 씨는~', '그는 ○○전자 기술총괄사장을 역임한 뒤~' 식으로 많이 쓴다. 아마도 역임을 역투(力投)나 역설(力說) 같은 말을 연상해 '(직무에) 힘껏 임하다' 정도로 해석해 쓰는 것 같다. 하나만 보여줄 때는 '거쳤다/맡았다/지냈다' 등을 적절히 골라 써야 하는데 어휘력이 달리니까 이를 알아채지 못한다. 앞에서 잘못 쓴 예도 '대변인을 여덟 번 맡아(지내)', '보건복지부 차관을 지낸(거친) ○○○ 씨는~', '기술총괄사장을 지낸 뒤~' 식으로 써야 할 곳이다.

현실적으로 신문에서 꼭 썼으면 좋을 한자도 모르는 경우가 왕왕 있다. 가령 젊은 사람들 가운데 '햄릿형(型) 성격'과 'S자형(形) 도로'에서 '형'을 구별해 쓸 수 있는 사람이 몇이나 될까. 그렇다고 모르니까 그냥 한글로만 써도 충분하다 넘겨 버리면 될 일일까. 그것이 진정 우리말을 살리고 키우는 길일까. 또, 공수(空輸)는 '항공수송'을 줄인 말로, 비행기로 무언가를 옮기는 것을 말한다. 그런데 좀 멀리서 가져온다 싶으면 자동차로 가져오든 배로 가져오든 무턱대고 '공수해 온다' 식으로 쓴다. 안타까운 점은 요즘 우리말 실태를 보면 이런 사례가 부지기수로 많다는 것이다.

100년을 이어 온 '민유총기' 망령

그런가 하면 있는 말을 제대로, 정확히 구사하지 못하는 경우도 많다. 살다 보면 누구나 한두 번씩 무언가를 임대하거나 임차하는 일이 있다. 얼마

전 집 앞 가게를 지나는데, '임대 문의'란 방문이 붙어 있는 걸 보았다. 순간 '임대 문의라…, 주인이 임대하는 것일 터인데 누구한테 문의한다는 것이지?' 라는 생각이 들었다. 누군가 점포를 빌리는 사람이 있으면 주인한테 '임차'에 관해 문의하라는 뜻으로 썼을 것이다.

사회가 발전할수록 세분화된 개념으로 언어를 더 정교하게 써야 하는데 우리는 있는 말도 구별 못하고 두루뭉술하게 쓴다. 그러니 "사무실이 비좁아 다른 건물을 '임대'해 쓰고 있다."라는 우스운 표현이 나온다. 사무실 공간이 부족할 때는 다른 사무실을 '임차'해 써야 하는데 이를 '임대'해 쓴다고 하는 것이다.

예전엔 '대통령이 사열을 받고 있다'는 식으로 망발을 하던 게 우리 언론의 언어 수준이었다. 그나마 요즘 신문에선 '사열하다'를 제대로 쓴다. 하지만 여전히 '자문을 구하다'란 표현은 많이 쓰인다. 이 말도 웃기는 표현이다. 자문을 마치 '현명한 답변' 정도로 생각하는 것 같다. 이렇게 맘대로 말을 쓰게 놔둬도 되는 것일까? 자문(諮問)은 전문가에게 의견을 묻는 일이다. 쉽게 말하면 질문과 비슷한 말이다. 그러니 '아무개가 (아무개에게 무엇을) 자문하다' 식으로 써야 논리적이고 과학적이며 어법에도 맞는 말이다.

더 기막힌 현실은 우리말에서 한자 의식은 많이 약해졌는데, 뿌리 깊은 일본식 한자어는 여전히 건재하다는 것이다. 이런 말들이 우리말을 해치는 중요한 병폐라는 것을 많은 이가 지적하지만 정작 이런 말을 쓰는 공무원 사회에는 마이동풍인 것 같다. 그런 말을 생각 없이 받아쓰는 신문이나 방송 등도 안타깝기는 마찬가지다.

'민유총기 일제신고'란 표현이 요즘도 인터넷에서 관련 기사로 검색되는 것을 보면 경찰청 등 관련 기관에서 여전히 쓰는 모양이다. 이 말은 100여 년 전 우리 신문에서도 보이니 그 뿌리는 일제가 만든 행정 용어일 터이다. 일제강점기 때 ≪동아일보≫에서 총독부 조사 자료를 인용해 보도한 기사(1928년 9월 23일 자)를 보면 '작년 말 현재 민유총기는 권총, 엽총을 합하여

사만 일천삼백칠십삼 명인데 그중에는…'이란 대목이 나온다. '민유총기'라니? 아마도 민간이 보유하고 있는 총기를 말하는 듯하다. 한자 사용에 대한 논란을 접어 둔다면, 이를 '民有총기'라고 쓰면 뜻은 쉽게 전달할 수 있을 것이다. 그런데 한글로 '민간 보유'를 줄여 '민유'라고 하면 일상적으로 쓰는 말이 아니라 뜻이 쉬이 떠오르지 않는다. 신문 언어든 일상의 언어든 독자가 그 말을 읽고 뜻을 미루어 짐작해야 한다면 언어로서의 효용성은 떨어진다.

'차집관거'도 정체불명의 암호 같은 말이다. 한글학회는 ≪한글새소식≫ 525호(2016. 5.)에서 독자 투고 형식으로 이 '차집관거'의 꼴사나움을 지적했다. 차집관거(遮集管渠)는 '막을 차(遮), 모을 집(集), 대롱 관(管), 도랑 거(渠)'로 된 말이다. 하수나 빗물을 모아 처리장으로 보내기 위해 만든 인공 도랑(물길)쯤으로 이해하면 된다. 투고자는 '대롱물길'을 제안했으나 이 역시 쉽게 들어오는 말은 아니다. 그보다는 '빗물도랑' 정도가 좋지 않을까(이 용어는 김도연 포스텍 총장이 ≪한국경제신문≫ 2016년 2월 22일 자 다산칼럼에서 제시한 것이다). 국민은 쓰지도 않고 알지도 못하는 말을 공무원은 태연히 쓰고 있는 것이 우리말의 아픈 현실이다.

과거 납 활자 시대에는 그나마 언어의 정체성이 확보됐다. 누구나 공감하는, 같은 말을 썼다. 말에 대한 의미 해석이 그만큼 보수적으로 유지됐기 때문일 것이다. 지금은 인터넷 시대를 넘어 모바일 시대다. 누구나 기자를 표방할 수 있는 1인 미디어 시대다. 소셜 네트워크 서비스(SNS)에는 무수한 실험 언어가 난무한다. 세대마다 쓰는 말이 다르고 공유하는 게 다르다. 자신들끼리만 아는 '또래 언어'가 판친다. 우리말의 의미 공유가 엷어질 수밖에 없고, 그로 인해 소통 실패를 가져오기 십상이다.

'솔까말(솔직히 까놓고 말해서), 낄끼빠빠(낄 때 끼고 빠질 때 빠지기), 복세편살(복잡한 세상 편하게 살자), 넘사벽(넘을 수 없는 4차원의 벽)….' 인터넷에는 이런 유의 조어가 헤아릴 수 없이 많다. 이들은 한순간 유행어로 끝나겠지만, 그마저도 소통에 걸림돌이 돼서는 안 된다. 자칫 우리말 질서와

체계를 무너뜨릴 위험성도 크기 때문이다.

사회가 진화, 발전함에 따라 새로운 개념이 속속 등장할뿐더러 기왕에 있던 개념도 세분화해 수많은 갈래를 만든다. 우리말이 이를 쫓아가느냐의 문제가 당장 눈앞의 과제다. 적절한 시점에 적절한 단어·용어가 없으면 외래어든 뭐든 물밀 듯 들어오고, 우리말도 휩쓸려 갈 수밖에 없다.

그런 점에서 언중의 생활 언어를 일선에서 담아내는 신문 언어는 여전히 자리 잡지 못하고 방황하는 모습이다. 신문 언어를 거르는 장치들은 언제나 제대로 작동하게 될까? 그렇다고 "쌍길 철, 오징어 윤…" 하던 시절로 되돌아갈 수도 없는 노릇이다. 신문 언어에 '철학'을 담아야 할 시대다.

국립국어원 소식

국립국어원 소식

1. 국립국어원 말다듬기위원회, 다듬은 말 선정

- 시에스(C.S. Customer Satisfaction)의 다듬은 말: **고객 만족**
- 데모데이(demoday)의 다듬은 말: **시연회**
- 플레이팅(plating)의 다듬은 말: **담음새**
- 젠트리피케이션(gentrification)의 다듬은 말: **둥지 내몰림**
- 리무버(remover)의 다듬은 말: **(화장) 지움액**

'시에스(C.S. Customer Satisfaction)'는 기업 경영에서 고객의 만족감을 높여 제품이나 서비스를 찾도록 하는 것을 뜻한다. 이를 '고객 만족'으로 다듬었다. '데모데이(demoday)'는 어떤 계획을 실시할 예정인 날 이전에 먼저 행사를 진행하는 날을 가리킨다. '시연회'로 다듬었다.

음식을 내기 직전에 먹음직스럽게 보이도록 그릇이나 접시 위에 담고 장식을 더하는 '플레이팅(plating)'은 '담음새'로 다듬었다. '젠트리피케이션(gentrification)'은 구도심이 번성해 중산층 이상의 사람들이 몰리면서 임대료가 오르고 원주민이 내몰리는 현상을 말한다. 이를 '둥지 내몰림'으로 다듬었다.

'리무버(remover)'는 특정한 물질을 제거하기 위해 사용하는 전용 물질이다. 이를 '(화장) 지움액'으로 다듬었다.

2. 정부·언론외래어심의공동위원회 심의 결과

- 정부·언론외래어심의공동위원회의 심의 결과는 국립국어원 누리집에서 확인하실 수 있습니다.
- 자료 위치: 국립국어원 누리집(http://korean.go.kr)
- '자료 찾기' → '연구 결과' → '기타 자료' → 검색: 검색어 '역대'

- 정부·언론외래어심의공동위원회 실무소위원회 2016년 제4차 심의 확정안 (2016. 3. 18.)
- 정부·언론외래어심의공동위원회 실무소위원회 2016년 제5차 심의 확정안 (2016. 3. 25.)
- 정부·언론외래어심의공동위원회 실무소위원회 2016년 제6차 심의 확정안 (2016. 4. 1.)
- 정부·언론외래어심의공동위원회 실무소위원회 2016년 제7차 심의 확정안 (2016. 4. 7.)
- 정부·언론외래어심의공동위원회 실무소위원회 2016년 제8차 심의 확정안 (2016. 4. 15.)
- 정부·언론외래어심의공동위원회 실무소위원회 2016년 제9차 심의 확정안 (2016. 4. 22.)
- 제126차 정부·언론외래어심의공동위원회 결정안(2016. 4. 27.)
- 정부·언론외래어심의공동위원회 실무소위원회 2016년 제10차 심의 확정안 (2016. 5. 10.)
- 정부·언론외래어심의공동위원회 실무소위원회 2016년 제11차 심의 확정안 (2016. 5. 13.)
- 정부·언론외래어심의공동위원회 실무소위원회 2016년 제12차 심의 확정안 (2016. 5. 20.)
- 정부·언론외래어심의공동위원회 실무소위원회 2016년 제13차 심의 확정안 (2016. 5. 27.)
- 정부·언론외래어심의공동위원회 실무소위원회 2016년 제14차 심의 확정안 (2016. 6. 3.)
- 정부·언론외래어심의공동위원회 실무소위원회 2016년 제15차 심의 확정안 (2016. 6. 10.)
- 정부·언론외래어심의공동위원회 실무소위원회 2016년 제16차 심의 확정안 (2016. 6. 17.)

3. 국립국어원, '한국수어 정책 토론회' 개최

국립국어원은 2016년 4월 8일 오후 1시 서울 여의도 이룸센터 누리홀에서 한국수화언어(이하 한국수어) 정책 토론회를 개최했다.

3.1. 〈한국수화언어법〉 시행 전 토론의 장 열려

지난 2월 3일 제정된 <한국수화언어법>은 2016년 8월 4일 시행을 앞두고 있다. 이에 국립국어원은 농인 당사자와 관련 전문가들을 모시고 아래와 같은 세 가지 주제를 중심으로 다양하고 체계적인 한국수어 정책에 대한 토론의 장을 마련했다.

(1) 한국수어 전문 조직 및 인력, 교육 과정 개발 등에 대한 제도 기반 마련
(2) 한국수어 연구, 실태 조사, 교육 자료 개발 등에 대한 수어 연구
(3) 수어 홍보 및 보급과 민간단체 지원 등 수어 확산

이 토론회는 우주형 나사렛대학교 교수가 좌장을 맡고 최상배 공주대학교 교수, 원성옥 한국복지대학교 교수 등이 발제를 맡았으며, 김정환 서울농아인협회 중랑구지부장, 진종순 명지대학교 행정학과 교수, 윤석민 국립국어원 한국수어연구자문위원회 위원 등이 토론을 진행하였다. 또한 객석의 농인 당사자들이 토론에 충분히 참여할 수 있는 시간도 마련되었다.

3.2. 한국 농인들의 언어권 확보 위한 법 제정

〈한국수화언어법〉(법률 제13978호) 정의(제3조)

- **한국수어**: 대한민국 농문화 속에서 시각·동작 체계를 바탕으로 생겨난 고유한 형식의 언어
- **농　　인**: 청각장애를 가진 사람으로서 농문화 속에서 한국수어를 일상어로 사용하는 사람

2015년까지만 해도 26만 명에 가까운 한국 농인들은 일상생활이나 전문 영역에서 한국수어로 소통할 권리를 얻지 못하고 살아왔다. 이들은 언어권(言語權) 확보를 위해 힘차고 끈기 있는 수어로 호소해 왔으며, 장애계 유관 단체들의 응원도 이어졌다. 지난 2013년 한국수어를 고유한 언어로 인정하고 대한민국 농인의 공용어로 선언하는 내용을 골자로 하는 4개의 법안이 국회에서 발의된 후 소관 부처인 문화체육관광부를 비롯하여 교육부, 행정자치부, 보건복지부 등 관계 부처도 해당 법 제정 필요성에 깊이 공감하고 협의해 나갔다. 그리고 마침내 이와 같은 민관 모두의 노력에 힘입어 선진적인 <한국수화언어법>이 2015년 12월 31일 국회를 통과하였다.

3.3. 한국수어 정책 방향에 대한 논의 이뤄져

지금까지도 많은 사람들은 한국수어를 청각에 장애가 있는 농인들이 소리를 대신하여 손짓 등 몸동작으로 흉내 내기를 하는 정도로만 알고 있다. 심지어 일부 드라마나 영화에서는 농인들이 한국어로 말하는 사람의 입 모양만 보고도 그 뜻을 쉽게 이해할 수 있는 것처럼 왜곡하고 있다. 그러나 한국어는 농인에게 전혀 생소한 언어이다. 마치 한국어 화자가 영어나 중국어를 대하는 것과 다름없는 외국어와 같아서 이런 외국어를 입 모양만 보고 알아듣는다는 것은 결코 쉬운 일이 아니다.

우리나라 농인의 모어(母語)는 한국수어이다. 한국수어는 한국의 농문화 속에서 시각·동작 체계를 바탕으로 생겨난, 한국어와는 구별되는 고유한 형식의 독립적 언어이다. 이번 토론회에서는 농인들이 한국수어로 불편 없이 일상생활을 영위하고 다양한 사회활동에 참여할 수 있게 하는 등 농인의 삶의 질 향상을 위해 정부가 추진해 나가야 할 정책 방향에 대한 다양한 논의들이 이루어졌다.

4. 또 하나의 언어 보물 상자, ≪한국수어사전≫ 새롭게 개통

4.1. 새로운 ≪한국수어사전≫의 필요성 대두

국립국어원은 한국수어를 모어(母語)로 사용하는 농인의 언어환경을 개선하기 위한 ≪한국수어사전(http://sldict.korean.go.kr)≫을 2016년 4월 20일 개통했다. 이 사전은 농인과 청인 모두 한국수어 일상용어와 전문용어에 대한 한국어 정보를 쉽게 찾아볼 수 있도록, 기존의 한국수어 웹 사전과 모바일 앱 사전, 그리고 책으로만 보급되던 전문용어 사전을 통합하여 새롭게 정비한 것이다.

국립국어원은 지난 2005년부터 (사)농아인협회와 더불어 한국수어 일상용어를 수록한 ≪한국수화(한국수어)사전≫과 법률, 교통, 종교 등 10개 분야 한국수어 전문용어 사전을 편찬해 왔다. 이 중에서 일상용어 사전은 인터넷 웹과 모바일 앱으로도 제공해 왔으나 사용 기기에 따른 제약이 있었고, 전문용어 사전은 종이로만 제작되어 활용도가 떨어졌다.

4.2. 국립국어원, 한국수어 자료 구축을 위해 노력 중

이번에 새롭게 개통한 ≪한국수어사전≫은 반응형 웹 사전으로 구축되어 다양한 기기에서 자유롭게 사용할 수 있고 그동안 종이 사전으로만 보급되어 오던 한국수어 전문용어도 포함하고 있어 사용자 편의성과 활용도를 높였다. 또한 사전 사용법을 수어 동영상으로 제공하고, 표제어를 클릭하지 않고 커서만 가져다 놓아도 수어 해설 영상이 바로 실행될 수 있도록 하여 보이는 언어인 한국수어를 모어로 사용하는 농인들이 좀 더 편하게 사전을 활용할 수 있도록 한 점이 눈에 띈다.

국립국어원은 이번 ≪한국수어사전≫에서 한 걸음 더 나아가 한국수어의 특성이 잘 반영되는 새로운 형식의 사전을 기획하고 있다. 앞으로 만들어질 ≪한국수어사전≫은 수형(한 손 또는 두 손이 결합하여 만들어 낸 의미 있는 형태), 수위(의미 있는 손 형태의 위치) 등 한국수어의 구성 요소가 잘 드러나도록 설계되어 농인이 더욱 편리하게 사전을 찾아볼 수 있도록 구축할 예정이다. 장기적으로는 지금의 한국어-한국수어 어휘 대조집 차원을 넘어서 뜻풀이도 한국수어로 제공하는 진정한 의미의 ≪한국수어사전≫이 만들어질 예정이다.

5. 국립국어원, 과천과학관과 한국수어 과학 정보 구축 업무 협약 체결

국립국어원과 국립과천과학관은 5월 31일(화요일) 오전 11시, 국립과천과학관에서 과학 정보 구축 및 보급을 위한 업무 협약식을 개최하였다. 두 기관은 과학 한국 건설을 위해서는 우리 국민 누구나 자유롭게 과학 문화를 누릴 수 있는 환경을 조성해야 한다는 데 인식을 같이하고 우리말로 쉽게 이해할 수 있는 과학 정보를 함께 구축해 나가기로 하였다. 이번 업무 협약에서 주목할 점은 장애로 과학 정보 습득에 어려움을 겪고 있는 청각장애인과 시각장애인을 위해 한국어 외에 한국수화언어(이하 '한국수어')와 한국점자로도 과학 정보를 구축하기로 한 것이다.

전문 분야에 대한 접근성을 높이기 위해서는 무엇보다도 사용자에게 익숙한 모어(母語)로 알기 쉽게 설명된 자료가 필요하다. 그런데 지금 우리나라에는 한국수어가 모어인 청각장애인을 위한 언어 자료가 턱없이 부족한 상황이며, 특히 과학 정보는 거의 없는 실정이다.

이에 국립국어원과 국립과천과학관은 청각장애인의 과학 정보 접근성을 꾸준히 높여 나가기로 하고, 그 첫 사업으로 올해는 국립과천과학관에 전시된 과학 체험물 50여 점에 대한 한국수어 해설 동영상을 제작하여 전시 관람에 활용할 예정이다. 국립국어원은 지난해 국립중앙박물관, 국립민속박물관, 국립한글박물관 전시물 650여 점에 대한 수어 해설 동영상을 제작하여 제공한 바 있으며, 올해는 대한민국역사박물관과 더불어 국립경주박물관, 국립부여박물관, 국립공주박물관, 국립진주박물관 등 지방 박물관 전시물에 대한 수어 해설 동영상 자료도 구축하여 관람객에게 서비스할 계획이다.

국립국어원과 국립과천과학관은 바르고 쉬운 한국어로 과학 자료를 제작하고 청각장애인에게는 그들의 모어인 한국수어로, 시각장애인에게는 손으로 읽는 한국점자로 해설 자료를 지속적으로 확충함으로써 우리 사회 구성원 누구나 과학 정보를 자유롭게 접할 수 있는 환경을 조성하기 위해 노력해 나갈 것이다.

6. 국립국어원, 제2회 국어사전 진흥 공모전 개최

국립국어원은 국어사전에 대한 국민의 관심을 불러일으키기 위해 지난해에 이어 2016년 국어사전 진흥 공모전 “함께 만들어 가요, 우리말 사전”을 개최한다. 이 행사는 누구나 사전 편찬에 참여할 수 있고, 편찬 과정을 볼 수 있는 사용자 참여형 국어사전 ≪우리말샘≫의 개통(2016년 10월 5일)을 앞두고, 낱말의 뜻풀이를 직접 해 봄으로써 우리말의 가치를 깨닫고 아름다운 우리말을 담은 사전을 함께 만들어 가자는 취지로 기획되었다.

“함께 만들어 가요, 우리말 사전” 공모전

- ㅇ 공모 부문: 창의적 뜻풀이(제시어 10개 중 5개 이상 뜻풀이)
  ※ 제시어: 한글, 멋, 꿈, 미래, 샘, 함께, 신나다, 해맑다, 정겹다, 만들다
- ㅇ 응모 기간: 2016년 8월 1일(월)~8월 15일(월)
- ㅇ 시상식: 2016년 10월 5일(수)/한국언론진흥재단 국제회의장(20층)
- ㅇ 수상작 전시회: 2016년 10월 8일~9일/2016 한글문화큰잔치 (광화문 광장)
- ㅇ “함께 만들어 가요, 우리말 사전” 누리집

(http://www.korean.go.kr/front/board/boardStandardView.do?board_id=4&mn_id=17&b_seq=1782&pageIndex=1)

6.1. 창의적 뜻풀이로 만들어지는 우리말의 미래

공모전은 제시된 10개의 낱말 중 5개 이상을 자기만의 개성으로 뜻풀이하는 것으로, 뜻풀이가 연상되는 사진이나 영상을 첨부할 수 있다. 제시어는 ‘한글의 멋을 담은 우리말이 자꾸자꾸 샘솟는 해맑은 미래를 꿈꾸고, 우리 모두 정겹게 소통하며 우리말 사전을 신나게 함께 만들어 가자’는 뜻을 담아 선정

되었다. 이 낱말의 선정에는 우리말을 사랑하고, 바르게 사용하는 기관의 대표(고은 겨레말큰사전남북공동편찬사업회 이사장, 문효치 한국문인협회 이사장, 정규성 한국기자협회 회장, 고영수 대한출판문화협회 회장, 김성규 한국어문학술단체연합회 대표, 고용우 전국국어교사모임 이사장)가 참여하였다.

6.2. 개인이나 단체, 외국인도 응모 가능

공모전에는 우리말을 사용하는 사람이면 누구나 참가할 수 있으며, 응모작 접수는 8월 1일부터 8월 15일까지이다. 심사는 공정성과 객관성을 위해 외부 심사위원(시인, 기자, 국어국문학 전문가, 국어 교사 등)이 진행하며, 대상 1명, 최우수상 2명, 우수상 4명, 장려상 8명, 단체상 3명 등 총 18명을 선정할 계획이다. 시상식은 10월 5일 ≪우리말샘≫ 개통 행사와 함께 열리며, 수상작은 '2016 한글문화큰잔치'의 일부로 진행될 '2016년 국어사전 진흥 공모전 수상작 전시회'에서 공개된다. 자세한 행사 내용은 국립국어원 누리집(http://www.korean.go.kr/front/board/boardStandardView.do?board_id=4&mn_id=17&b_seq=1782&pageIndex=1)에 안내되어 있다.

6.3. ≪우리말샘≫ 개통 행사에서 수상작 뜻풀이 입력 시연

대상 수상자는 ≪우리말샘≫ 개통 행사에서 자신이 뜻풀이한 낱말을 직접 입력함으로써 '≪우리말샘≫ 최초 어휘 등록자'로 기록될 것이다. 2016년 국어사전 진흥 공모전은 한국어 사용자가 새롭게 사용하는 낱말을 사전에 올리고 고치는 과정을 반복하면서 '함께 만들어 살아 움직이는 사전'을 위한 준비의 첫걸음이 될 것이다. 그리고 낱말의 폭넓은 의미와 다양한 쓰임을 다시 한 번 되새겨서 우리말의 가치와 아름다움을 깨닫는 기회가 될 것으로 기대된다.

7. 인사 이동

7.1. 임용

- 홍혜진(학예연구사 시보): 신규 임용, 어문연구과 근무(4월 16일)

7.2. 전보 발령

- 김동안(서기관): 문화체육관광부 → 한국어진흥과(1월 29일)
- 정시화(서기관): 문화체육관광부 → 교육연수과(2월 1일)
- 원경덕(행정주사): 기획운영과 → 문화체육관광부 재정담당관(4월 11일)
- 서광철(서기관): 문화체육관광부 감사관실 → 교육연수과(4월 16일)
- 박성욱(행정주사보): 문화체육관광부 출판인쇄산업과 → 기획운영과 (4월 27일)
- 이주형(행정주사보): 기획운영과 → 문화체육관광부 국제체육과(5월 25일)
- 이현우(행정주사보): 문화체육관광부 문화여가정책과 → 기획운영과 (5월 25일)
- 유은숙(행정서기): 국립민속박물관 → 기획운영과(6월 2일)

7.3. 승진

- 이보라미(학예연구사): 학예연구사 → 학예연구관(5월 25일)

7.4. 대우공무원

- 김미경(행정서기): 7급(행정주사보) 대우(4월 1일)

7.5. 휴직

- 김아영(학예연구사): 육아휴직 연장(2016년 4월 3일부터 2017년 4월 2일 까지)

≪표준국어대사전≫ 정보 수정 내용

2016년 1분기

표제항 (가나다 순)	수정 전	수정 후	비교
난형열	—	난형-열(卵形熱)[난 : -녈]「명사」『의학』 열원충이 적혈구를 침습하여 적혈구가 팽대해져 계란형이 되고 열원충이 적혈구 내에서 48시간 동안 분화되어 열이 오르는 말라리아.「참고어휘」삼일열;사일열;열대열.	표제어 추가
랍스터	-	랍스터(lobster)「명사」=바닷가재.	표제어 추가
아시가바트	-	아시가바트(Ashgabat)「명사」『지명』중앙아시아 투르크메니스탄에 있는 도시. 카라쿰 사막 서쪽의 이란 국경에 가깝고 트란스카스피안(Trans Caspian) 철도가 걸쳐 있는 요충지로, 직물·식료품 가공업이 발달하였다. 투르크메니스탄의 수도이다.	표제어 추가
좀목형	좀-목형(一荊)「명사」『식물』 → 좀모형.	-	표제어 삭제
좀모형	좀-모형	좀-목형	표제어 수정
녹섬광	녹-섬광	녹색-섬광	표제어 수정
끝장	일의 마지막. ≒기극02(紀極). ¶{끝장에} 이르다/처음에는 진지하던 토론회가 {끝장에는} 난장판이 되었다./너에게 무슨 일이 생기면 우리는 {끝장이야}./서두르다가 결정적인 순간을 그르치면 만사는 {끝장이다}./{끝장에} 가서 이긴 자는 함성을 올리고 진 자는 서로 얼싸안고 비탄에 잠긴다.≪선우휘, 사도행전≫	「1」일이 더 나아갈 수 없는 막다른 상태. ≒기극02(紀極). ¶{끝장에} 이르다/처음에는 진지하던 토론회가 {끝장에는} 난장판이 되었다./{끝장에} 가서 이긴 자는 함성을 올리고 진 자는 서로 얼싸안고 비탄에 잠긴다.≪선우휘, 사도행전≫ 「2」실패, 패망, 파탄 따위를 속되게 이르는 말. ¶너에게 무슨 일이 생기면 우리는 {끝장이야}./서두르다가 결정적인 순간을 그르치면 만사는 {끝장이다}.	뜻풀이 추가

나[10]	[1] 「1」((받침 없는 체언이나 부사어, 연결 어미 '-아, -게, -지, -고', 합성 동사의 선행 요소 따위의 뒤에 붙어))마음에 차지 아니하는 선택, 또는 최소한 허용되어야 할 선택이라는 뜻을 나타내는 보조사. 때로는 가장 좋은 것을 선택하면서 마치 그것이 마음에 차지 않는 선택인 것처럼 표현하는 데 쓰기도 한다. 「2」((받침 없는 체언이나 부사어 뒤에 붙어))마치 현실의 것인 양 가정된 가장 좋은 선택이라는 뜻을 나타내는 보조사. 빈정거리는 뜻이 드러난다. 「3」((수량이나 정도를 나타내는, 받침 없는 체언이나 부사어 뒤에 붙어))수량이 크거나 많음, 또는 정도가 높음을 강조하는 보조사. 흔히 놀람의 뜻이 수반된다. 「4」((수량의 단위나 정도를 나타내는, 받침 없는 체언이나 부사어 뒤에 붙어))수량이나 정도를 어림잡는 뜻을 나타내는 보조사. 「5」((수량의 단위나 정도를 나타내는, 받침 없는 말 뒤에 붙어))많지는 아니하나 어느 정도는 됨을 나타내는 보조사. 「6」((종결 어미 '-다', '-ㄴ다', '-는다', '-라' 따위에 붙어))화자가 인용하는 사람이 되는 간접 인용절에서 인용되는 내용에 스스로 가벼운 의문을 가진다든가 인용하는 사람은 그 내용에 별 관심이 없다는 뜻을 나타내는 보조사. 흔히 빈정거리는 태도나 가벼운 불만을 나타낸다. 인용자와 인용 동사는 생략될 때가 많다. 「7」((받침 없는 체언이나 부사어 뒤에 붙어))여러 가지 중에서 어느 것을 선택하여도 상관없음을 나타내는 보조사. 맨 뒤에 나열되는 말에는 붙지 않을 때도 있다. 「8」비교의 뜻을 나타내는 보조사. 뒤 절에는 결국 같다는 뜻을 가진 말이 온다.	[1] 「1」((받침 없는 체언이나 부사어, 연결 어미 '-아, -게, -지, -고', 합성 동사의 선행 요소 따위의 뒤에 붙어))마음에 차지 아니하는 선택, 또는 최소한 허용되어야 할 선택이라는 뜻을 나타내는 보조사. 때로는 가장 좋은 것을 선택하면서 마치 그것이 마음에 차지 않는 선택인 것처럼 표현하는 데 쓰기도 한다. 「2」((받침 없는 체언이나 부사어 뒤에 붙어))마치 현실의 것인 양 가정된 가장 좋은 선택이라는 뜻을 나타내는 보조사. 빈정거리는 뜻이 드러난다. 「3」((받침 없는 체언이나 부사어 뒤에 붙어))어떤 대상이 최선의 자격 또는 조건이 됨을 뜻하는 보조사. ¶{언니나} 되니까 너한테 양보하지./ {나나} 되니까 그 일을 해결할 수 있는 거다. 「4」((수량이나 정도를 나타내는, 받침 없는 체언이나 부사어 뒤에 붙어))수량이 크거나 많음, 또는 정도가 높음을 강조하는 보조사. 흔히 놀람의 뜻이 수반된다. 「5」((수량의 단위나 정도를 나타내는, 받침 없는 체언이나 부사어 뒤에 붙어))수량이나 정도를 어림잡는 뜻을 나타내는 보조사. 「6」((수량의 단위나 정도를 나타내는, 받침 없는 말 뒤에 붙어))많지는 아니하나 어느 정도는 됨을 나타내는 보조사. 「7」((종결 어미 '-다', '-ㄴ다', '-는다', '-라' 따위에 붙어))화자가 인용하는 사람이 되는 간접 인용절에서 인용되는 내용에 스스로 가벼운 의문을 가진다든가 인용하는 사람은 그 내용에 별 관심이 없다는 뜻을 나타내는 보조사. 흔히 빈정거리는 태도나 가벼운 불만을 나타낸다. 인용자와 인용 동사는 생략될 때가 많다. 「8」((받침 없는 체언이나 부사	뜻풀이 추가

		어 뒤에 붙어))여러 가지 중에서 어느 것을 선택하여도 상관없음을 나타내는 보조사. 맨 뒤에 나열되는 말에는 붙지 않을 때도 있다. 「9」비교의 뜻을 나타내는 보조사. 뒤 절에는 결국 같다는 뜻을 가진 말이 온다.	
라[07]	((‘이다’, ‘아니다’의 어간이나 어미 ‘-으시-’, ‘-더-’, ‘-으리-’ 뒤에 붙어)) 까닭이나 근거 따위를 나타내는 연결 어미.	((‘이다’, ‘아니다’의 어간이나 어미 ‘-으시-’, ‘-더-’, ‘-으리-’ 뒤에 붙어)) 「1」까닭이나 근거 따위를 나타내는 연결 어미. 「2」((‘아니다’의 어간 뒤에 붙어))앞 절의 내용이 뒤 절의 내용과 대조적임을 나타내는 연결어미. ¶그 일은 철수가 한 게 {아니라} 영희가 한 거야.	뜻풀이 추가
뭐	[ㅣ]「대명사」 「1」‘무어01[ㅣ]「1」’의 준말. 「2」‘무어01[ㅣ]「2」’의 준말.	[ㅣ]「대명사」 「1」‘무어01[ㅣ]「1」’의 준말. 「2」‘무어01[ㅣ]「2」’의 준말. 「3」((주로 ‘-지 뭐야’, ‘-지 뭐예요’, ‘-지 뭐니’, ‘-지 뭡니까’ 꼴로 쓰여))이미 일어난 뜻밖의 일이나 상황에 놀라거나 후회하거나 아쉬워하면서 그것을 강조함을 나타내는 말. ¶제가 글쎄 오늘 지각했지 {뭐예요}./그 여자, 얼마나 예쁜지 내가 한눈에 반했지 {뭐야}./오늘 준비물을 깜빡 잊었지 {뭐니}.	뜻풀이 추가
운항하다	【…을】 배나 비행기가 항로를 따라 다니다. ¶그 선장은 새로운 항로를 {운항할} 계획을 세웠다./안개가 심하게 낀 날은 여객기를 {운항하지} 않는다.	[1] 【(…을)】 배나 비행기가 정해진 항로나 목적지를 오고 가다. ¶그 섬에 가려면 하루에 두 번 {운항하는} 여객선을 이용하면 된다./정부가 비행을 허가할 노선은 이전의 여객기가 {운항했던} 노선과 거의 같다. ‖ 그 선장은 새로운 항로를 {운항할} 계획을 세웠다./안개가 심하게 낀 날은 제주를 {운항하는} 여객기가 뜨지 않는다. [2] 【…을】 배나 비행기 따위를 운용하다. ¶화물선을 {운항하다}/에이 380을 {운항하다}.	뜻풀이 추가

이나[02]	[1] 「1」((받침 있는 체언이나 부사어, 합성 동사의 선행 요소 따위의 뒤에 붙어))마음에 차지 않는 선택, 또는 최소한 허용되어야 할 선택이라는 뜻을 나타내는 보조사. 때로는 가장 좋은 것을 선택하면서 마치 그것이 마음에 차지 않는 선택인 것처럼 표현하는 데 쓰기도 한다. 「2」((받침 있는 체언이나 부사어 뒤에 붙어))마치 현실의 것인 양 가정된 가장 좋은 선택이라는 뜻을 나타내는 보조사. 빈정거리는 뜻이 드러난다. 「3」((수량이나 정도를 나타내는, 받침 있는 체언이나 부사어 뒤에 붙어))수량이 크거나 많음, 혹은 정도가 높음을 강조하는 보조사. 흔히 놀람의 뜻이 수반된다. 「4」((수량이나 정도를 나타내는, 받침 있는 체언이나 부사어 뒤에 붙어))수량이나 정도를 어림잡는 뜻을 나타내는 보조사. 「5」((수량의 단위나 정도를 나타내는, 받침 있는 말 뒤에 붙어))많지는 않으나 어느 정도는 됨을 나타내는 보조사. 「6」((받침 있는 체언이나 부사어 뒤에 붙어))여러 가지 중에서 어느 것을 선택해도 상관없음을 나타내는 보조사. 맨 뒤에 나열되는 말에는 붙지 않을 때도 있다. 「7」((받침 있는 체언이나 부사어 뒤에 붙어))비교의 뜻을 나타내는 보조사. 뒤 절에는 결국 같다는 뜻을 가진 말이 온다.	[1] 「1」((받침 있는 체언이나 부사어, 합성 동사의 선행 요소 따위의 뒤에 붙어))마음에 차지 않는 선택, 또는 최소한 허용되어야 할 선택이라는 뜻을 나타내는 보조사. 때로는 가장 좋은 것을 선택하면서 마치 그것이 마음에 차지 않는 선택인 것처럼 표현하는 데 쓰기도 한다. 「2」((받침 있는 체언이나 부사어 뒤에 붙어))마치 현실의 것인 양 가정된 가장 좋은 선택이라는 뜻을 나타내는 보조사. 빈정거리는 뜻이 드러난다. 「3」((받침 있는 체언이나 부사어 뒤에 붙어))어떤 대상이 최선의 자격 또는 조건이 됨을 뜻하는 보조사. ¶{사월이나} 되니까 날씨가 따뜻하다./우리 {선생님이나} 되니까 네 사정 보고 배려해 주시는 거지. 「4」((수량이나 정도를 나타내는, 받침 있는 체언이나 부사어 뒤에 붙어))수량이 크거나 많음, 혹은 정도가 높음을 강조하는 보조사. 흔히 놀람의 뜻이 수반된다. 「5」((수량이나 정도를 나타내는, 받침 있는 체언이나 부사어 뒤에 붙어))수량이나 정도를 어림잡는 뜻을 나타내는 보조사. 「6」((수량의 단위나 정도를 나타내는, 받침 있는 말 뒤에 붙어))많지는 않으나 어느 정도는 됨을 나타내는 보조사. 「7」((받침 있는 체언이나 부사어 뒤에 붙어))여러 가지 중에서 어느 것을 선택해도 상관없음을 나타내는 보조사. 맨 뒤에 나열되는 말에는 붙지 않을 때도 있다. 「8」((받침 있는 체언이나 부사어 뒤에 붙어))비교의 뜻을 나타내는 보조사. 뒤 절에는 결국 같다는 뜻을 가진 말이 온다.	뜻풀이 추가
그해	「1」앞에서 이미 이야기하여서 말하는 이와 듣는 이가 알고 있는 과거의 어느 해. ≒당세[02]「2」. ¶윤	말하는 이와 듣는 이가 알고 있거나 말하는 이만 알고 있는 과 거의 어느 해. ≒당세[02]「2」. ¶윤씨 부인	뜻풀이 수정

	씨 부인이 절에서 돌아온 {그해} 동짓달. 나이 열세 살에 치수는 성례를 치르었다.≪박경리, 토지≫ 「2」말하는 이가 이야기하고자 하는 과거의 어느 해. ≒당세[02]「3」. ¶어느 날 밤이었소. {그해} 내가 학교 가게 될 것이란 말을 듣고 있었으니까 일곱 살 되던 봄이었죠.≪이병주, 행복어 사전≫	이 절에서 돌아온 {그해} 동짓달. 나이 열세 살에 치수는 성례를 치르었다.≪박경리, 토지≫/어느 날 밤이었소. {그해} 내가 학교 가게 될 것이란 말을 듣고 있었으니까 일곱 살 되던 봄이었죠.≪이병주, 행복어 사전≫	
깔아뭉개다	「1」무엇을 밑에 두고 세게 누르거나 지나면서 누르다.	「1」무엇을 밑에 두고 짓이겨질 정도로 세게 누르다.	뜻풀이 수정
디엠지	→ 디엠제트(DMZ).	=비무장 지대「2」.	뜻풀이 수정
비속어	=속어(俗語)「1」.	격이 낮고 속된 말.	뜻풀이 수정
아슈하바트	중앙아시아 투르크메니스탄에 있는 도시. 카라쿰 사막 서쪽의 이란 국경에 가깝고 트란스카스피안(Trans Caspian) 철도가 걸쳐 있는 요충지로, 직물·식료품 가공업이 발달하였다. 투르크메니스탄의 수도이다.	'아시가바트'의 옛 이름.	뜻풀이 수정
올리다	[2]「9」『컴퓨터』컴퓨터 통신망을 통하여 파일이나 자료를 전송하다.	[2]「9」『컴퓨터』컴퓨터 통신망이나 인터넷 신문에 파일이나 글, 기사 따위를 게시하다.	뜻풀이 수정
이[32]	「1」((일부 형용사 어간 뒤에 붙어))부사를 만드는 접미사. ¶{많이}/{같이}/{높이}.	「1」((일부 형용사 어근이나 어간 뒤에 붙어))부사를 만드는 접미사. ¶{깊숙이}/{수북이}/{끔찍이}/{많이}/{같이}/{높이}.	문법 정보 수정
정성[08]	(停聲)	(丁聲)	원어 수정
정후은성	(停後殷聲)	(丁後殷聲)	원어 수정