

한국어교육에서 일반 말뭉치 활용의 확대를 위한 제언*

고경태(고려대)

<목 차>

1. 서론
2. 한국어교육 내에서의 일반 말뭉치 활용 연구의 현황
3. 일반 말뭉치 활용에 따르는 문제
4. 일반 말뭉치의 활용 확대를 위한 제언
5. 결론

<국문초록>

이 글에서는 한국어교육에서 말뭉치의 활용이 좀 더 확대될 필요가 있음을 주장하고, 이를 위해 필요한 것들이 무엇인지를 다루었다. 말뭉치가 한참 각광을 받을 당시에는 말뭉치의 효용이 크게 기대되었으나 실제로 한국어교육에서 이를 활용한 연구는 적은 편이다. 그 이유에 대해 본고는 첫째, 말뭉치 활용을 위한 적절한 응용 프로그램(문맥 색인 프로그램)의 부재, 둘째, 말뭉치 자체에 대한 오해 또는 올바른 인식의 부재를 꼽았다. 그리고 말뭉치를 통한 한국어교육 연구의 활성화에 있어서 다양한 정렬 방식을 지원하는 문맥 색인기

* 이 글은 필자가 평소에 한국어교육과 관련된 말뭉치 관련 연구를 눈여겨보면서, 말뭉치 활용 연구에 관하여 일반적으로 과대평가와 과소평가가 동시에 존재하는 데 대해 깊이 고민하던 중에 작성되었다. 필자의 문제제기에 유익한 비평과 조언을 주신 익명의 심사 위원들께 깊이 감사를 드린다.

가 개발될 필요가 있다는 것과, 말뭉치는 무조건 커야 한다든지 구어 말뭉치만이 유용하다든지 하는 생각 대신, 연구 주제와 범위에 맞게 균형성이 있는 말뭉치를 이용하는 것이 중요하다고 강조하였다.

주제어 : 코퍼스, 세종 코퍼스, 일반 말뭉치, 직관, 균형성

1. 서론

본고는 한국어교육 연구 논의 중 말뭉치 관련 연구들을 살핍으로써, 말뭉치가 기존 논의에서 어떻게 활용되었는지를 고찰하고, 교육 내용의 발굴과 점진 측면에서 일반 말뭉치(general corpus)¹⁾ 활용의 확대 제안하는 데 목적이 있다.

주지하다시피 말뭉치를 이용한 연구는 실제 자료(authentic data)를 통해 연구자의 직관을 보조하고, 더 나아가 직관으로 닿지 않은 언어 사용의 실재를 보여준다는 점에서 오래 전부터 각광을 받아 왔다. 외국인을 위한 한국어교육은 한국어에 관한 직관이 없는 외국인들을 대상으로 하여 이상적으로는 한국어 원어민 수준의 한국어 직관을 갖게 한다는 목표를 갖고 있다. 그런 점에서 실제 한국어 모습을 그대로 보여주고 있는 말뭉치는 한국어교육계에서도 일찍이 높은 활용도가 기대돼 왔다.

그러나 막상 한국어 교사의 입장에서, 혹은 한국어 학습자의 입장에서 말뭉치를 이용한 실질적인 사례나 응용을 체험하기는 어려운 것 같다. 아마도 대개의 한국어교육 연구자나 교사들에게 있어, 말뭉치란 아직도 쉽게 다가서기 어려운 대상으로 인식되고 있을지도 모른다. 한국어교육에서 말뭉치의 활용에 대해 본격적으로 논의하기 시작한 것이 2000년을 전후한 시기라고 보면, 불과 10년밖에 되지 않는 현 시점에서 교실이나 기타 교육 현장에서 체험할 수 있는 성과를 기대하는 것은 성급한 일일 수도 있다. 그러나 다른 한 편에서 볼 때,

1) ‘일반 말뭉치’란 특정한 연구 목적에 따라 적절히 균형성을 갖추고 있는 말뭉치를 일컫는 것으로(McEnery et al. 2006:59), 적절한 균형성을 갖추고 있다는 점에서는 ‘균형 말뭉치’(balanced corpus)라고도 할 수 있다. 여기서는 학습자 말뭉치(learner corpus)와 상대되는 개념으로서 모국어 화자의 발화를 연구 목적에 따라 균형적으로 담은 말뭉치를 뜻하는 것으로 한다.

좀 더 ‘실제적인 한국어’를 가르치려고 노력하는 연구자들과 교사들에게 ‘실제 자료의 보고’인 말뭉치의 활용도는 좀 더 제고될 필요가 있어 보인다.

본고는 이러한 문제제기에서 출발하여, 한국어교육에서 일반 말뭉치의 활용의 확대와 관련하여 몇 가지 기본적인 논의를 전개해 보고자 한다. 이를 위해 우선적으로 기왕에 이뤄진 말뭉치 관련 연구에서 보이는 흐름들을 살피고, 말뭉치 연구가 활성화되기 위한 기본적인 방향을 모색해 보고자 한다. 이는 앞으로 한국어교육에서 일반 말뭉치가 갖게 될 역할과 의의를 확고히 한다는 점에서 의의가 있으리라 생각된다.

2. 한국어교육 내에서의 일반 말뭉치 활용 연구의 현황

한국어교육에서 말뭉치의 연구가 본격화된 것은 강현화(2000), 유석훈(2001), 서상규 외(2002) 등을 전후로 하고 있다고 생각된다. 이 시기에는 주로 말뭉치 연구의 의의와 효용 등 일반론적인 논의들이 두드러지며 특히 학습자 말뭉치(learner corpus) 구축의 필요성이 주요하게 부각되고 있다. 일반 말뭉치의 교육적 효용에 대해서는 강현화(2000:59)에서 ‘단순한 문법 지식으로서 는 설명이 부족한 어휘의 공기 관계 등에 유의한 어휘 교육이 가능’하다고 하여, 일반 말뭉치 활용의 의의를 밝히고 있음을 볼 수 있다.

그러나 오늘날에 이르기까지 양적인 측면에서 일반 말뭉치 연구가 활기를 띠고 있다고 말하기는 어렵다. 대략적으로 본다면, 현재로서는 실제 한국어교육에 유용한 교육 내용 발굴 및 점검에 무게를 두고 있는 듯하나, 그마저도 활발한 양상을 띠고 있다고 보기는 어려워 보인다. 이런 이유로 현재로서는 각종 연구들 속에서 특정한 흐름을 정확히 짚어낼 수는 없으나, 대략적으로 세 부분으로 가름해 볼 수는 있다. 우선 말뭉치를 이용하여 유의 관계에 있는 어휘의 용례를 분석하고 있는 연구로서 봉미경(2005)과 유현경(2007), 강현화(2005) 등을 꼽을 수 있고, 특정 어미나 어휘와의 호응 관계를 이루는 부사를 다룬 연구들로서, 강현화(2000), 임유종(2007) 등을 들 수 있다. 마지막으로 Sinclair(1999)의 연어(collocation) 개념이나 Lewis(2001)의 어휘적 접근법(lexical approach) 등의 개념에 기반한 논의라 할 수 있는데, 이는 한송화·강현화(2004) 등에서 소개돼 있으며, 이와 유사한 관점에서의 구절론(phraseology)

의 측면까지 고려하면 강현화(2005) 등도 포함된다. 최근에는 연구 대상이 좀 더 확장되는 경향도 보이는데, 고경태(2007)는 명사화소 ‘-음’과 ‘-기’를, 홍운혜(2007)는 보조사 ‘-(이)나’와 ‘-(이)나마’에 관한 연구를 선보이고 있고, 변우영(2007), 강현화(2008) 등은 화행 분야로까지 연구 범위를 확장하고 있다.

이상의 선행 연구를 좀 더 깊이 들여다보면 몇 가지 특징적인 점을 발견할 수 있다. 첫째, 주로 연구 대상이 되는 한국어 형태나 어휘를 미리 선정해 놓은 연구가 많다는 점이다. 이것은 말뭉치 활용이 주로 용례 검색 중심으로 이뤄지고 있다는 점과 관련이 깊어 보인다. 다시 말해, 미리 연구 대상이 되는 한국어 어휘나 형태를 선정하여 놓고, 그의 용례를 검색하여 교육 내용의 점검과 정밀화를 기하고 있는 논의가 많다는 것이다. 둘째, 연구 대상이 되는 말뭉치 선택에서 대개 21세기 세종계획 기초 말뭉치(이하 ‘세종 말뭉치’) 특히 1천만 어절의 교육용 말뭉치가 사용되고 있는데, 형태소 정보라든지 구문 정보 등을 부착하지 않은 순수한 발화 자료 그대로의 원시 말뭉치(raw corpus)가 사용되고 있다는 점이다. 그리고 연구마다 조금씩 차이는 있으나 말뭉치의 신뢰도를 논함에 있어 문어와 구어의 비율을 고려하는 경우가 많다는 점도 꼽을 수 있다.

그런데 이러한 모습은 말뭉치 활용이 가져다주는 효용의 일부일 뿐이어서, 말뭉치 활용이 본격적인 궤도에 이른 상태라고 보기에는 무리가 있다. 용례를 검색하고 간단히 빈도를 구하는 것 외에도 말뭉치를 통해 유의미한 정보를 더 얻을 수 있으며, 꽤 오래 전부터 배포되고 있는 ‘형태소분석기’를 사용하면 원시 말뭉치보다 좀 더 정교한 검색을 가능케 하는 형태 분석 말뭉치를 만들어 사용할 수 있다.²⁾ 후술하겠지만, 여러 사항들을 다각도로 검토한다면 일반 말뭉치의 활용은 지금보다는 좀 더 확장될 수 있다. 그러나 현 단계에서는 한국어교육을 위한 일반 말뭉치 활용에 있어 제한이 따르는 것도 사실이다.

2) ‘지능형형태소분석기’는 21세기 세종계획 공식 홈페이지에서 내려 받을 수 있다. 완성형 텍스트 파일을 읽어 들여 손쉽게 형태 분석을 해 주며, 변환된 형태 분석 말뭉치는 한 행에 하나의 어절을 기록하는 수직 말뭉치 형태를 띤다.

다만, 이 프로그램을 통한 형태소 분석은 다소 많은 오류를 내포하고 있다. 형태소 분석 시 지나치게 고빈도 동사로 간주하지 않고 ‘흔치 않은’ 명사로 분석하려는 경향이 있는 것도 문제다. 또한, 1음절로 구성된 어절이나 용언의 활용형의 경우 우연히 다른 단어와 형태가 같아지는 경우가 있어서 수작업을 통해 오류를 고쳐 주어야 하는 문제가 있다. 그렇지만 오류를 적절히 수정한다면 꽤 쓸 만한 형태 분석 말뭉치를 얻을 수 있다. 이를 다운로드 받으려면 http://www.sejong.or.kr/dist_frame.php 접속.

3. 일반 말뭉치 활용에 따르는 문제

차재은(2003:423)에 따르면 실제적인 말뭉치 활용이 활성화되지 못하는 이유에 대해서, 일선 교육자들이 말뭉치를 직접 이용하기에는 제한이 있다는 점과, 제한된 자료이나마 이에 대한 소개가 체계적으로 이뤄지지 못하고 있는 점을 꼽았다. 전자의 문제는 말뭉치 조사 연구에 필요한 컴퓨터 프로그램 등의 도구(tool) 및 이를 실제적으로 운용할 수 있는 전산적 지식을 가리킨다면, 후자는 한국어교육에서 필요한 말뭉치 자료의 선택과 관련이 있다. 오늘날의 일반 말뭉치 활용이 갖는 한계를 차재은(2003)과 비슷한 관점에서 살핀다면 다음과 같다.

먼저 말뭉치에 접근할 수 있는 프로그램, 즉 문맥 색인기(concordancer)와 관련되는 문제이다. 문맥 색인기란 사용자로부터 입력 받은 검색어(대상어)를 중심으로, 말뭉치에서 해당 검색어가 사용된 문맥을 보여주는 프로그램을 일컫는다. 이렇게 만들어진 것을 문맥 색인(concordance)이라 하는데, 문맥 색인은 단순히 검색어에 대한 실제 용례이기도 하지만, 규모화된 말뭉치에서 추출된 경우에는 특별한 의미를 갖기도 한다.

그러나 선행 논의에서는 물론 가장 널리 쓰이는 ‘글잡이2’는 일정 규모를 넘는 말뭉치에서 검색할 때에는 동일한 검색어임에도 불구하고 매번 조금씩 다른 용례를 출력하기도 하며, 용례 검색 빈도도 조금씩 다르게 나타나는 문제가 있다. 이를 대체할 만한 것으로는 SynKDP(‘깎작새’)를 들 수 있는데, 현재로서는 가장 정확한 문맥 색인 결과를 보여주나 공식적으로 배포하지 않고 있어 알음알음으로 구해야 하는 문제가 있다. 이러한 문제들은 분명 말뭉치 활용에 한계를 갖는 부정적인 원인이라 할 만하다.

최근에는 글잡이2를 대신하여 ‘Hanmaru’라는 문맥 색인 프로그램이 배포되고 있다. 그러나 형태 분석 말뭉치를 이용하여 검색할 때에 고작 ‘한 개 형태소’에 대한 용례 검색만을 지원할 뿐이어서 실제적으로 형태 분석이 되어 있지 않은 말뭉치를 대상으로 용례를 검색할 때보다 나은 점이라고는 고작 ‘중의성 어절’을 해소해 줄 수 있다는 점 정도뿐이다. 형태 태그를 부착한 말뭉치를 가지고 일반적으로 이용할 수 있는 것이 제한돼 있다는 점은 분명 문제이다. 또 한 Hanmaru는 검색어(대상어)가 한 문장에 두 개 이상 존재할 때 그중 가장

먼저 있는 하나만을 보여 준다는 문제도 안고 있다.³⁾

말뭉치 활용의 한계와 관련된 둘째 문제는 말뭉치 자체에 대한 이해도가 전반적으로 높지 않은 현 상황과 관련이 깊다. 말뭉치에 대한 이해도가 낮다는 점은 기존 연구에서 대체로 말뭉치의 선정과 관련하여 문어와 구어의 비율 문제 정도만을 언급하고 있다는 데에서 추론이 가능하다. 말뭉치의 선정이나 구축에서 제일 먼저 고려할 것은 말뭉치의 규모도 아니고 말뭉치에 반영된 문어와 구어의 비율도 아니기 때문이다. 물론 이런 문제가 발생한 데에는 기존에 말뭉치의 활용법이 폭넓게 소개되지 못한 것과도 관련이 있어 보인다. 연구 초기에 말뭉치의 활용이 갖는 의의나 효용이 소개되는 일은 많았지만, 정작 말뭉치가 가져야 하는 조건이나 말뭉치 연구에서 유의할 점 등, 말뭉치 연구 및 말뭉치 연구의 이해에서 선결돼야 하는 지식들이 제대로 소개되지 못했다는 것이다.⁴⁾ 연구자들이 말뭉치를 용례 검색을 위한 도구 정도로만 활용하고 있는 것도 이러한 맥락에서 이해될 수 있다.

이에 덧붙여 세종 말뭉치에 대한 부정적 인식도 거론할 만하다. 국가적인 사업으로 구축된 세종 말뭉치에 대해 많은 연구자들이 부정확하고 오류가 많아 이용하기 어렵다는 생각을 갖고 있는 것 같다. 이론적 정합성과 정교함을 강조하는 패러다임에서는 사소한 오류도 문제가 될 수 있다. 그러나 그 불신이 말뭉치를 통한 연구 그 자체까지 확대되는 것은 아닌지 우려가 될 때도 있다. 물론 세종 계획에서 배포한 프로그램들에서도 오류가 발견되고 있고, 세종 계획이 완료된 이후로는 아무 디버깅도 이뤄지지 않고 있다는 것도 문제이다. 형태 분석 말뭉치를 만드는 과정에서 중의성을 보이는 수많은 어절들에 대한 수작업이 이뤄졌는데 그 수작업마저도 틀릴 수 있다는 것은 특히 말뭉치에 대해 기대를 갖고 바라보던 이들에게는 매우 아쉬운 부분일 수 있다. 문제는 이러한 경향이 말뭉치의 활용에 대한 기대를 꺾는 데 한몫을 하고 있다는 것이다. 어찌면 위와 같은 문제들로 인해 말뭉치 활용의 긍정적 기대보다는 말뭉치 활용

3) 본고에서 시험적으로 만든 말뭉치를 통해 조사한 것으로 총 16회 발견된 단어가 Hanmaru에 서는 총 9회만 발견되어 조사하던 중 발견하게 된 사실이다.

4) 이는 어쩌면 초기의 말뭉치 관련 논의들이 대규모 말뭉치를 구축하는 단계에서 이뤄졌기 때문일지도 모른다. 대규모 말뭉치는 많은 비용과 인력을 투입해야 하는 만큼, 당시에 말뭉치 관련 지식은 일반에 폭넓게 공유되어야 할 필요성이 없었을지도 모른다. 확실히 이 당시의 말뭉치 활용론이 규모화된 말뭉치를 겨냥하고 있는 것도 이와 무관치 않다.

에 대한 우려가 커지고 있는 것은 아닌지 걱정스럽다.

4. 일반 말뭉치의 활용 확대를 위한 제언

4.1. 쉽게 이용할 수 있는 말뭉치 접근 도구의 개발

말뭉치 활용이 확대되기 위한 그 첫 번째는 말뭉치를 이용하여 자신이 원하는 결과를 탐색 및 검토할 수 있는 응용 프로그램이 마련되어야 한다는 것이다. 말뭉치에 담긴 언어 자료들의 양과 상관없이 이용할 수 있는 프로그램이 필요하다는 것이다. 또한 기왕의 글잡이² 등과 같은 문맥 색인기와 같이 기존에 구축된 말뭉치는 물론, 연구자 스스로 만든 D.I.Y. 말뭉치도 탐색할 수 있는 문맥 색인기가 만들어져야 함은 물론이다.

문맥 색인 프로그램이 절실한 까닭은, 이것이 없는 경우 말뭉치 이용에 늘 제약이 따르기 때문이다. 1천만 어절이 넘는 세종 말뭉치를 일반 뷰어나 텍스트에디터 프로그램으로 열어서 용례를 검색하기란 만만치 않음은 말뭉치를 가지고 연구를 해 본 이라면 누구나 공감하는 사실일 것이다. 시간도 오래 걸릴뿐더러, 그 결과의 정확성을 기하기도 어렵다.

말뭉치를 통한 연구는 단순히 텍스트 문서를 열어 단어나 표현을 검색하는 것과는 다른 작업이므로 애초부터 텍스트에디터나 워드프로세서 등의 프로그램으로 접근하기가 어렵다. 말뭉치는 일반적으로 기대되는 텍스트의 양보다 더 많은 양의 자료를 대상으로 검색을 하기 때문에, 수많은 용례 속에서 내재하는 질서나 체계, 규칙 등을 찾는 데 도움이 된다. 이러한 작업을 일반 텍스트에디터 등으로 하게 된다면 후처리, 후분류라는 대단히 수고로운 작업이 뒤따른다.

텍스트에디터 등, 일반적으로 사용하는 응용 프로그램을 이용하여 문맥색인을 만드는 법을 일러주고 있는 연규동 외(2003)가 있기는 하다. 그러나 이러한 방법은 절차가 복잡하여 컴퓨터에 익숙하지 않은 이들에게는 제한이 따를 수 있다. 더욱이 이런 종류의 지식은 컴퓨터와 각종 응용 프로그램에 익숙한 이가 아니면 도저히 생각해 낼 수 없는 성질의 것이다. 어떤 면에서 말뭉치 활용 연구가 소수의 연구자들에게 국한돼 보이는 것도 이런 데에서 말미암는다고 할 것이다.

최소한 세종 말뭉치와 같이 국가적으로 만들어진 경우에는 폭넓은 연구자들이 의해 다방면으로 활용될 필요가 있을 것이다. 현재의 Hanmaru는 이미 그 용도를 ‘검색’ 정도로 제한한 듯한 인상을 준다. 기존에 알려진 SynKDP, 그리고 예전 세종계획에서 개발된 글잡이²를 보고 좀 더 개선됐기를 기대하는 입장에서 Hanmaru를 다뤄 보면 실망스러운 면이 없지 않다. 애초부터 활용을 ‘검색’ 정도로 제한하지 말고, 검색한 것들을 SynKDP처럼 다양하게 ‘정렬’할 수 있는 옵션을 두었으면 좋았을 것이다. 모쪼록 여러 연구자들이 쉽게 접할 수 있도록 프로그램의 완성도가 현재 상태보다는 높아져야 한다는 것만큼은 확실하다.

4.2. 활용도에 대한 인식 확대

일찍이 한국어교육에서 말뭉치의 효용을 논한 서상규 외(2002:129~30)에서는 한국어 모어 화자들의 발화를 담은 일반 말뭉치(general corpus)에 대해 ‘학습의 대상이 되는 표준적인 한국어’로서의 의미를 부여하고, ‘한국어 교재나 학습 사전의 개발, 교육용 기초·기본 어휘 선정을 위한 어휘 빈도 조사, 어휘 중요도 평정 등’에 쓰일 수 있음을 보이고 있다.⁵⁾ 또한, 한송화(2000:59)와 변우영(2007:151)에서도 한국어 학습자가 실제적인 용례를 충분히 접하면서 한국어를 배울 필요성이 있으며, 여기서 말뭉치 활용은 큰 역할을 할 것으로 기대하는 모습을 볼 수 있다.

일반 말뭉치를 통해 기대되는 효용은 크게 세 가지로 구분해 볼 수 있다. 먼저, 각급 한국어교재에서 문법 또는 어휘를 가르치기 위해 제시하는 용례들을 좀 더 실제에 가깝게 만드는 데 기여할 수 있다. 변우영(2007)에서 ‘감사에 대한 대답’과 관련하여 한국어 교재에 나타난 것과 한국어 모어 화자의 말뭉치 속에서 나타난 것을 대조함으로써 일부 교재가 실제성에서 다소 차이를 보이는 것을 지적한 것은 바로 말뭉치의 일반적 활용이라 할 수 있는 용례 검색과 관련된다. 또한, 기존에 잘 알려져 있는 여러 어휘적 덩어리들, 가령 ‘-에 대하여’와 같은 경우 실제 교육 시에 제시되어야 할 기본 형태는 어떤 것이 좋을지 검증해 볼 수 있다. 일반적으로 동사나 형용사에 후행하는 어미로서 가장 많은

5) 서상규 외(2002)에서는 ‘표준 말뭉치’라고 언급하고 있으나 그 본질은 본고의 일반 말뭉치와 다르지 않다.

빈도를 차지하는 것은 관형격 어미 ‘-ㄴ’인 것으로 알려져 있다(김홍규·강범모 2000:93). 따라서 ‘-에 대해서’보다 ‘-에 대한’의 빈도가 더 높다면, 지금까지 교육에서 기본형으로서 ‘-에 대한’을 우선적으로 고려해 볼 수 있을 것이다.

다른 한 가지는 한국어교육의 내용을 좀 더 풍부하게 구성할 수 있다는 데 있다. 문법 교육을 진행하면서 예측하지 못한 학습자의 오류들을 발견하기란 어렵지 않은 일이다. 어떤 경우에는 복잡한 규칙으로 제시하기보다는 ‘자주 어울려 쓰이는 용례’들을 우선적으로 가르치는 것을 고려해 볼 수 있으며, 또한, 용례에 대한 정밀한 관찰을 통해 문법 규칙을 좀 더 정교하게 만들 수도 있다. 이를테면 추측이나 회의를 나타낸다고 알려진 ‘-나 싶다’는 ‘이 옷이 크지 않나 싶다’와 같이 부정의 ‘-지 않(았)’ 또는 ‘추측’의 ‘-겠-’과 과 자주 어울려 쓰이는 것으로 나타나 있으나(국립국어원 2005:113), 실제로 ‘-겠-’이 쓰이는 경우는 적어 보였고, ‘-지 않(았)나 싶다’와 좀 더 많이 쓰이거나 유사한 빈도를 보이는 것으로 ‘명사-이/가 아니었나 싶다’가 있다. 아래는 1천만 어절 세종 형태 분석 말뭉치에서 추출한 ‘-아니었나 싶다’와 ‘-이었나 싶다’의 용례이다.⁶⁾

<그림 1> ‘명사-아니었나 싶다’의 용례 (일부)

80년에 들어서가	아니었나	싶다.
가는 날 같은 것이	아니었나	싶다.
핑계일 뿐이	아니었나	싶다.
가장 행복한 순간이	아니었나	싶었다.
차갑고 무관심한 곳이	아니었나	싶었는데
말은 바로 이런 뜻이	아니었나	싶다.

6) 본고에서는 말뭉치 활용의 의의를 보이는 데 목적이 있으므로, 본문에 제시된 각종 문맥 색인을 추출한 절차에 대한 상술은 피하고자 한다. 다만 본문의 문맥 색인들은 세종 형태 분석 말뭉치에 대해 소정의 가공을 가한 것이며, 이러한 가공법은 각종 세종 말뭉치 보고서에서 찾을 수 있음을 밝혀 둔다.

한편, 익명의 심사위원으로부터 위 말뭉치에서 ‘아니었나’의 ‘-나’ 어미는 종결어미인 ‘EF’가 부착되어 있어야 하지 않는가고 질의를 받았다. 필자가 아는 한, EF 태그는 일괄적으로 종결 부호인 마침표, 물음표, 느낌표 앞에서만 붙고 내부절의 종결어미에는 EC, 즉 연결어미로 보고 있는 것으로 알고 있다.

말뭉치 태그 자체에 대한 논쟁은 말뭉치 활용 측면에서는 다소 유보할 수 있다. 내부절의 서술어 말미에 부착된 어미를 종결 어미로 볼 것인지 연결 어미로 볼 것인지와 상관없이 그 태그가 ‘일정한 규칙대로’ 부착되었는가가 중요한 것이다. 일정한 규칙대로 부착되었다면 그들을 모두 걸러내는 일이 가능하지만, 그렇지 않다면 일일이 수작업을 통해 걸러야 하는 번거로움이 수반되기 때문이고, 그만큼 말뭉치의 활용에 치명적인 문제가 될 수 있기 때문이다.

참으로 멋진 사람이	아니었나	싶다.
때에 나온 말이	아니었나	싶다.
쑥맥이	아니었나	싶습니다.
바로 존재의 기쁨이	아니었나	싶습니다.
안이한 임전태세가	아니었나	싶다.
충실한 대리 역이	아니었나	싶기도
저의 특수한 사정이	아니었나	싶다.
먹던 것은 역시 알사탕이	아니었나	싶다.

이밖에도 ‘명사-이/가 있나 싶다’ 역시 비슷한 빈도로 발견된다. 이는 ‘뭐 저런 사람이 다 있나 싶어’와 같은 용례에서 보이는 것으로서 아래에 그 일부를 보이도록 한다.

<그림 2> ‘명사-이/가 있나 싶다’의 용례(일부)

자신감에 차 있을 수	있나	싶어서
뭐 이런 남자가 다	있나	싶어서
저런 무식한 남자가 다	있나	싶어,
그래도그럴 수가	있나	싶어
하나라도 억울한 이가	있나	싶어
시골에 아직도 이런 구멍가게가	있나	싶을
서울에 이런 곳이	있나	싶을
홍대 앞에도 이런 곳이	있나	싶을
이렇게 많은 교향악단이	있나	싶을 것이다.
나는 모아 둔 돈이	있나	싶었지.
성현이에게 그런 모습이	있나	싶어
내가사위복은	있나	싶었는데,

그간에는 ‘-나 싶다’의 ‘-나’가 연결어미라는 사실에 유의하여, 주로 이에 선행되는 성분으로 용언을 염두에 두었던 결과인지는 모르지만, 실제 말뭉치 용례 검색에서는 ‘명사’가 선행되었을 때의 ‘-이/가 아니(었)나 싶다’와 ‘-이/가 있나 싶다’ 역시 비슷한 빈도를 보였음을 확인할 수 있다. 이것은 새로운 문법 교육의 내용으로 추가해 볼 수 있다.

말뭉치를 통한 용례 검색은 기존의 교육 내용에 새로운 내용을 보충해 주기도 한다. 이를 테면, 각급 한국어 교육 현장에서 하나의 단어처럼 제시되는 ‘-기 (이)전에는’ 주로 동사가 앞서 위치할 때를 가리키는 것으로 교육되고 있는

것 같다. 「한국어 문법」2에서도 물론, 대부분의 한국어 교육 사전류에서 이들은 한결같이 ‘-기 전에’의 앞뒤 사건의 시간적 선후 관계만을 서술하고 있기 때문이다. 한국어 모국어 화자의 직관으로도 ‘-기 전에’는 역시 시간적 선후 관계에 있는 사건의 연결과 관계가 있어 보인다. 그러나 다음의 말뭉치 결과는 좀 더 다른 ‘-기 전에’의 면을 보여 준다.

<그림 3> ‘-기 전에’의 용례 (일부)

단짝 친구이	기 전에	건달이야.
문학은 글이	기 전에	말이며, 길쌈이고 농사
정 총리는 총리이	기 전에	교수이다.
총리이	기 전에	대학 강단에 선
역사적 서명 그것이 정부이	기 전에	북한의 내면적 계략에
여성이	기 전에	인간이며 인간은 누구나
이 세국은 조정의 곡식이	기 전에	나라의 국록을 먹는
우리는 우리 무대의 일원이	기 전에	먼저 고등학생이다.
하지만 저 사람은 남자이	기 전에	아픈 사람 아닌가.
콤플렉스에서 벗어날 것이며, 여자이	기 전에	인간으로 살아갈 것이다.
여자이	기 전에	인간으로 살기 위해서
여자이	기 전에	인간으로 우리 사회의
학문은 경쟁이	기 전에	협동이다.
동맹자로 규합할까 라는 전술의 문제이	기 전에	, 노동자 계급을
윤리 도덕에 관한 문제이	기 전에	애정에 관한 문제라
한국의 수치이	기 전에	실로 일본의 비인도
대한 인과론 법칙은 법칙이	기 전에	명제이고 주장이다.
말로라든가 지드이	기 전에	, 카뮈와 썬르트트리기에

‘-기 전에’ 앞에 ‘명사+이다’가 연결된 용례도 제법 많이 발견된다. 가령 ‘그녀는 장관이기 전에 한 아이의 엄마이다’와 같은 용법이 그것이다. 말뭉치는 이처럼 직관으로 쉽게 닿지 않는 상세한 용례를 발견하는 데도 적잖은 도움을 줄 수 있다.

이상에서 언급한 효용들은 실제 선행 연구에서도 공통적으로 인식되고 있는 것으로서, 여기에 하나를 더 추가한다면 좀 더 아직 발굴하지 못한 교육 내용을 탐색하는 데 도움이 될 수 있다는 점을 들 수 있다. 이것은 ‘한글’을 검색어로 하는 것이 아니라, 통사 정보를 입력하여 검색하는 방식에서 찾을 수 있

다. 위 <그림3>이 그와 같은 방식으로 만들어진 문맥 색인이다. 위 그림은 명사화소 ‘-기’에 명사가 후행하는 경우를 찾은 예의 일부이다.

이런 점에서 원시 말뭉치 외에도 형태 분석 말뭉치를 적극 활용할 필요가 있다. 이를테면, ‘-에 대하여’나 ‘-에 관해’ 등과 같은 문법적 연어 구성에 착안하여, ‘격조사+동사+어미’의 형태를 갖춘 모든 용례를 찾는다면, 현재 한국어 교재에서 다루지 않은 다른 유형의 문법적 연어 구성을 발견할 가능성이 높다. 다시 말해 기존의 여러 문법 항목이 주는 암시를 근거로 하여 새로운 교육 내용을 발굴할 수 있다는 것이다. 또한, 문장을 2~3개씩의 단위로 분리해 내어, 분리된 각각의 빈도를 헤아리는 데 n-gram을 이용한 말뭉치 조사 연구를 하면, 연어 관계를 보이는 형태들을 발견하는 데도 도움이 된다. 이는 형태 태그가 부착된 말뭉치를 이용한다면 충분히 가능한 일이다.

한편, 일반 말뭉치에 있어 이미 주어진 대규모 세종 말뭉치 외에도 연구자 스스로 만드는 말뭉치(D.I.Y. corpus) 연구도 적극 고려할 필요가 있다. 특히, 특정 장르별로 구축한 적정 규모의 말뭉치는 한국어교육에 필요한 다양하고 세분화된 교육 내용 구축이라는 측면에서 좀 더 적극적인 연구가 필요하리라 생각된다. 예컨대, 한국어 교실에서 흔히 보이는 미디어 수업을 위해, TV뉴스의 발화 자료를 가지고 소규모의 말뭉치를 만들고, 여기에서 자주 사용되는 어휘나 형태 또는 표현들을 발굴하는 것이나, 신문의 사회면 뉴스를 모아 말뭉치를 만들고 여기서 유의미한 교육 내용을 발굴하는 것 역시 일반 말뭉치 활용으로 기대할 수 있는 효용이라고 할 수 있다.

4.3. 말뭉치에 대한 이해 선행

연구자 스스로 개발하여 사용하는 말뭉치 연구는 아직까지 한국어교육계에서 시도된 바는 없지만 충분히 탐구할 만한 분야라 생각된다. 이때 대두되는 문제는 말뭉치의 규모와 말뭉치의 균형성의 문제이다. 이는 말뭉치 자체에 대한 선결적 논의라 할 수 있는데, 이는 그간에 충분히 소개되지는 않은 면이 있다고 본다.

현대적인 의미에서 말뭉치가 가져야 할 기본적인 요건은 크게 세 가지를 꼽는다. 첫째는 기계 가독형 자료(Machine Readable Data)여야 할 것, 둘째는 저

작권 문제가 없어야 할 것, 그리고 남은 한 가지는 연구 목적에 부합할 것 등이다(McEnergy et al. 2006:71~2). 파일 형태로 된 말뭉치를 저작권 문제없이 입수할 수 있다면, 남은 문제는 연구 목적에 맞는 말뭉치의 선택 또는 구축과 관련된다고 할 것이다.

그런데 한국어 교육에 있어서 종종 말뭉치 자체에 대한 이해의 부족에서 비롯된 오해들이 있어 보인다. 이는 크게 세 가지를 꼽을 수 있다. 첫째, ‘말뭉치 이니만큼 일단 규모가 커야 한다’, 둘째, ‘한국어교육에서 중요한 것은 문어 자료가 아니므로 말뭉치의 효용은 제한적이다’, 마지막으로 셋째 ‘아직 빈도만큼 확실한 기준은 없다’ 등이 그것이다. 그러나 이들은 결코 항진 명제가 아니다. 아래에서 이와 관련된 문제들을 살펴보자.

4.3.1. 말뭉치 규모의 문제

그간의 논의들을 보면, 대체로 한국어교육에서 논하는 말뭉치들은 대규모를 지향하고 있음을 볼 수 있다. 가령, 유석훈(2000)이나 서상규 외(2000)에서처럼 학습자 말뭉치를 대규모화하고 정교한 오류 태그를 부착한다면, 분명 교육적으로 큰 함의를 주는 결과들을 얻을 수 있을 것이다. 그러나 항상 대규모여야 할 필요는 없다. 말뭉치의 규모는 연구 목적과 더불어, 실제적인 말뭉치 조사에 필요한 고려 사항에 맞물려 결정될 성질의 것이다(McEnergy et al. 2006:71).

연구 목적이 어휘보다는 문법 면과 관련이 깊다면, 어휘 관련 연구에 비해 좀 더 작은 규모를 고려해 볼 수 있다. Biber(1993)는 영어의 경우 과거-현재-미래 시제의 빈도를 헤아리는 데 1000어절 정도의 말뭉치로도 충분한 연구 결과를 이뤄내기도 하였다. 비단 문법 형태소만이 아니라, 문법적 기능이 강한 연어인 ‘-에 대하여’나 ‘-에 따라서’ 등과 같은 문법화(grammaticalization) 과정을 겪고 있는 소위 표현 항목⁷⁾ 등은 무작위 추출된 68개의 문장에서도 어렵지 않게 검출된다. 이를 아래에 보이도록 한다.

7) ‘표현 항목’은 이미혜(2005)에서 문법 교육 항목 중에서 몇 개의 어절로 구성된 것들을 가리키는 용어로 고안되었지만, 이 용어는 마치 ‘일반 의사 표현’에 사용되는 항목까지도 아우르는 것처럼 보여서 학제간 공유할 만한 용어로는 적합하지 않아 보인다. 이 본질을 밝히기 위해서는 최소한 ‘문법 표현 항목’ 정도가 온당할 것이다.

<그림 4> 1,602어절에서 검색한 ‘격조사+동사+어미’의 문맥 색인(concordance)의 결과(일부)

89	ING+장비/NING+의/JKG 사용/NING 기준/NING 등/NNB+	예/JKB 관하/VV+ㄴ/ETM	규정/NING+예/JKB+는/JX 소방/NING+기본/NING+법/
90	VA+ㄴ/ETM 상황/NING+에서/JKB 인명/NING 구조/NING+	예/JKB 나서/VV+어아/EC	탈/IC 소방대/NING+가/JKS 어발/VA+계/EC 활거민/
91	이/JKS 소방대/NING+가/JKS 크/VA+ㄴ/ETM 길가/NING+	예/JKB 대기하/VV+도록/EC	오정하/VV+고/EC 6/SN+시/NNB+5/SN+분/NNB 건
92	VV+고/EC 일/VX+는/ETM 8/SN+가지/NNB 사할/NING+	예/JKB 대하/VV+아/EC	제대로/MAG 규명하/VV+지/EC 앞/VX+는다면/EC 우
93	+이/VCP+며/EC /SS+용산/NNP 참사/NING+사건/NING+	예/JKB 대하/VV+아/EC	검찰/NING+수사/NING+
94	ING 부트/VV+ㄴ/ETM 살인/NING 진압/NING 경위/NING+	예/JKB 대하/VV+아/EC	철저히/MAG 수사하/VV+아아/EC 용산/NNP 참사/
95	+이/VCP+며/EC /SS+용산/NNP 참사/NING+사건/NING+	예/JKB 대하/VV+아/EC	이명박/NNP 정부/NING+가/JKS 책임져야/NF 용산/
96	+이/VCP+며/EC /SS+용산/NNP 참사/NING+사건/NING+	예/JKB 대하/VV+아/EC	검찰/NING+수사/NING+
97	/NING+를/JKO 사용하/VV+있/EP+다는/ETM 의혹/NING+	예/JKB 대하/VV+아서/EC	+는/JX 진상/NING+
98	/NING+를/JKO 사용하/VV+있/EP+다는/ETM 의혹/NING+	예/JKB 대하/VV+아서/EC	+도/JX 수사/NING+담당/NING+의/JKG 밀장/NING+
99	/JKB 죽/VV+거나/EC 다치/VV+ㄴ/ETM 피해자들/NING+	예/JKB 대하/VV+아서/EC	+는/JX 가족하/VA+ㄴ/ETM 수사/NING+
100	/NING+이/JKS 경찰/NING+과/JC 총력/NING+강제/NING+	예/JKB 대하/VV+아서/EC	+는/JX 어따하/VA+ㄴ/ETM 형사/NING+처벌/NING+
101	/NING+를/JKO 사용하/VV+있/EP+다는/ETM 의혹/NING+	예/JKB 대하/VV+아서/EC	+도/JX 절차상/NING 문제/NING+
102	/JKB 죽/VV+거나/EC 다치/VV+ㄴ/ETM 피해자들/NING+	예/JKB 대하/VV+아서/EC	+는/JX 가족하/VA+ㄴ/ETM 수사/NING+
103	+이/VCP+며/EC /SS+Q/SS+지체/NING+의/JKG 진술/NING+	예/JKB 따르/VV+면/EC	+Y/SS 망루/NING+
104		이/NNP	검찰/NING+은/JX 현장/NING+
105	NING+의/JKG 거듭되/VV+는/ETM 방수/NING 요청/NING+	예/JKB 따르/VV+아/EC	검찰/NING+
106	ETM 행위/NING+로/JKB 엄격하/VA+ㄴ/ETM 절차/NING+	예/JKB 따르/VV+아/EC	국회/MAG 제한하/NING+
107	+의/JKG 소방장비/NING+를/JKO 시위/NING+진압/NING+	예/JKB 사용하/VV+ㄴ/ETM	것/NNB+은/JX 경찰관/NING+직무/NING+집행/NN
108	경찰/NING+과/JC 소방/NING+5P 인명/NING 구조/NING+	예/JKB 사용하/VV+아아/EC	할/IC 소방장비/NING+마져/JX 시위/NING 진압/NING
109	/JKB 직접/MAG 고인/NING+과/JC 유족/NING 앞/NING+	예/JKB 사직하/VV+ㄴ/ETM	것/NNB+

위 <그림 4>는 논의를 위해 본고에서 만든 말뭉치로서, 인터넷 신문에서 추출한 1,602어절의 소규모 말뭉치를 만들고, 이를 텍스트 에디터에서 정규식(regular expression)을 이용하여 ‘격조사+동사+어미’의 유형을 모두 검색한 후, 검색어를 중심으로 오름차순으로 정렬한 결과의 일부이다.⁸⁾ 굵은 선 안쪽을 보면 알 수 있듯이, 다른 유형의 ‘격조사+동사+어미’에 비해 빈도가 높아 보이는 패턴(pattern)이 다수 발견되는데, 그것은 ‘-에 대하여’에서 비롯된 ‘-에 대해, -에 대해서’ 그리고 ‘-에 따르면’과 관련된 ‘-에 따르면, -에 따라’ 등이다. 여기에는 어떤 동사가 실질적인 의미보다는 특정 격조사와 어미와 더불어 문법적 의미로 기능할 경우, 일반 동사들이 가질 수 있는 ‘격조사+동사+어미’의 패턴보다 빈도가 높을 것이라는 가정이 전제돼 있다.

일반적으로 문법 형태소 연구를 할 때에는 어휘 형태소를 연구할 때보다 더 작은 규모의 말뭉치로도 충분하다. 만일, <그림 4>와 같은 연구, 즉 신문에 나타난 문법 항목을 발굴한다는 측면에서 말뭉치 연구를 하기 위해 좀 더 규모를 확대한다면, 신문에서 자주 쓰이는 어휘들을 발굴할 때보다는 좀 더 작은 규모의 말뭉치로도 가능하리라는 예상이 가능하다. 이는 Santos(1996:11)의 언급

8) 형태 분석은 21세기 세종계획에서 배포한 ‘지능형형태소분석기’를, 텍스트 에디터는 EmEditor, 그리고 문맥 색인 정렬은 Microsoft Excel을 사용하였고, 수직 말뭉치를 수평 말뭉치로 변환하는 데에는 EXCEL과 EmEditor를 이용하였다.

처럼 ‘어휘 형태소는 장르나 텍스트 작성 시기 등 여러 가지 요인들의 영향을 받을 수 있기 때문에 규모가 클수록 좋지만, 문법 형태소는 텍스트의 장르나 특성에 상관없이 편재하기 때문에’ 가능한 것이다.

한편, 한국어교육은 교육이라는 특별한 목표를 지향하고 있는 만큼, 좀 더 다양한 연구 목적에 따라 소규모의 말뭉치를 연구자 스스로 구축해 볼 수 있다. 규모가 작아도 매우 특화된 말뭉치는 수백 만 어절의 말뭉치보다도 유용할 수 있다(Shimazumi and Berber-Sardinha 1996, McEnery et al. 2006:72 재인용). 한국어 모어 화자들의 발화 자료 중 장르를 한정하여 소규모로 만든 말뭉치 연구도 앞으로 꼭 필요하리라 본다. 일반 말뭉치 연구는 아니지만 학습자 말뭉치와 관련하여서는 초급과 고급 학습자의 부사 사용 양상을 비교하기 위하여 각각 26,523개와 91,142개의 문장으로 구성된 말뭉치를 구축하여 이를 연구에 활용한 유현경·서상규(2002), 그리고 초·중·고급 한국어 학습자의 글쓰기에서 64,980어절의 말뭉치를 구축해 명사+동사 꼴의 언어 관계를 살핀 이승연·최은지(2007) 등을 꼽을 수 있다.

이러한 논리를 좀 더 확장하면, 좀 더 조건을 세부화·특성화하는 경우에는 말뭉치의 규모 역시 그만큼 더 작아질 수 있다고 생각해 볼 수 있다. 가령, 텍스트의 장르를 제한하고, 동일한 주제로 글쓰기를 한 초급과 고급의 학습자를 비교할 때에는, 막연히 초급과 고급의 학습자의 작문 비교를 하는 경우보다 더 작은 규모의 말뭉치 구축도 가능하다. 앞 절에서도 언급한 바와 같이 TV 뉴스 텍스트에서 나타나는 일반적인 문법적 특징을 포착하기 위해서 해당 텍스트만 모아 소규모의 말뭉치를 만들고, 그에 필요한 듣기 교육 내용을 마련하는 것도 하나의 활용 예가 될 수 있다. 이를 위해서 항상 대규모화된 말뭉치를 이용해야 할 필요는 없을 것이다. 요컨대 말뭉치를 통한 연구 목적이 세밀할수록 말뭉치의 규모도 좀 더 작아질 수 있으므로, 말뭉치의 규모 자체가 절대적인 가치를 갖는다고 볼 수는 없다.

4.3.2. 말뭉치의 적정성에 대한 이해

말뭉치 규모와 더불어 고민해야 할 문제는 자신이 구축하려 하거나 혹은 이미 갖고 있는 말뭉치가 연구에 적절한지를 판단하는 일이다. 이때 고려해야 할

점이 ‘균형성’(balance)이다. 균형성은 말뭉치가 대표성(representativeness)을 가지기 위한 필수 조건으로서, 보통 텍스트의 범위(range)와 관련된다. 그러나 정작 말뭉치의 균형성을 정확히 판명해 줄 과학적인 근거는 없다는 문제가 있다(McEnery et al. 2006:16). 가장 분명한 기준이라고 하면, 텍스트의 종류 또는 장르에 따른 균형성의 확보를 언급할 수는 있으나, 텍스트를 어떤 장르로 구분할 것인가에 따라서 그 판단도 약간씩 달라질 수 있기 때문에 절대적 근거로 세우기는 어려운 면이 있다.

대표적인 균형 말뭉치(balanced corpus)로서 거론되는 세종 말뭉치는 대규모이기 때문에, 종종 균형성을 확보하기 위해서는 어느 정도 규모를 갖추어야 하는 것으로 인식될 수도 있다. 그러나 균형성을 판단하는 절대적 기준은 ‘연구 목적’과의 부합성이다.

여기서 다시 앞서 <그림 4>에서 들었던 작은 말뭉치 문제로 돌아가 보자. 여기서 예로 들은 말뭉치는 고작 1,602어절만으로 구성되어 있다. 이 정도로 작은 규모의 말뭉치로 어떤 연구 결론을 얻을 수 있는지를 말하기는 쉽지 않을 것이다. 만일 이를 기준으로 하여 ‘신문에 나타난 표현 항목’을 논한다면 이는 직관적으로도 합당하지 않을 것이다. 그러나 ‘최근 화제를 모았던 특정 사건과 관련된 신문 뉴스’로⁹⁾ 한정한다면 이야기가 달라진다. 이 작은 말뭉치에서 신문에서 나타나는 독특한 문법 내용을 찾을 수도 있고, 자주 사용된 명사나 동사 등을 조사해 볼 수도 있다. 마찬가지로의 관점에서 ‘신문 사회면에 나타난 표현 항목 발굴’을 목적으로 한다면, 최근 몇 달간의 사회면 뉴스만을 채집하여 말뭉치를 만들어도 충분히 활용할 가치가 있을 것이다. 그러나 이 말뭉치를 가지고 ‘신문에 나타난 표현 항목 연구’라고 일반화할 수는 없다.

이는 ‘연구 목적에 따른 적정 규모’를 판단하는 문제가 균형성과도 밀접한 관련이 있음을 보여주는 예이다. 지금까지 균형성과 관련하여서는 주로 구어와 문어의 비율을 많이 중시한 경향이 있었다. 그러나 균형성이란 말뭉치에 포함된 텍스트의 범위(range)의 문제이며, 단지 구어와 문어의 혼합비만을 일컫는 용어는 아니다. 대규모의 세종 말뭉치를 가리켜 균형 말뭉치(balanced

9) 이는 2009년 2월 2일부터 8일까지 보도됐던 용산 철거 사건과 관련된 다양한 인터넷 기사로만 구성된 것으로서, 가능한 한 다양한 신문사로부터 균등한 비율로 구축했다. 또한 문장 내 패턴의 특징을 살피기 위해 헤드라인과 기자 정보 및 출처 정보는 삭제한 것임을 밝혀 둔다.

corpus)라 하는 것은 ‘다양한 연구 목적을 만족시키기’ 위해 다양한 장르의 텍스트를 고려한 결과이지, 균형성을 확보하기 위해 규모를 키운 것은 아니다.¹⁰⁾ 요컨대 ‘균형성’이란 전적으로 연구 목적과 관계가 있는 개념이다(McEnery et al. 2006:13~6). 오히려 편중된 텍스트로만 규모화하게 되는 것이 더 큰 문제가 될 수 있다. 바로 전술한 것처럼 신문 사회면으로만 구성된 말뭉치를 갖고 신문 전체의 문제인 것처럼 살피면 안 되는 까닭은 바로 이 균형성 기준에 어긋나기 때문이다.

균형성과 관련하여 다른 문제로 들어가 보자. 한국어 구어 능력의 향상을 위해서 구어만으로 구성된 말뭉치를 사용하여 유의미한 교육 내용을 발굴하는 것은 어떨까? 가령, 구어에서 자주 사용되는 문법적 언어들을 발굴한다고 하면, 구어 말뭉치는 좋은 선택이 될 수 있다. ‘이유’의 연결 어미처럼 쓰이는 ‘-아/어 가지고/갖고’라는 표현 항목은 구어에서만 발견될 것으로 예상되기 때문이다. 따라서 한국어 모어 화자의 구어 말뭉치를 최대한 이용한다면 구어에서 표현 항목을 추출하고 그 빈도를 세어 연구 결과로 내 보일 수도 있을 것이다. 그러나 한국어 학습자가 가장 많이 직면하는 일상적인 의사소통 상황을 중시한다면, 구어 말뭉치라고 해도 준구어와 순구어를 구별해야 할 필요가 있을 수도 있다. 준구어 중에서도 방송에서의 내레이션, 일반 강연 등 화자의 발화가 두드러지는 경우는 균형성을 해칠 수 있기 때문이다. 이로 인해서 말뭉치의 규모가 작아지는 것은 문제가 될 수 없다. 규모는 크지만 균형성이 없는 말뭉치의 신뢰도에 문제가 생긴다는 것이 오히려 더 큰 문제이다.

또한 구어의 향상을 목표로 하는 한국어 교실을 지향한다고 하더라도, 문어 중심으로 구성된 말뭉치가 전혀 무용하다고 할 수도 없다. 구어가 갖는 일반적 특징, 즉 조사나 필수 성분의 빈번한 생략을 고려하면, 문법 연구의 측면에서 구어 말뭉치 연구는 부적절할 수 있다.

10) 세종 말뭉치는 가장 균형적으로 구축되었다고 알려진 1억 어절의 영국의 BNC(British National Corpus)를 참고하여 만들어졌다. 이 말뭉치는 미국, 폴란드, 러시아 등의 국영 말뭉치 구축 시 표본이 되었다(Sharoff 2003).

4.4. 직관의 적용

말뭉치 연구는 가장 객관적인 연구처럼 여겨질 수 있으나, 연구자의 직관(intuition) 역시 중요하게 적용되어야 한다. McEnery et al.(2006)에서는 말뭉치가 가져야 할 균형성을 판단하는 문제에 있어, 균형성을 담보할 만한 과학적 근거가 없기 때문에, 균형성을 확보하기 위해서는 ‘직관’에 의존할 수밖에 없음을 논하고 있다. 이에 대해 말뭉치의 균형성이란 ‘신뢰의 문제’이지 ‘사실의 문제’가 아니라고 단언하고 있다. 따라서 말뭉치 연구의 신뢰도를 평가할 때에는 독자들의 직관 역시 필요하다. 독자들은 말뭉치 조사 연구를 읽으면서 말뭉치의 규모를 판단하기에 앞서, 해당 말뭉치가 연구 목적에 비춰 볼 때 균형성을 갖추었는지를 판단할 수 있어야 한다는 것이다.

다른 한편으로 직관은 말뭉치 연구 결과에 대한 해석에서도 필요하다. 언어 자료에 대한 객관적인 수치를 보여주는 빈도는 거의 모든 말뭉치 연구에서 언급되고 있다. 탐구 대상이 무엇이나에 따라 다를 수 있으나, 결론적으로 말하자면 빈도 그 자체는 객관적 수치일 뿐, 그 자체가 어떤 특별한 가치를 갖지는 않는다. 남윤진(2002:362)에서처럼 빈도는 +/-의 이가적 특성을 가진 것이 아니라, 기본적으로 정도성의 특징을 띠고 있다는 점을 유념할 필요가 있다. 빈도는 그 자체로서 무엇을 판단하는 데 필요한 근거를 제공할 뿐, 그 자체가 결정의 척도로 사용되기는 어렵다는 것이다.

가령, 위에서 들었던 <그림 4>의 경우에 ‘-에 대하여’나 ‘-에 따라’와 같은 문법화된 표현 항목들의 빈도는 상대적으로 다른 동사들에 비해 높지만, 그 빈도는 각각 11개와 4개일 뿐임에 주목해 보자. ‘-에 따라’는 ‘(회견)-을 열고/열어’의 빈도 7보다 낮으며, ‘-□에도 불구하고’의 빈도는 겨우 3이다. ‘-을/를 통해/통한’의 빈도는 10으로서 비교적 다른 유형에 비해 높지만, ‘-을/를 비롯하여’는 2개의 빈도만을 갖고 있다. 만일, ‘-에 대하여’, ‘-에 비해’ 등과 같은 잘 알려진 표현 항목들이 주는 암시를 근거로 하여 ‘격조사+동사+어미’ 꼴을 갖는 교육 항목을 추출한다고 할 때에, 빈도라는 잣대만을 적용한다면 문제가 될 수 있을 것이다.

말뭉치 연구는 분명 객관적인 연구이기는 하나, 성공적인 말뭉치 조사 연구에서는 연구자의 직관(intuition)이 필수적이다. 이에 대해 Leech(1991:14)에서는 ‘직관을 거부했던 50년대 학자들도, 말뭉치를 거부했던 60년대 학자들도 말

말뭉치와 직관의 조화를 통한 성공적인 연구에는 다다를 수 없었다'고 말한 바 있다. 실제로 직관의 중요성은 과거 한국어교육계에서 있었던 기초 어휘 선정의 문제에서도 논의된 바 있다. 백봉자(2000)에서 활용 가치가 있는 어휘의 선정에서는 빈도수나 교사의 직관이 작용해야 한다고 언급한 것도 이와 동일한 맥락에서 이해될 수 있을 것이다.

요는 이러하다. 말뭉치를 활용한 연구는 컴퓨터를 통해 이뤄지는 연구이니 만큼 객관성이 매우 두드러진 연구라고 생각하기 쉽다. 그러나 말뭉치만큼 직관을 통해 보고 해석해야 하는 것도 없다. 말뭉치에서 유의미해 보이는 결과를 찾는 것은 컴퓨터가 아니라 사람이기 때문이다.

5. 결론

이상에서 본고는 지난 10년간 한국어교육에서 다양한 일반 말뭉치 연구가 진행되었지만, 그 활용의 폭이 넓지 않음을 지적하고, 일반 말뭉치의 활용 확대를 위한 논의를 진행하였다. 그간 일반 말뭉치 활용이 제한을 겪을 수밖에 없던 데에는 그간 말뭉치의 활용에 대한 방법론적인 논의가 부족했기 때문이기도 하다. 이에 본고는 보다 다양한 장르에서 교육 내용 발굴과 점검을 위해서 일반 말뭉치 활용이 좀 더 활발해질 필요가 있다는 측면에서, 크게 말뭉치 활용도에 대한 인식 확대가 필요함을 주장하였다. 또한 말뭉치 조사 연구가 컴퓨터의 활용이 필요한 만큼 주로 소수의 연구자들에게서 이뤄지고는 있으나, 그를 읽고 판단하는 다수의 독자들에게도 말뭉치 연구의 효용이 인식될 필요가 있으므로, 말뭉치 자체에 대한 이해가 필요하며, 더불어 말뭉치 연구에는 연구자와 독자의 직관이 중요함을 지적했다.

본론에서 언급됐지만, 말뭉치 연구는 매우 객관적이면서도, 연구 목적에 부합하는 적절한 말뭉치를 구축 또는 선정하기 위해서는 항상 고민이 뒤따를 수밖에 없다. 그러나 말뭉치가 아무리 크고 균형적이라 하더라도 '이미 발화된 자료 중 일부'라는 면에서 한계를 갖는다. 이런 점에서 '말뭉치를 활용한 연구의 결론을 내릴 때에는 자신이 이용한 말뭉치를 바탕으로 해야 한다'(McEnery et al. 2006:121)는 것이나, '특정 말뭉치를 이용하여 내린 결론들은 일반적 사실이라기보다 연역적 결과이다'(Hunston 2002:23)의 언급은 말뭉치 조사 연구의 활성화를 위해 모두가 알아두어야 할 명제가 되리라 본다.

〈참고문헌〉

- 강현화. 「어휘접근적 문법교수를 위한 표현문형의 화행기능 분석 : ‘-어야 하다, 되다’를 사례로」. 『한국어 의미학』 26 (한국어 의미학회, 2008) : 21~46.
- 강현화. 「중,고급 학습자를 위한 감정 기초형용사의 유의 관계 변별 기제 연구 : 기쁨, 슬픔을 나타내는 형용사의 통합관계를 중심으로」. 『한국어 의미학』 17 (한국어 의미학회, 2005a) : 43~64.
- 강현화. 「코퍼스를 이용한 부사의 어휘 교육 방안 연구」. 『이중언어학』 17 (이중언어학회, 2000) : 57~75.
- 강현화. 「한국어학습을 위한 프레지올러지 유형에 대한 연구(1) : 강조 표현을 중심으로」. 『이중언어학』 24 (이중언어학회, 2004) : 4966.
- 강현화. 「한국어학습을 위한 프레지올러지에 관한 연구(2) : 한중 대조 자료를 중심으로」. 『한국어 교육』 16-1 (국제한국어교육학회, 2005b) : 1~21.
- 고경태. 「명사화소 ‘-음’과 ‘-기’의 패턴 연구 : 조사 및 체언과 공기하는 패턴을 중심으로」. 『이중언어학』 38 (이중언어학회, 2007) : 1~25.
- 고경태. 「한국어 동사 교육 연구 : 동사의 통사 및 의미 교육을 위한 패턴 (pattern) 선정을 중심으로」. 고려대 박사학위 논문, 2008a.
- 고경태. 「한국어 교육을 위한 명사화소 ‘-음’, ‘-기’ 연구 : 용언과의 패턴을 중심으로」. 『문법교육』 8(한국문법교육학회, 2008b) : 1~29.
- 국립국어원. 『외국인을 위한 한국어 문법 2』. (커뮤니케이션북스, 2005).
- 김홍규 · 강범모. 『한국어 형태소 및 어휘 사용 빈도의 분석 1』. (고려대학교 민족문화연구원, 2000).
- 남윤진. 「국어 연구와 빈도 정보」. 『한국어와 정보화』(태학사, 2002).
- 백봉자. 「한국어 어휘 교육」. 제8회 언어정보개발연구원 월례학술발표회 발표 요지, 2000.
- 변우영. 「말뭉치의 교육적 활용에 대한 연구 : 화행 분석을 중심으로」. 『한국어 교육』 18-2 (국제한국어교육학회, 2007) : 135~154.
- 봉미경. 「시간부사의 어휘 변별 정보 연구 : 유의어 “방금”과 “금방”의 분석을

- 중심으로」. 『외국어로서의 한국어교육』 30 (연세대 한국어학당, 2005) : 113~139.
- 서상규 · 유현경 · 남윤진. 「한국어 학습자 말뭉치와 한국어 교육」. 『한국어 교육』 13-1 (국제한국어교육학회, 2002) : 127~156.
- 서세정. 「한국어 선어말 어미 ‘-겠-’의 영어 대응 양상 연구 : 영한 병렬 말뭉치를 이용하여」. 『언어정보와 사전편찬』 22 (연세대학교 언어정보연구원, 2008) : 193~215.
- 연규동 · 박진호 · 최운호. 『인문학을 위한 컴퓨터』 (태학사, 2008).
- 유석훈. 「외국어로서 한국어 학습자 말뭉치 구축의 필요성과 자료 분석」. 『한국어 교육』 12-1 (국제한국어교육학회, 2001) : 165~179.
- 유현경. 「사이, ‘동안’의 어휘 정보 연구」. 『한국어 교육』 18-1 (국제한국어교육학회, 2007) : 315~335.
- 유현경 · 서상규. 「한국어 학습자 말뭉치에 나타난 부사 사용에 대한 연구」. 『이중언어학』 20 (이중언어학회, 2002) : 213~234.
- 이승연 · 최은지. 「한국어 학습자의 어휘적 언어 사용 연구」. 『이중언어학』 34 (이중언어학회, 2007) : 299~321.
- 임유중. 「부정 표현과 호응하는 부사의 사용 양상과 한국어 교육」. 『한민족문화연구』 22 (한민족문화학회, 2007) : 187~214.
- 차재은. 「국어정보학의 응용 방안 : 한국어 교육과 국어 교육을 중심으로」. 『한국어 교육』 14-2 (국제한국어교육학회, 2003) : 423~443.
- 한송화 · 강현화. 「언어를 이용한 어휘 교육 방안 연구」. 『한국어 교육』 15-3 (국제한국어교육학회, 2004) : 295~318.
- 홍윤혜. 「보조사 ‘-(이)라도’의 의미와 분포적 특성 : ‘-(이)나’와 ‘-(이)나마’와의 비교를 중심으로」. 『한국어교육』 18-2 (국제한국어교육학회, 2007) : 487~506.
- Biber, D. 「Representativeness in corpus design」. *Literary and Linguistic Computing* 8/4, 2003 : 243~257.
- Hunston, S. *Corpora in Applied Linguistics*. Cambridge Univ. Press. 2002.
- Hunston, S., Francis, G. *Pattern Grammar*. John Benjamins. 2002.
- Leech, G. 「The state of art in corpus linguistics」. *English Corpus Linguistics*.

1991 : 1~18.

Lewis, M. Teaching Collocation : Further Developments in the Lexical Approach, Language Teaching Publications. 2001.

McEnery, T., Xiao, R., Tono, Y. Corpus-Based Language Studies : An advanced resource book, Routledge. 2006.

Santos, D. 「Tense and aspect in English and Portuguese : a contrastive semantical study」, Ph. D. thesis, Universidade Tecnica de Lisboa. 1996.

Sharoff, S. 「Methods and tools for development of the Russian Reference Corpus」, Corpus Linguistics Around the World, Amsterdam : Rodopi. 2003.

Sinclair, J. ed. How to Use Corpora in Language Teaching. John Benjamins. 2004.

Sinclair, J. Corpus Concordance Collocation, Oxford Univ. Press. 1991.

〈Abstract〉

A Suggestion for the enlargement of the general corpus-based researches in Korean language teaching

Koh, Kyoungtae

This study aims to emphasize the importance of the general corpus-based studies and to give suggestions for enlargement of them in Korean language teaching. By Korean language educators, the general corpora have been expected that they could give appropriate uses of words or morphemes and help to compose teaching contents by virtue of plentiful authentic data. However, there are few studies that take full advantage of general corpora, and this is partly because of the insufficient information or knowledge on the general corpus itself. According to this issue, I suggest that very specialized small sized corpora or D.I.Y. corpora, i.e., the corpus built with TV news or society section in newspapers etc., could be useful to compose the teaching contents, and emphasize that the appropriateness of corpus adopted in a research must be evaluated not by the size but by the balance of it. In addition, for a successful study, it would be needed not only the balanced corpus suitable for its intended use, but also the intuition of corpus-based researchers.

Keywords : corpus, Sejong corpus, general corpus, intuition, balance

이 논문은 2010년 7월 31일에 투고되었으며, 2010년 8월 12일에 심사완료되어 2010년 8월 16일에 게재가 확정되었음.