말뭉치와 언어학

최재웅*
고려대학교 언어학과 교수

1. 4차 산업 혁명

아주 일상적이면서도 이번 특집의 주제와 직결되는 소재부터 살펴보자. 2017년 초부터 국내에 '4차 산업 혁명'이란 말이 크게 회자되기 시작하면서 한국 사회를 흔들고 있는 것으로 보인다. 본래 정치계에서 대통령 선거 캠페인의 일환으로 주목받기 시작한 것으로 알고 있지만, 언론계와 산업계도 가세하면서 순식간에 대한민국을 휩쓸고 있다고 해도 과언이 아닌 듯하다. 학계도 예외는 아니다. 필자가 받은 지난 상반기의 학회 개최 안내 프로그램 대여섯 개를 보면 하나같이 학회 주제로 '4차 산업 혁명'이라는 어구가 중심을 차지하고 있다.[1] 이러한 추세가 전 세계적인 현상인가? 세상이 벌써 4차 산업 혁명 시대로 접어들고 있다는 말인가? 아니면 극히 일부에서만 논의되고 있는 주제인데 필자가 과잉 반응을 보이고 있는가? 이에

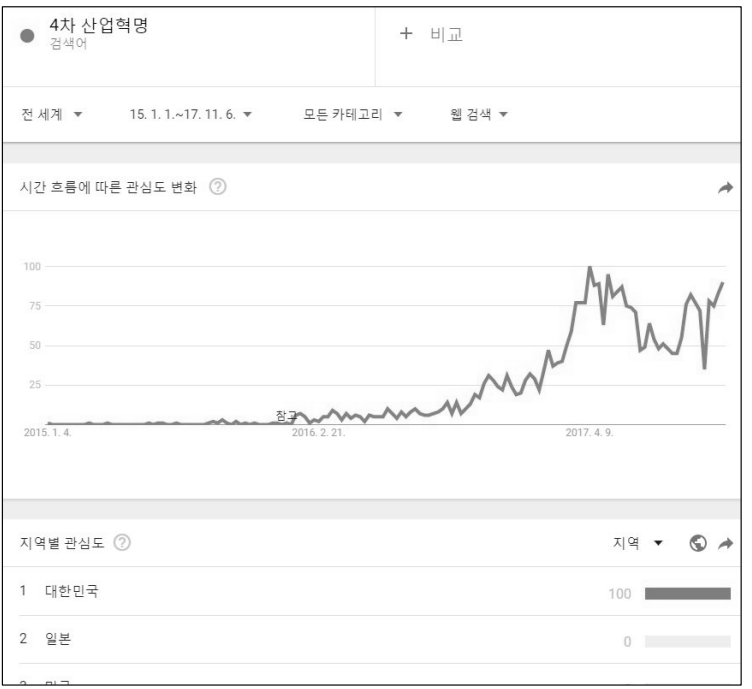
* 본 논문은 편집자의 요청에 따라 필자가 발표한 논문 '말뭉치와 언어연구[최재웅(2014), 한국어학 63, 71~102쪽]'를 상당 부분 발췌하여 재작성한 것이다. 일부(1장, 2장 전반부)는 새롭게 작성하였으며 논문 구조도 재구성하고 일부 수정하였다. 본문에 나오는 영어 약자로 된 말뭉치나 도구는 웹 검색으로 쉽게 찾을 수 있는 경우 웹 사이트 주소를 병기하지 않았다. 그 밖에 본문에 언급된 웹 사이트는 2017년 11월에 접속하였다.

1) 이러한 현상은 학술 논문에서도 뚜렷하다. 국내 대표적인 학술 논문 서비스 사이트 중 하나(http://dbpia.co.kr, 2017년 11월 10일 접속)에서는 '많이 이용된 논문' 5개 중 1위부터 3위까지의 주제가, 그리고 '인기 급상승 논문' 상위 5개 중 4개가 '4차 산업 혁명'을 주제로 한 것이다.

대한 답을 얻으려면 누구에게 물어보아야 하는 것인가?

이와 같은 의문은 필자만이 가진 의문일까? 웹이 보편화되면서 궁금한 내용을 검색 엔진에 확인해 보는 것이 일상화되었다. 그러한 검색어의 추이를 보여 주는 간편한 도구도 쉽게 접속이 가능하다.

그림 1 구글 트렌드 검색 결과: 4차 산업 혁명



'구글 트렌드(Google trends)'의 검색 결과인 <그림 1>[2]을 보면 2016년 하반기부터 '4차 산업 혁명'이라는 검색어 질의 횟수가 눈에 띄게 증가하다가 2017년 상반기에 관심이 크게 증대되었다는 점이 확연히 드러난다. 구글

2) https://www.google.com/trends/, 2017년 11월 6일 접속.

사용자들이 궁금해서 던져 보는 검색어 전체의 비중에서 '4차 산업 혁명'이 극적으로 증가하였다는 것이다. 즉, 이런저런 경로를 통해 필자가 어렴풋이 형성하게 된 시대의 흐름 중 하나를 <그림 1>은 아주 명확하게 드러내 주고 있다. 또한 위 그림의 하단에 나오는 '지역별 관심도'는 '4차 산업 혁명'에 대한 폭발적인 관심이 상대적으로 한국에서 특히 두드러진 현상이라는 점을 보여 준다.[3)]

'왜 이런 현상이 생겨났는가?', '이런 현상이 무엇을 의미하는가?', '4차 산업 혁명이란 무엇인가?' 등의 질문도 흥미롭겠지만 본 논문에서의 일차적인 관심사는 '구글 트렌드'의 검색 방식이다. 그 검색 방식을 어떻게 구축하였기에 그렇게 손쉽게 시대의 흐름을 확인해 볼 수 있는 것인가. 그것은 개념상으로는 매우 간단한 방식이다. 일반 사용자가 입력하는 검색어를 모두 모아 거대 규모의 '검색어 말뭉치'를 구축한 뒤에 그것으로부터 빈도 목록(검색어와 검색어별 빈도수)을 산출해서 그 목록을 기초로 <그림 1>에서처럼 특정 검색어의 상대적 분포를 화면에 보여 주는 것이다. 일별로 상대적 비중을 계산하여 해당 기간 중에 가장 비중이 컸던 날을 100으로 하고, 나머지는 그에 비례하여 그보다 작은 수치를 부여한 후 그것을 도표로 만들면 <그림 1>이 된다. 이는 개인의 직관이나 경험만으로는 도저히 만들어 낼 수 없는 구체적인 결과이다.

이와 같은 방식은 '검색어 말뭉치'로 한정되지 않는다. 예를 들어 언어를 대상으로 연구하는 학문인 언어학에서도 자료, 즉 텍스트를 대규모로 모아 말뭉치를 구축한 뒤에 언어와 관련한 궁금한 점을 검색해 보는 방법을 개발해 볼 수 있을 것이다. 실제로 그러한 방식은 이미 오래전부터 시도되고 활용되어 왔다. 이처럼 언어 연구를 위해 구축된 말뭉치 및 말뭉치 활용을

3) <그림 1>은 한국어로 검색한 결과만 보여 주기 때문에 해당 주제에 대한 다른 나라에서의 관심도는 공평하게 진단해 주지 못한다. 그러나 영어로 검색해 보아도 결과는 크게 달라지지 않는다. 검색어 '(the) 4th industrial revolution'로 검색한 결과에서의 '지역별 관심도'를 보더라도 한국에서의 관심도가 상대적으로 매우 크다는 점을 확인할 수 있다.

개괄해 보는 것이 본 논문의 목적이다. 논문의 구성은 다음과 같다. 2절에서는 말뭉치의 분류 기준 및 말뭉치 활용 도구에 대하여 논하면서 일부 주요 말뭉치를 소개한다. 3절에서는 언어학 및 언어 연구의 시각에서 말뭉치가 어떤 의미가 있으며 어떤 역할을 하는지 논한다. 4절은 결론이다.

2. 말뭉치(코퍼스)

말뭉치는 간단히 정의하자면 텍스트를 모아 놓은 것이다. 여기에서 텍스트란 폭넓게 해석되는 것으로 일상 대화를 전사해 둔 자료부터 신문 기사, 소설 등 문자로 작성된 모든 것을 포괄하는 개념이다. 또한 요즈음은 당연히 전자 파일 형태로 저장된 텍스트를 전제로 한다. 이러한 말뭉치는 어떤 종류가 있으며 어떤 기준으로 분류되는지, 그리고 말뭉치 활용을 위한 도구에는 무엇이 있는지 등을 이 절에서 살펴보기로 한다.

말뭉치를 만들 때 첫 번째로 고려해 보아야 할 사항은 무슨 목적에서 말뭉치를 만드는가이다. 예를 들어 사람들이 인터넷에서 어떤 내용을 궁금해하고 검색하는지를 알아보고자 한다면 앞에서 본 검색어 말뭉치를 구축해야 할 것이고, 한국어 소설에서 가장 흔히 쓰이는 형용사들을 찾아보고 싶다면 한국어 소설 말뭉치를 구축해야 할 것이다. 이처럼 말뭉치 구축에는 목적이 전제되고 있기 때문에 구축된 말뭉치를 분류하는 데도 해당 말뭉치가 어떤 목적에서 구축되었는지가 중요한 기준이 된다. 이러한 기준은 크게 말뭉치 내용면과 형식면으로 나누어 볼 수 있다.

그 둘을 차례로 살펴보기 전에 말뭉치 구축과 관련한 한 가지 추가 변수를 잠시 검토해 본다. 이른바 현실적인 제약이다. 이는 특히 말뭉치의 크기를 결정하는 가장 큰 변수로 작용한다. '이 세상의 모든 텍스트로 구성된 말뭉치'는 누구나 쉽게 생각해 볼 수 있는 목표이고 말은 간단하지만 실제로는

거의 불가능하다.[4] "현실적으로 구축 가능한가?"라는 질문에 답하기 위해서는 당연히도 현실적인 제약이 무엇인가를 생각해 보아야 한다. 그중 대표적인 제약은 주어진 여건을 고려해 볼 때 실질적으로 대상 텍스트를 구할 수 있는가이다. 구어의 경우엔 자료 모으는 데 따르는 제약이 문어에 비해 훨씬 더 커진다. 이와 같은 자료 수집상의 제약은 수집 가능한 자료의 규모와 직접적인 상관성이 있다. 또한 설령 텍스트를 구할 수 있다 하더라도 그 텍스트가 전자화되어 있지 않다면 전자화 과정에 많은 물적, 인적 자원이 소요될 수 있다는 점도 고려되어야 한다. 이에 더해서 점차 중요하게 인식되고 있는 요인은 윤리적, 법률적 제약이다. 텍스트를 구할 수 있고, 그것을 전자화하는 문제까지 해결되었다 하더라도 개별 텍스트별 저자로부터 해당 텍스트의 사용 허락을 받을 수 있는가라는 문제가 발생한다. 일반 텍스트의 경우 지식 재산권 보호가, 그리고 구어 자료의 경우에는 사생활 보호가 중요한 쟁점으로 부각된다.

2.1. 말뭉치 분류: 내용 기준

내용에 따른 말뭉치 분류에서 첫 번째 고려해 볼 요인은 말뭉치를 구성하는 텍스트의 언어이다. 세상에는 수많은 언어가 있고 각 언어별로 말뭉치가 구축될 수 있다. 물론 실질적으로는 웬만한 규모의 말뭉치는 극히 소수의 주요 언어로 한정되어 있는 것이 현실이다. 단일어로 된 말뭉치가 대부분이지만, 언어 간 비교를 위한 목적에서 구축된 말뭉치들도 있다. 두 언어 간 대조를 목적으로 한 병렬 말뭉치가 있고, 또 동일한 텍스트를 여러 언어로 옮기거나 번역한 텍스트를 함께 엮어 놓은 다국어 말뭉치가 있다.[5]

---

4) 구글 엔그램(https://books.google.com/ngrams)과 '코퍼스로서의 웹(Web as Corpus, Kilgarriff & Grefenstette, 2003)' 등이 현재 이러한 목적에 가장 부합하는 말뭉치이다.

5) 웹상에서 접속 가능한 대규모 공개 병렬/다국어 말뭉치로 OPUS(http://opus.nlpl.eu/, Tiedemann, 2012)를 들 수 있다. 병렬 말뭉치는 일반적으로 문장 수준에서 상호 대응 하도록 구성되어 있다.

말뭉치 구축에 가해지는 인적, 물적 제약을 고려해 볼 때, 많은 경우 특정 장르에 속하는 자료로 한정하여 말뭉치를 구성하는 방식이 통용된다. 자료 수집 과정이 비교적 용이하다고 볼 수 있는 신문 기사 말뭉치가 각 언어별로 상당수 크고 작은 규모로 구축이 되었고, 또 구축 기관의 입장에서 비교적 손쉽게 구할 수 있는 하위 장르별 대상 텍스트가 말뭉치 자료로 활용되었다. 이처럼 다양한 층위의 장르를 대상으로 한 제한된 목적의 특수 말뭉치가 많이 구축되어 있다.

말뭉치 구축을 특정 장르의 범위 내에서만 시도할 이유는 물론 없다. 개별 언어의 보편적 특성과 패턴을 연구하기 위해서는 개별 장르를 넘어서는 범언어적 차원의 말뭉치를 구축하는 것이 바람직하다. 이러한 말뭉치를 일반 말뭉치, 또는 참조 말뭉치라 칭한다. 일반 말뭉치 구축은 크게 두 가지 방식으로 나누어 볼 수 있다. 하나는 가능한 한 최대 분량의 텍스트를 모으는 방식으로 소위 '관찰 코퍼스(monitor corpus)'를 구축하여 수집 텍스트를 수시로 늘려 나가는 방식이다. 대표적으로 '영어 뱅크(Bank of English, Hunston, 2002)'가 있고, 장르가 신문과 잡지로 한정되어 있기는 하지만 NOW(https://corpus.byu.edu/now/)는 2010년 이후에 영어로 발간된 웹 자료가 매일 4~5백만 단어 분량 정도가 추가 업데이트되고 있으며 이를 즉각 활용할 수 있도록 되어 있다.

또 다른 일반 말뭉치 구축 방식은 수집 텍스트의 균형성과 대표성에 대한 틀을 미리 만들어 놓은 후 그 틀 내의 각 항목별로 텍스트를 수집하는 것이다. 언어 연구 차원에서 말뭉치 구축의 첫 번째 본격적인 시도라 할 수 있는 브라운 코퍼스(http://clu.uni.no/icame/manuals/)는 바로 두 번째 방식으로, 말뭉치 구성을 위한 표본 추출 틀(sampling frame)을 정한 후 그에 따라 텍스트를 일정 분량씩 수집하였다. 영국 영어 균형 말뭉치로는 BNC(http://www.natcorp.ox.ac.uk/)가, 미국 영어는 COCA(https://corpus.byu.edu/) 가 대표적이다. 특히 COCA를 비롯하여 마크 데이비스

(Mark Davies)가 구축한 다양한 말뭉치는 가히 영어 및 스페인어 말뭉치의 보고라 할 수 있다. 한국어 연구의 대표적 말뭉치인 세종 말뭉치 중 현대 문어 및 구어 말뭉치도 균형성과 대표성을 감안하여 구축되었다(https://ithub.korean.go.kr, 황용주·최정도, 2016).

언어 변화도 언어 연구의 주요 대상이고, 그러한 목적에서도 말뭉치가 구축된다. 지나간 특정 시점의 언어 사용 양상이나 시간의 흐름에 따른 언어 사용 양상의 변화를 알고 싶다면 역사 말뭉치, 또는 통시 말뭉치(diachronic corpus)를 만들어 연구한다. 영어의 경우 위에서 언급한 BYU 사이트의 COHA가, 한국어의 경우 세종 역사 말뭉치가 쉽게 이용 가능한 것들이다. 반대로 현재 사용 중인 언어의 일반적 특성을 알아보고 싶다면 최근 사용된 다양한 텍스트를 확보하여 공시 말뭉치(synchronic corpus)를 구축하여 활용한다. 앞에서 언급한 대부분의 말뭉치가 공시 말뭉치다.

그 밖에도 말뭉치 구축을 위한 다양한 목적을 생각해 볼 수 있을 것이고, 그런 목적에 부합되는 크고 작은 말뭉치가 필요할 것이다. 아래는 지금까지 기술한 분류 기준에 따라 말뭉치의 종류를 정리해 본 것이다.

(1) 구성 내용에 따른 말뭉치 분류 기준

가. 언어별: 영어, 중국어, 일본어, 한국어…

나. 언어 비교: 병렬, 다국어…

다. 장르별: 구어(대화, 강의, 설교…), 문어(신문 기사, 소설, 블로그, 편지…), 반구어(방송 뉴스 대본, 희곡…)…

라. 일반성: 일반/참조, 특수…

마. 균형성: 균형, 모니터…

바. 공시성/통시성: 공시, 역사/통시…

2.2. 말뭉치 분류: 형식 기준

이번에는 형식에 따른 말뭉치 분류에 대하여 살펴보기로 한다. 말뭉치는 구성 텍스트를 모은 뒤에 어느 정도의 정제 과정을 거친다. 정제 작업에서 고려해야 할 핵심적인 사항은 텍스트별 메타데이터 정보를 부여하는 일과 텍스트 주석이다. 메타데이터는 개별 텍스트별 출처와 저자, 제작 연도 등과 같은 문서의 이력에 관한 맥락적 정보를 가리킨다. 흔히는 국제적인 표준인 에스지엠엘(SGML) 또는 엑스엠엘(XML)식을 따른다. 이는 텍스트 주석에서도 마찬가지다.

구축 목적에 따라 다양한 종류의 텍스트 주석 체계를 채택하게 되고 이를 말뭉치에 적용한다. 초기에는 극히 기본적인 문서 정보나 문단 단위 구분 정도만 담은 원시 말뭉치로부터 시작하여 형태 분석 말뭉치까지 구축하는 것이 큰 흐름이었다. 거기에 더해 구문 주석을 더한 말뭉치도 어느 정도 규모로 구축되어 많이 활용되고 있다. 의미·화용적인 정보의 주석은 좀 더 어려운 단계로, 현재 다양한 방식으로 추진되고 있다. 정리하자면 대체로 아래와 같은 스펙트럼으로 주석 방식들을 구분해 볼 수 있다.

(2) 구성 형식에 따른 말뭉치 분류 기준

가. 원시(기본 정보 주석)

나. 형태 분석/중의 표시

다. 구문 분석(트리뱅크)

라. 논항 정보 주석

마. 시제-공간 정보 주석

바. 화용-담화-텍스트 정보 주석

사. 기타 특수 정보 주석: 습득/학습 관련, 언어 간 대응, 통시 정보 등

문장 구조를 논할 때 동사를 중심으로 한 논항 관계가 구조의 뼈대를 이루고 있다는 점은 거의 이견이 없다. 특히 문장 요소 간 의미적 관계를 체계적으로 파악하기 위한 방편으로 문장 내 술어와 그 논항 간의 관계를 명시적으로 표시해 보고자 하는 노력이 (2라) 논항 정보 주석의 핵심 내용이다. 대표적으로 프로프뱅크(PropBank)와 프레임넷(FrameNet)을 들 수 있다.[6] 논항 관계를 연구하는 언어학자들의 경우엔 대부분 예시적으로만 자료를 다루기 때문에 실제 대규모 자료에 그런 예시적인 방식이 확장 적용될 수 있을 것이라 추정만 할 수 있을 뿐이다. 반면 논항 정보 주석을 위해서는 실제 쓰인 언어 자료에서 발견되는 무수한 애매한 예들을 하나씩 모두 검토해야 하는 어려움이 있다.

(2마)에 언급된 시제 정보의 경우 국제적인 공조 연구(TimeML)를 예로 들 수 있다. (2바)와 관련한 주석에서는 예를 들어 구어 말뭉치에서 화자 및 구어적 특성을 어떻게 형식화하여 부착하느냐 하는 문제를 비롯하여, 지시 관계나 화행 등 전형적인 화행적 정보를 어떻게 표준화하여 주석을 붙일지 등의 문제가 있다. 또는 담화/텍스트 전개상에 드러나는 문장 간 의미 관계(예: 원인 관계)도 이미 말뭉치로 일부 구축되어 있다. (2사)의 예로는 언어 습득자나 학습자가 보이는 오류를 포함한 여러 특성들을 말뭉치에 표시하는 것이 있다.

말뭉치 주석과 관련하여 주목할 만한 큰 흐름으로 이종 주석 체계 간의 상호 운용성(interoperability)에 대한 필요성이 부각되고 있다. 같은 종류의 정보라 하더라도 개인별 주석 방식에서부터 그룹별 주석 방식, 국제 컨소시엄별 주석 방식 등 다양한 주석 방식이 혼재하는 상황에서 그러한 이종 체계의 주석 방식에 따라 구축된 말뭉치들을 좀 더 효과적으로 연계하여 다룰 수 있는 방편이 있다면 매우 바람직한 일이 될 것이다. 또는 A라는

6) 예를 들어 프로프뱅크(PropBank)의 경우 펜 트리뱅크(Penn Treebank)의 통사 구조 분석 자료에다가 논항별로 의미역을 추가하는 방식으로 구축되었다(Palmer, et al., 2005).

주석 방식으로 구축된 말뭉치를 B라는 주석 방식으로 일괄적으로 변환할 수 있는 도구가 개발된다면 필요에 따라 요긴하게 활용될 수 있을 것이다. 이러한 흐름은 그동안 다양하게 구축된 언어 자원을 연구자들이 모두 쉽게 접속하여 쓸 수 있게 사이버 자원 구조(cyberinfrastructure)로 만들어 가자는 원대한 비전하에 추진되는 작업으로 수렴된다고 볼 수 있다(Bender & Langendoen, 2009). 이와 관련하여 이미 오래전부터 언어 자원 주석 방식의 표준화 작업이 국제 표준 기구(ISO)의 주관하에 진행 중이다. 언어 자질 구조, 논항 정보, 시제 정보, 화행 정보 등 다양한 언어 정보 주석 체계에 대한 표준화 작업이 일부 이루어진 것도 있다(Kiyong Lee, 2014).[7]

2.3. 말뭉치 활용 도구

언어 연구의 관점에서 본다면 말뭉치 자체가 곧 연구를 위한 도구다. 그러나 말뭉치를 대상으로 놓고 본다면 그 대상을 이용하기 위한 방법이 필요하고, 그런 점에서 말뭉치 활용 도구가 필수적이다. 주요 언어를 중심으로 여러 언어 자원이 많이 구축되면서 당연히 따르는 문제는 그러한 자원을 활용할 도구로 무엇이 있는가이다. 그러한 활용 도구의 관점에서 가능한 방법들을 아래와 같이 구분해 볼 수 있다.

(3) 말뭉치 활용 도구 분류

가. 전용 도구

나. 범용 도구

다. 프로그래밍 언어: 파이선(Python), 펄(Perl), 루비(Ruby)…

라. 프로그래밍 패키지

마. 통계

7) 관련 웹 페이지 http://www.iso.org/tc 참고

바. 통계 도구: 아르(R), 에스피에스에스(SPSS)…

말뭉치 구축 초기에는 개별 말뭉치별로 그 말뭉치를 활용하기 위한 목적으로 개발된 전용 도구를 함께 배포하는 경향이 있었다. 물론 그러한 경우에도 그 도구가 다른 말뭉치에 활용될 가능성이 배제된 것은 아니었으나, 1차적인 도구 개발 목적은 해당 말뭉치에 특화된 활용 도구였다. 반면 애초부터 다양한 자원에 활용할 수 있도록 특정 말뭉치와 연계하지 않고 독자적으로 개발된 도구들이 사용자 수를 많이 늘려 가고 있다. 그러한 범용 도구로 앤트콘크(AntConc)와 워드스미스(WordSmith)를 들 수 있다. 매케너리와 하디(McEnery & Hardie, 2011)에서는 이러한 범용 프로그램을 4세대 도구라 부른다.

위에 언급한 두 가지 범용 도구는 초보-중급 수준의 사용자들이 웬만한 규모의 말뭉치 분석에 활용할 만한 최선의 도구라 판단된다. 말뭉치 분석에 필요한 기본 기능들은 현재 거의 표준화되어 있는 상태로 위의 두 범용 도구에는 그러한 기능이 포함되어 있다. KWIC 형태의 맥락 표시 검색어 목록(concordance), 어휘/어절 목록, 연어(collocation) 목록, 어휘 연쇄/엔그램(N-gram) 목록, 핵심어(keyword) 목록 등을 매우 손쉽게 추출해 주는 기능들이 이에 해당된다.

웹을 통한 말뭉치 활용도 많이 시도되는 방식이다. 매케너리와 하디(2011)에서 4세대 도구라 부르는 이 방식은 특히 말뭉치 배포 및 관리와 연결된 문제를 일부 해결해 주는 방식으로도 각광을 받고 있다. 이와 관련한 대표적인 사이트는 마크 데이비스가 구축해 놓은 비와이유(BYU) 말뭉치(https://corpus.byu.edu/)로 해당 포털을 통해 언어 연구자들은 여러 종류의 주요 말뭉치를 손쉽게 접근할 수 있다.[8] 위에 제시된 4세대 도구의 기능들

8) 비와이유(BYU) 사이트에는 말뭉치별로 다운로드를 받는 방식도 마련되어 있다.

을 비롯해서 더 다양한 도구가 웹상에 마련되어 있어서 해당 말뭉치에 대한 다각적인 연구를 할 수 있도록 가능성을 열어 준다.

어떤 도구도 그러하듯이 각 도구는 도구별로 개발자가 마련해 둔 기능만 사용할 수 있다는 제약이 있다. 그러한 한계에 구애받지 않고 마음껏 말뭉치를 자유자재로 다루고 싶다면 (3다)에서 언급된 프로그래밍 언어가 필요하다. 프로그래밍 언어 중에서는 대표적으로 파이선(Python), 펄(Perl), 루비(Ruby) 등 소위 스크립트형 언어가 언어 처리에는 매우 유용한 편이다.[9] 물론 언어학자에게 프로그래밍 언어는 다소 이질적인 대상이고, 그것을 배운다는 것이 가외의 부담이라는 점에서 선뜻 내키지는 않는 방식이지만, 프로그래밍 언어의 본질이 논리학이라는 점에서 언어학도들에게 아주 먼 대상은 아니라고 본다. 더군다나 처음 진입 장벽만 잘 넘어갈 경우 초급이나 중급 수준 정도의 프로그래밍 능력이면 KWIC 산출 등 어느 정도 수준의 작업은 어렵지 않게 해낼 수 있다. 언어학자를 위한 프로그래밍 입문서가 등장하여 이러한 필요성에 부응하고 있다[Hammond(2003), Bird, et al. (2009), Weisser(2009)]. 또한 프로그래밍 언어 관련 사용자 집단이 전 세계적으로 놀라울 정도로 활성화되어 있다는 점도 프로그래밍을 이용하는 사람들에게는 매우 고무적인 일이다. 입문자를 위한 수많은 튜토리얼(tutorial)부터 시작하여 어떤 종류의 궁금증에도 그에 대한 답이 이미 모두 인터넷상에 게시되어 있다고 해도 과언이 아니다.

프로그래밍 언어에는 때로 매우 활용도가 높은 모듈 및 패키지가 이미 개발된 경우가 있다. 이러한 패키지를 활용하는 것이 처음부터 독자적으로 개발하는 것에 비해 비교할 수 없을 정도로 효율적이다. 예를 들어 패키지로 개발된 NLTK는 파이선 사용자라면 활용해 볼 만한 매력적인 종합 언어

9) 기존의 도구들도 알고 보면 펄(Perl)이나 파이선(Python) 등 스크립트 언어로 개발된 것이다. 앤트콘크(AntConc)의 경우 현재 사용되는 버전 3.4까지는 펄(Perl)로 개발되었고, 버전 4.x부터는 대규모 메모리 처리 및 인터페이스 개발의 편의를 위해 파이선(Python)으로 바꾸어 개발될 예정이라 한다.

처리 도구이고, 펄의 경우에는 어휘망 워드넷(WordNet) 활용에 요긴한 일련의 도구들이 피더슨(Pedersen)에 의해 개발(http://www.d.umn.edu/~tpederse/)되었다.

말뭉치를 비롯한 언어 자원을 본격적으로 활용하는 데는 통계가 필수 불가결한 요소가 된다. 통계는 크게 기술 통계학과 추론/분석 통계학으로 나누어 볼 수 있다. 기술 통계학은 말 그대로 현상을 기술하는 수준의 통계로 주로 단순 빈도, 상대 빈도, 평균 등 중심 경향성을 드러내는 값, 그리고 그러한 것을 시각적으로 잘 드러내는 도표 등을 다룬다. 이와 달리 추론 통계학은 유의성(significance)을 논하게 된다. '빈도나 평균이 이러이러하다'를 보여 주는 것을 넘어서서 어떤 빈도 분포가 유의미한 분포라 할 수 있는지 이미 통계학에서 정립된 기준에 따라 유의미한 차이와 그렇지 못한 차이를 구분해 주는 척도를 활용한다. 이러한 통계법으로는 카이스퀘어 검정[chi-square(χ^2)], 티 검정(t-test), 로그 가능도 검정(log-likelihood) 등이 비교적 잘 알려진 것들이다. 국내 말뭉치 연구에서 최근 몇 년간 활용되고 있는 핵심어(keyword) 추출에도 카이스퀘어 검정이나 로그 가능도 검정 같은 척도가 활용된다. 그러나 점차 더 고난도의 통계 사용이 언어학 논문에서도 주류를 이루어 가고 있다. 주로 선형 분석이나 회귀 분석 같은 다요인 또는 다변량 통계 분석법이다. 비교적 잘 알려진 것들로 로지스틱 회귀 분석(logistic regression), 로그리니어 회귀 분석(loglinear regression), 군집 분석(cluster analysis)[10] 등이 있고, 최근 들어서는 '혼합 효과(mixed effects)' 통계법 사용이 확장되는 추세라고 한다(Gries, 2013).

통계 처리를 위해서는 적절한 통계 패키지 활용이 필수적이다. 기본적인 수준의 통계 처리는 마이크로소프트 엑셀(MS Excel)에서도 가능하나, 오랫동안 사회과학 쪽에서는 에스피에스에스(SPSS) 사용이 거의 절대적이었다.

10) 군집 분석은 엄격히 말해서 가설 검정 방식이 아니라 가설 생성 방식으로, 흔히 탐색적(exploratory) 방법으로 분류된다(Gries, 2013).

그러나 에스피에스에스의 경우 개인 연구자가 구입하기에는 부담스러운 높은 가격대로 인해 사용상의 제약이 많은 편이었다. 그러한 상황이 '아르(R)'라는 통계 패키지의 등장으로 상당히 바뀌었다. 무료로 배포되고 쉽게 설치가 가능하면서도 에스피에스에스와 대등한 수준의 통계 기능을 갖추고 있으며, 거기에 더해 뛰어난 그림 출력 기능 등을 갖추고 있다. 아르는 그동안 사용자 수가 폭발적으로 증가하여 현재는 연구자들에게 대표적인 통계 패키지로 간주되고 있다. 통계 언어학을 본격적으로 소개하는 주요 저서들에서도 아르를 이용하고 있다[Baayen(2008), Johnson(2008), Gries(2009), Gries(2013)].

특히 다른 통계 프로그램과 달리 아르는 말뭉치 처리, 분포 정보 추출, 통계 분석 등을 모두 하나의 환경에서 처리할 수 있다는 점이 큰 강점이다. 또 아르에는 프로그래밍 기능까지 포함되어 있어서 반복되는 일련의 작업을 프로그램화해 둘 경우 아무리 복잡한 절차라도 명령 하나로 쉽게 처리된다. 또한 전 세계적으로 아르 사용자 집단이 매우 활성화되어 있으며, 말뭉치 처리를 위해 특화된 기능을 담은 아르 패키지들도 인터넷에서 어렵지 않게 구해 쓸 수 있다.

3. 말뭉치 활용 언어 연구

일반적인 관점에서 본다면 말뭉치는 언어 이론을 위한 도구이다. 그러나 일부 학자들은 말뭉치와 언어 이론이 서로 분리될 수 없다는 관점을 취하기도 한다. 본 절에서는 말뭉치 구축 및 활용에 대한 몇 가지 관점을 제시해 본 후, 관점별로 어떤 관련 연구들이 있고 말뭉치 활용 연구의 동향은 어떠한지를 훑어보기로 한다.

3.1. 말뭉치와 이론 언어학

말뭉치에 대한 언어학계의 반응은 아직까지도 매우 다양하다고 본다. 대체로 다음과 같은 관점들로 정리해 볼 수 있다.

(4) 가. 말뭉치는 말뭉치일뿐 언어 연구와 무관
나. 이론 전개에 필요한 예시 자료 추출 자원
다. 언어적 일반화를 도출하거나 뒷받침하기 위한 자원
라. 말뭉치가 곧 언어 이론

(4가)와 같은 입장을 취하는 대표적인 학자로 촘스키(Chomsky)를 들 수 있다. “Corpus linguistics doesn’t mean anything.”[11]이라는 말이 그러한 입장을 명확하게 보여 주고 있고, 또는 말뭉치를 바탕으로 하는 확률적 접근에 대한 아래와 같은 촘스키의 언급도 많이 회자된다.

(5) “But it must be recognized that the notion of ‘probability of a sentence’ is an entirely useless one, under any known interpretation of this term.”(Chomsky, 1969)

이러한 극단적인 입장이 아마도 아직도 이론 언어학계의 주요 관점과 정서를 대변하고 있다고 해도 과언이 아닐 것이다. 이와 반대의 극단에는 (4라)와 같은 입장이 있다. 매케너리와 하디에서 언급된 신퍼스식(neo-Firthian) 입장이 그 예로, 그 관점에 따르면 설명적, 기술적 일반화의 유일한 자원은 말뭉치이다. (4가)의 입장이 화자 직관을 유일한 언어 자원의 원천으로 간주하는 것에 비해 정반대로 말뭉치가 그런 지위를 지니고 있다는

11) 인터뷰에서의 촘스키 언급(Andor, 2004: 97)

것이다.[12] 이러한 입장은 연구자의 주관을 철저히 배제하고 실제 자료에 근거해서만 문법을 구축하고자 했던 1950년대까지의 구조주의 연구 방법과 일맥상통하고 있다.

말뭉치 언어학의 연구는 위의 두 극단적인 입장보다는 (4나)나 (4다)와 같은 절충적인 방식으로 더 많이 진행되고 있다고 볼 수 있다. (4나)에서처럼 필요한 예시 자료 창고로서 말뭉치가 사용되는 예는 오래전부터 무수히 많은 편이다. 그러나 이러한 소극적인 태도에서 벗어나 좀 더 본격적으로 말뭉치에서 추출한 자료에 근거해서 이론을 도출하는, 그러면서도 연구자의 직관이나 기존의 이론적 틀과 연결시켜 보려는 연구들도 점차 꽤나 활발하게 이루어지고 있다. 또는 기존 이론에서 한 주제에 관해 단편적으로 논의되던 사항들을 엮어서 보다 종합적인 관점에서 말뭉치와 통계를 바탕으로 모형을 만들고 이론을 구축해 가는 방법들도 적극적으로 모색되고 있는 편이다.[13]

3.2. 말뭉치 활용 연구 분야

언어학에는 이미 확립된 여러 하위 분야들이 있다. 말뭉치 활용은 이러한 하위 분야의 구분에 구애받지 않고 모든 하위 분야에 걸쳐 연구에 활용되고 있다. 이를 가늠해 볼 수 있는 한 가지 방편으로 말뭉치 언어학 핸드북 2권(Lüdeling & Kytö, 2009)에 나오는 말뭉치 활용에 관한 개괄 논문들의 제목들을 살펴보기로 하자.[14]

---

12) 예를 들어 뢰메르(Römer, 2005: 7)는 다음과 같이 주장한다. “The investigation is highly committed to the data it starts from and [...] it tries to derive observational and theoretical findings from there, always trying not to lose contact with the corpora.”

13) ‘corpus-as-method’나 ‘corpus-based research’ 등으로 불리는 연구 방법이 대세라고 할 수 있다. 이런 경우엔 대개들 고급 통계를 활용한다[매케너리와 하디(McEnery & Hardie, 2011)].

14) 또 다른 말뭉치 언어학 핸드북인 오키프와 맥카시(O’Keeffe & McCarthy, 2012)나 바이버와 레펜(Biber & Reppen, 2015)도 좋은 참고가 된다.

표 1 Lüdeling & Kytö(2009)에 나오는 말뭉치 활용 개관 논문 목록

장	저자	주제
36	Marco Baroni/Stefan Evert	Statistical methods for corpus exploitation
37	Marco Baroni	Distributions in text
38	Douglas Biber	Multi-dimensional approaches
39	Antal van den Bosch	Machine learning
40	Hermann Moisl	Exploratory multivariate analysis
41	R Harald Baayen	Corpus linguistics in morphology: Morphological productivity
42	W Detmar Meurers/Stefan Müller	Corpora and syntax
43	Anatol Stefanowitsch/Stefan ThGries	Corpora and grammar
44	Sabine Schulte im Walde	The induction of verb frames and verb classes from corpora
45	Michael Hoey	Corpus linguistics and word meaning
46	Richard Xiao	Theory-driven corpus research: Using corpora to inform aspect theory
47	Michael McCarthy/Anne O'Keeffe	Corpora and spoken language
48	Anders Lindström/Robert Eklund	Cross-lingual influence: The integration of foreign items
49	Tuija Virtanen	Corpora and discourse analysis
50	Michael P Oakes	Corpus linguistics and stylometry
51	Anne Curzan	Historical corpus linguistics and evidence of language change
52	Christian Mair	Corpora and the study of recent change in language
53	Lieselotte Anderwald/Benedikt Szmrecsanyi	Corpus linguistics and dialectology
54	Josef Schmied	Contrastive corpus studies
55	Silvia Hansen-Schirra/Elke Teich	Corpora in human translation
56	Harold Somers	Corpora and machine translation

장	저자	주제
57	Holger Diessel	Corpus linguistics and first language acquisition
58	Stefan Evert	Corpora and collocations
59	Paul Clough/Rob Gaizauskas	Corpora and text re-use
60	Constantin Orasan/Laura Hasler/Ruslan Mitkov	Corpora for text summarisation
61	Douglas Biber/James K Jones	Quantitative methods in corpus linguistics

위 표의 가운데 열(column)에 등장하는 연구자 명단에는 공히 말뭉치 언어학을 선도하는 학자들 중 상당수가 망라되어 있다. 개별 논문의 제목인 오른쪽의 세 번째 열에 담긴 내용들을 일별해 보면 우선 통계나 언어 공학과 관련된 주제들이 많이 눈에 띈다(36~40, 58, 60~61). 그만큼 통계가 중요한 비중을 차지하고 있다는 말이 될 것이다. 나머지는 주로 기존 언어학적 주제와 관련된 연구로 형태론(41), 통사론(42, 43), 의미론(44~46), 담화 분석(49, 50), 역사 언어학(51, 52), 방언학(53), 번역학(55, 56), 언어 습득(57), 대조 연구(54) 등이 제시되어 있다. 컴퓨터가 대부분의 모든 학문 분야에 주요한 연구 도구나 방법론으로 자리 잡아 가는 것과 마찬가지로 말뭉치를 활용하는 연구가 언어학의 모든 하위 분야에 걸쳐 활용되고 있다고 볼 수 있다.

위에서는 부각되고 있지는 않지만 사회 언어학적 연구에서 말뭉치가 매우 중요한 자원으로 활용될 수 있고 그러한 연구가 많이 이루어진 바 있다. 특히 맥락적 변인이 잘 마련된 말뭉치의 경우 말뭉치를 구성하는 텍스트 내 어휘나 문법적 특징과 성별 등 사회적 변인 사이의 상관관계를 대규모로 연구하는 데 큰 기여를 할 수 있다. 사회 언어학적 차원의 말뭉치 활용에 대한 교과서적인 베이커(Baker, 2010)에도 나올 정도로 이 분야에 대한 연구도 활성화되어 있다.

또한 위 목록에 화용론과 관련한 연구가 명시적으로 제시되어 있지 않으나 특히 구어 말뭉치를 활용한 화용적 측면의 연구가 이루어지고 있다. 예를 들어 대화 시 말 차례 유지와 관련한 언어적 특징 또는 화행적 특징 등을 연구하는 데 구어 말뭉치는 훌륭한 자원이 된다[Adolphs(2008), Rühlemann(2012)]. 또는 화자의 태도를 드러내는 '의미 운률(semantic prosody)'에 대한 연구도 흥미로운 주제로, 이는 특히 댓글 등의 논조를 효율적으로 파악하려는 산업계 쪽에서의 '논조 분석(sentiment analysis)'과도 일맥상통한다.

사전학이나 사전 편찬이란 주제가 언어 이론적 논의에서는 좀 벗어나 있는 편이지만, 말뭉치가 사전 편찬에 결정적인 기여를 하고 있다는 점은 결코 과소평가될 수 없다고 본다. 이는 이론적 관점에서 어휘 의미를 체계화하려는 노력과도 통하는 것으로 그동안 말뭉치가 가장 괄목할 만한 기여를 한 분야로는 아무래도 어휘 의미론 및 사전 편찬이라고 해도 과언이 아닐 것이다.

3.3. 협의의 언어 연구, 광의의 언어 연구

언어에 대한 연구라고 하면 언어학자들은 우선 언어학자들이 제시한 이론부터 생각하게 될 것이다. 그리고 그러한 이론의 연장선상에서 언어 현상을 바라보는 경향이 있다. 그러나 언어에 대한 관심과 탐구는 언어학자들만의 전유물이라고 할 수 없다. 구글을 비롯한 거대 규모의 회사들이 사활을 걸고 씨름하는 문제도 크게는 웹 페이지에 담긴 언어를 이해하여 검색 엔진의 효율성을 높이자는 데 있다. 웹의 창시자라 할 팀 버너스 리(Tim Berners-Lee)가 차세대 웹을 '시맨틱 웹(Semantic web)'이라 명명한 것이 우연이 아니다(Berners-Lee, et al., 2001). 그 말에 담긴 '시맨틱(Semantic)'이 언어학 하위 분야 '의미론(Semantics)'과 많이 다른 것처럼

생각될는지 몰라도 근본적으로는 언어의 의미 문제를 가지고 고민하는 것이라는 점에서 다르지 않다. 웹 자원에 의미적 정보를 추가하고 보강하여 보다 정확하고 효과적인 검색이 가능하도록 하자는 것이 '시맨틱 웹(Semantic web)'의 주요 목표로 되어 있다.

광의의 언어 연구로 구글 번역도 들 수 있다. 자동 번역기를 만드는 것은 오랫동안 언어학자 및 언어 공학자들의 목표였으나 제대로 이루지 못한 꿈이었다. 언어학자들은 당연히도 언어에 내재하는 규칙을 찾아 그것을 번역 시스템에 구현하는 것에 관심을 쏟았으나 이러한 방식은 극히 제한적으로만 목표를 이룰 수 있었을 뿐 대규모의 실제 자원에 적용하는 데는 성공하지 못했다. 반면 같은 과제에 대하여 구글에서는 언어의 구조나 의미에 의존하지 않고 거의 전적으로 통계적인 방식으로만 시스템을 만들어 개방하였고, 이는 여러 실제 현장에서 사용할 만한 수준이라고 평가되고 있다. 이러한 통계적인 방식은 두말할 나위 없이 말뭉치를 전제로 한다. 언어학자로서 관심을 가져 볼 만한 부분은, 과연 통계에 무슨 비결이 있기에 전통적인 언어학적 접근법으로는 풀지 못하는 문제가 실용 가능한 수준으로까지 풀릴 수 있게 된 것일까라는 점이다. 언어의 비밀을 파헤치고 이해하는 방식이 전통적인 이론만이 전부가 아닐 수도 있다는 열린 사고가 필요하다고 본다. 따라서 전통적인 의미의 언어학적 연구를 '협의의 언어 연구'라고 본다면 그런 테두리 밖에서 매우 활발하게 전개되고 있는 통계적 접근 방식의 언어 연구도 포함된 '광의의 언어 연구'가 있다. 그런 데까지도 언어학적 관심의 지평을 넓히고자 하는 것이 말뭉치 언어학과 관련된 한 가지 흐름이라 판단된다.

4. 결론

본 논문에서는 거시적인 관점에서 말뭉치 언어학과 관련된 중요 개념, 도구와 쟁점 및 관련 동향을 살펴보았다. 말뭉치에 전혀 신경을 쓰지 않고도 할 수 있는 언어학 연구 주제도 수없이 많고, 경우에 따라서는 말뭉치가 거의 도움이 되지 않는 주제들도 물론 있다. 그러나 대부분의 언어 연구에 말뭉치가 어떤 형태로든 기여할 측면이 있을 것이라는 점은 점차 분명해지고 있다. 많은 경우 연구의 효율성과 객관성의 향상에 결정적인 기여를 한다. 특히 언어 현상의 법칙성 추출이나 자료 정제 등에 탁월하다. 물론 언어학적 논의의 중요한 근거를 찾아내는 데도 많이 활용되고 있다. 직관에 의존할 때와 비교할 수 없을 정도로 큰 규모의 언어 자원 활용이 가능해진다.

또한 만일 언어학적 연구가 실험실 내의 소규모 실험 형태로만 머무르지 않고 그것을 실제 현장에 활용하는 데까지 감안해야 한다면 말뭉치 활용이 더욱 중요해진다. 언어 정보에 대한 관심과 연구가 더 이상 언어학자들만의 전유물이 아닌 시대에 들어선 지 이미 한참을 지났다. 언어 연구의 전문가 집단에서도 그러한 '외부'의 움직임에 관심을 가지는 사람들이 지속적으로 배출될 필요가 있다고 본다. 그러한 방향으로 관심의 폭을 넓힐 수 있는 중요한 통로 중 하나로 말뭉치와 통계를 들 수 있다. 언어 공학이나 산업계에서 이루어지는 언어 처리가 기본적으로 거대 말뭉치인 웹 자원과 그것을 다룰 수 있게 해 주는 통계이기 때문이다. 그리고 그 연장선상에 4차 산업 혁명과 언어학 사이의 접점이 맺어질 수 있을 것이다.

참고 문헌

황용주·최정도(2016), '21세기 세종 말뭉치 제대로 살펴보기 - 언어정보나눔터 활용하기', ≪새국어생활≫ 26-2, 국립국어원, 73~86쪽.

Andor, József(2004), "The master and his performance: An interview with Noam Chomsky", Intercultural Pragmatics 1-1: 93-111.

Adolphs, Svenja(2008), Corpus and Context: Investigating Pragmatic Functions in Spoken Discourse, John Benjamins Publishing Company.

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

Baker, Paul(2010), Sociolinguistics and Corpus Linguistics. Edinburgh University Press.

Bender, Emily M. and D. Terence Langendoen(2009), "Computational Linguistics in support of linguistic theory", Linguistic Issues in Language Technology - LiLT.

Berners-Lee, Tim, James Hendler & Ora Lassila(2001), "The Semantic Web", Scientific American, May 2001. 29-37.

Biber, Douglas and Randi Reppen(2015), The Cambridge Handbook of English Corpus Linguistics, Cambridge University Press.

Bird, Steven, Ewan Klein, & Edward Loper(2009), Natural Language Processing with Python. O'Reilly Media.

Chomsky, Noam(1969), "Quine's empirical assumptions", In Davidson and Hintikka (eds.), Words and Objections, Humanities Press.

Gries, Stefan Th(2009), Quantitative Corpus Linguistics with R: A Practical Introduction, Routledge.

Gries, Stefan Th(2013), Statistics for Linguistics with R: A Practical Introduction, de Gruyter Mouton; 2nd revised edition.[최재웅·홍정하 역(2013), 언어학자를 위한 통계학: R 활용, 고려대학교 출판부.]

Hammond, Michael(2003), Programming for Linguists: Perl for Language Researchers, Blackwell.

Hunston, Susan(2002), Corpora in Applied Linguistics, Cambridge University Press.

Johnson, Keith(2008), Quantitative Methods in Linguistics, Wiley-Blackwell.

Kilgarriff, Adam. and Gregory Grefenstette(2003), “Introduction to the special issue on the Web as Corpus”, Computational Linguistics 29 (3): 333 - 47.

Kučera, Henry and Nelson Francis(1967), Computational Analysis of Present Day American English, Providence, RI: Brown University Press.

Lee, Kiyong(2014), “ISO Standards on language resources”, Manuscripts.

Lüdeling, Anke and Merja Kytö(2009), Corpus Linguistics HSK 29.2 (Handbooks of Linguistics and Communication Science) Volume II, Mouton de Gruyter.

McEnery, Tony and Andrew Hardie(2011), Corpus Linguistics: Method, Theory and Practice, Cambridge University Press. [최재웅 역, 코퍼스 언어학: 방법, 이론, 실제, 고려대학교 출판부, 2018년 1월 발간 예정.]

O’Keeffe, Anne and Michael McCarthy(2012), The Routledge Handbook of Corpus Linguistics, Routledge.

Palmer, Martha, Daniel Gildea & Paul Kingsbury(2005), “The Proposition Bank: An annotated corpus of semantic roles”, Computational Linguistics 31.1: 71-106.

Römer, Ute(2005), Progressives, Patterns, Pedagogy: A Corpus-driven Approach to English Progressive Forms, Functions, Fontexts and Didactics(Studies in Corpus Linguistics), John Benjamins Publishing Company.

Rühlemann, Christoph(2012), “What can a corpus tell us about pragmatics?”, In O’Keeffe & McCarthy(2012), 288-301.

Tiedemann, Jörg(2012), “Parallel data, tools and interfaces in OPUS”, In Proceedings of the 8th International Conference on Language Resources and Evaluation(LREC 2012).

Weisser, Martin(2009), Essential Programming for Linguistics. Edinburgh University Press.