한국어 의미학 43, 107-132 (2014. 3. 30.)

구어 어휘의 의미 연구 방법론*,**

-말뭉치 기반 사용역 변이 연구를 중심으로-

안 의 정
(연세대학교)

<Abstract>

Ahn, Euijeong. 2014. A Study of Methodology for the Meaning of Colloquial Vocabulary-Focused on the study of corpus-based register variation. *Korean Semantics, 43.* This study aims at finding the methodology for the meaning of colloquial vocabulary by arranging the history of research of corpus-based register variation. Many studies using corpora have focused on variation by comparing two or more dialects, channels, genres, sublanguages or varieties of language in order to uncover the main differences and similarities between them. In one language, there are many types of sublanguage existed and this sublanguage can be divided and understood in the aspects of genre, language register, and style. In this study, I'd like to determine how to suggest the methodology for the study of meaning by studying the variation of corpus linguistics. To understand the language register is to understand various usages of language forms in the discourse level out of sentence level. Therefore, we can understand the meaning of spoken language vocabulary in the discourse level by focusing on the study of language register variation including spoken language. In Chapter 2, I

* 이 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2009-361-A00027).

** 이 논문은 2014년 2월 13일 중앙대학교에서 열린 제34차 한국어의미학회 전국학술대회에서 주제 발표 논문(주제: 구어의 의미론)으로 발표된 초고를 수정한 것이다.

organized the relationship between corpus linguistics and semantic study as a basic theoretical discussion. In Chapter 3, thereafter, I organized what type of variation study was processed following the corpus annotation. I organized the study of various registers by dividing into raw corpus, POS tagged corpus, and speech corpora. Finally, in the Chapter 4, I introduced a methodology for the study of meaning targeting sense tagged corpus.

핵심어: 말뭉치언어학(corpus linguistics), 구어 말뭉치(spoken corpus), 언어 변이(language variation), 언어 사용역(language register), 텍스트 장르(text genre), 텍스트 유형(text type), 형태주석 말뭉치(POS tagged corpus), 음성 말뭉치(speech corpora), 의미주석 말뭉치(sense tagged corpus), 담화표지(discourse marker)

1. 서론

이 글의 목적은 말뭉치 기반 사용역 변이 연구사를 정리하여 구어 어휘의 의미 연구 방법을 모색하는 것이다. 1960년대 구미에서 시작된 말뭉치언어학은 이미 중요한 언어 연구 방법론으로 자리잡아 가고 있다. 국내에도 1억 어절 이상의 대량의 말뭉치가 다양한 유형으로 구축되었으며, 분석 도구 및 방법론에 대한 연구도 활발히 진행되고 있다.

말뭉치언어학의 여러 방법론 중에서 변이(variation)에 대한 연구는 방언/채널/장르/하위언어의 영역에서, 또는 언어 변천의 측면에서, 언어의 다양성을 발견하고 이들 간의 차이점과 유사점을 발견하는 것이다.[1] 하나의 언어 속에는 다양한 유형의 하위 언어가 존재하고 이 하위 언어는 장르, 언어 사용역, 문체(스타일)의 측면에서 분화되고 파악될 수 있다.[2]

그렇다면 말뭉치언어학의 변이 연구를 통해 어떻게 의미 연구 방법론을 제시할 수 있는가? 말뭉치언어학적 의미 연구는 김진해(2006:77)에서 논의한

1) Baker 외(2006:166)의 용어 정리를 참고하였다.

2) 언어학에서 변이에 대한 연구는 사회언어학을 중심으로 시작이 되었고, 다양한 말뭉치가 수십 년간 축적이 되면서 말뭉치언어학적 변이 연구가 활발히 진행되고 있다.

바와 같이 새로운 연구 대상을 발견하게 되었다는 범위 확장의 의의를 가지고 있으며, 의미의 생성과 존재 방식에 대한 근원적 질문을 찾아낼 수 있게 한다. 범위 확장이란 기존의 직관을 중심으로 한 연구에서 객관적인 자료로 관찰의 대상이 전환된 것을 말한다. 그리고 김유정(2011:101~107)에서는 언어 사용역을 활용한 의미 분석 기준을 제시[3], 언어 사용역을 파악한다는 것은 문장 차원을 벗어나 담화 차원에서 언어 이형의 다양한 용법을 파악한다는 것을 의미한다고 하였다. 따라서 구어를 포함한 언어 사용역의 변이 연구를 중심으로 구어 어휘가 담화상에서 가지는 의미를 파악할 수 있다.

아울러 말뭉치언어학에서는 본질적으로 언어가 심리적 현상이 아닌 사회적 현상이며, 의미도 사회적 현상으로 파악하고 있다(김진해, 2006:83). 여기서 사회적 현상이란 구체적으로 무엇을 말하는가? 이는 의미의 이해가 고정된 것이 아니라 사회 구조에 따라 변화를 보이며 맥락 의존적이라는 의미로 설명될 수 있다. 따라서 다양한 장르와 사용역의 변이에 따른 말뭉치 연구는 구어 어휘가 특정 상황에서 어떻게 쓰이고 어떻게 이해되는가를 객관적으로 보여줄 수 있는 방법이 될 것이다.

이 글의 구성은 다음과 같다. 먼저 2장에서는 기초적인 이론적 논의로 말뭉치언어학과 의미 연구에 대해 정리해 보고자 한다. 이어 3장과 4장에서는 말뭉치의 유형(원시/주석)과 주석의 유형(형태분석/의미주석)에 따라 어떠한 변이 연구가 진행되었는지를 정리하고자 한다. 4장에서는 주석 말뭉치 중에서도 새로운 의미주석 말뭉치를 토대로 한 의미 연구 방법을 소개하고자 한다. 새로운 의미주석 말뭉치란 다의어 차원까지 의미주석된 말뭉치로 3장에서 다룬 자료와 마찬가지로 구어, 문어로 구성되어 사용역 변이 연구를 할 수 있는 자료이다.[4]

---

3) 여기서 언어 사용역을 활용한 의미 분석 기준은 아래와 같이 담화의 장과 담화의 매체, 담화의 형식으로 나누어 제시되었다.
-담화의 장 : 담화 주제(무엇을 이야기하는가?), 담화 목적(제보/호소/책무/접촉/선언 기능)
-담화의 매체 : **담화 수단(구어/문어), 담화 유형(장르)**
-담화의 형식 : 청화자 성별, 청화자 관계1(지위), 청화자 관계2(친밀도), 청화자 태도, 발화 장면, 담화 문체 등.

4) 의미 주석 말뭉치를 4장에 분리한 이유는 본고의 연구 목적이 의미 연구 방법론의 소개이며, 의미 주석을 활용한 구어 어휘 연구가 가장 중요하다는 판단에서 그 개념과 구축 방법, 활용

2. 이론적 논의

2.1. 말뭉치언어학과 의미 연구

말뭉치언어학적 의미 연구의 특징과 내용에 대한 논의는 Sinclair(2007), Hoey(2009), 김진해(2006) 등에서 이루어졌는데, McEnery 외(2006:103~104)에서는 말뭉치가 의미 할당에 객관적인 기준을 제공하며, 퍼지 범주와 변화도(gradience)의 증거를 제공할 수 있다고 하였다. 또, 어휘 차원에서의 의미 연구는 의미 운율(semantic prosody)[5]과 의미 선호(semantic preference)에 대한 연구가 진행되고 있다고 하였다.

의미 운율이란 "어떤 환경에서 암시되는, 문자적 의미에 얹혀서 실현되는 의미이며, 통사적, 화용적 측면을 반영한 관습적인 태도나 견해(Louw(1993), 김혜영, 강범모(2010:105)에서 재인용)"를 말한다. 그리고 의미 선호는 의미 패턴(semantic pattern)이라고도 하며, 어떤 단어형과 관계가 있는 의미 부류를 말한다. 예를 들어 'rising'은 'work/money'의 의미 부류와 함께 쓰이는 경향이 있어 "incomes, prices, wages, earnings, unemployment"와 같은 단어와 잘 결합한다.[6]

말뭉치언어학적 의미 연구를 살펴보기에 앞서 먼저 말뭉치언어학의 기본 성격을 정리해 보자. 김진해(2006:78)에서는 말뭉치언어학의 기본 성격을 다음과 같이 제시하고 있다.

(1) 말뭉치언어학의 기본 성격
- 빈도(frequency)를 통한 양적 연구
- 언어사실주의
- 하위언어의 존재 각인
- 언어의 가부가 아닌 통계 또는 개연성(확률)의 문제
- 문맥주의 = 전체론적, 비분석적 입장

방법을 강조하기 위함이다.

5) 얹힘의미소로 번역하기도 한다(한영균, 고은아, 2011:386).

6) Baker 외(2006:144)의 용어 정리를 참고하였다.

빈도는 말뭉치언어학에서 자주 쓰이는 용어이며 말뭉치의 통계적 분석에서 가장 중요한 요소이다. 실제로 국내의 많은 말뭉치 기반 연구에서 이 수치만으로 모든 현상을 해석하고 있을 정도이다. 말뭉치언어학이 방법론에 있어서 언어의 가부가 아닌 빈도 분석을 통한 통계 또는 확률로 표현된다는 것은 의미의 구분이 절대적이 아니라 변화의 양상에 있다는 것을 보여주는 것이며, 이는 말뭉치언어학적 의미 연구의 중요한 경향이라고 할 수 있다. 예를 들어 Altenberg(1994)에서는 고빈도 기능어 'such'의 사용 분포를 말뭉치 분석을 통해 파악함에 있어 강조의 'such(intensifier)'[7]가 LLC[8]에서는 사적 대화 및 공적 담화에서 비슷하게 나타날 가능성이 있는 반면, 확인의 'such (identifier)'[9]는 사적 대화에서는 사용 빈도가 현저히 낮아짐을 시기적으로 다른 말뭉치를 통해서 발견하였다. 이 부분에서 말뭉치 기반 분포 분석의 가치가 드러난다. 즉, 말뭉치 분석은 이렇게 기능의 전이를 보이는 연구 대상을 분석함에 있어, 절대적인 구분에 의한 것이 아니라 기능적 변화 경향을 파악할 수 있게 한다[10]. 이러한 연구는 최근의 구어성에 대한 국내 연구인 서상규(2013ㄱ)에서도 잘 드러나는데, 여기서는 어휘의 구어성을 말뭉치언어학의 정도성으로 해석하여 빈도 분석을 기반으로 한 스펙트럼을 제시하고 있다.

2.2. 사용역의 개념과 연구 방법

여기서는 3장의 연구사 정리의 범위 설정을 위해 사용역의 개념에 대해 정리해 보고자 한다. Biber & Conrad(2009:16)에서는 장르와 사용역, 문체의 연구 경향 및 관점을 다음과 같이 비교하고 있다.

---

7) 다음이 강조의 예이다. 예) **such** a beautiful daughter(Altenberg, 1994:225)

8) The London Lund Corpus을 줄인 것이다. 50만 어절로 구성되어 있으며 기본적인 철자법 전사와 함께 상당히 자세한 운율 전사가 되어 있다. 따라서 이 말뭉치는 운율 연구뿐 아니라, 어휘, 문법, 대화 분석 등에 이용되고 있고, 문어 말뭉치인 The Lancaster-Oslo/Bergen corpus와 비교하는 연구에서 광범위하게 이용되었다(안의정, 1998:9).

9) 다음이 확인의 예이다. 예) he struck the plaintiff's cycle in **such** a way as to break the plaintiff's right leg(Altenberg, 1994:227)

10) 안동환 역(2010:230) 참고.

<표 1> Biber&Conrad(2009)의 장르/사용역/문체의 연구 경향 비교

	장르	사용역	문체
대상 텍스트의 범위	전체	텍스트에서 발췌한 샘플	텍스트에서 발췌한 샘플
언어학적 특성	특정 표현, 수사적 구조화, 포맷	모든 어휘-문법적 자질	모든 어휘-문법적 자질
언어학적 특성의 분포	텍스트에서, 또는 텍스트의 특정 위치에서 보통 1회 출현	변이에 따라 텍스트 전체에 자주, 전반적으로 나타남	변이에 따라 텍스트 전체에 자주, 전반적으로 나타남
해석	자질은 관습적으로 장르와 연관됨 : 예상되는 포맷이지만 종종 기능적이지 않음	자질은 사용역에서 중요한 의사소통적 기능을 제공	자질은 직접적으로 기능적이지 않음 : 이들은 미학적인 가치가 있어 선호된다.

예를 들어 상업적 편지의 장르 연구라고 한다면 전체 텍스트에서 예상되는 텍스트적 관습, 즉 편지의 시작 형식부터 끝맺음 방식까지 모두 분석하는 연구가 된다(Biber & Conrad, 2009:17). 이렇듯 장르와 사용역은 파악하고자 하는 언어학적 특성과 그 분포가 다름을 알 수 있다.

대부분의 국내 연구에서 사용역은 장르의 하위 개념으로, 언어 사용 상황에 따라 달리 사용되는 언어의 다양한 양상을 포괄하는 용어로 쓰이고 있다(남길임, 차지현, 2010:92). 여기서 말하는 사용역과 관련된 언어의 다양한 양상에는 사용자(남자, 여자; 어른, 아이 등), 언어 사용의 양식(구어, 문어), 매체(신문, 잡지, 전자출판물), 담화의 주제(인문, 사회, 자연, 일상생활), 언어 사용의 목적(정보 전달, 예술적 감흥) 등이 있다(강범모 외, 2000:17). 본고의 목적은 구어 말뭉치를 바탕으로 한 사용역 변이 연구를 정리하는 것으로, 실제로 말뭉치 기반 연구가 이루어진 다음과 같은 구어 관련 사용역이 포함된 연구를 대상으로 하고자 한다.

(2) 구어와 관련된 사용역
: 구어/문어, 공적/사적([+공식성]/[-공식성]), 대화/독백([+상호성]/[-상호성]), 남/여

그런데 말뭉치 기반 연구가 활성화되기 이전에도 구어 자료를 분석한 구어 어휘 연구는 꾸준히 진행되어 왔다. 그러나 본고에서는 연구 방법론의 측면에서 볼 때 변이 연구의 관점에서 사용역에 따라 달리 나타나는 구어 의미 파악에 초점을 맞춘 것이다. 이 때 사용역의 유형에는 구어/문어의 대별도 있을 수 있지만, 더 세분하여 위의 (2)와 같이 발화 환경과 관련된 다양한 사용역이 설정될 수 있다. 이는 말뭉치가 갖추어져 뒷받침이 되어야만 연구가 가능하다.

사용역이 달라진다는 것은 무엇을 의미하는가? 예를 들어 Lindemann & Mauranen(2001)에서는 Michigan Corpus of Academic Spoken English라는 학술 관련 구어 말뭉치를 연구함에 있어, 5개의 대화 사례[11]로 나누어 구성되었다는 점을 이용하였다. 이렇게 다른 대화 사례들은 모두 학술적 텍스트로서의 특징을 가지면서, 특정 담화표지-여기서는 'just'를 살펴보았다.-의 5가지 담화 기능[12]의 빈도를 측정해 본 결과 대화 사례에 따른 차이가 드러나게 되었다. 이 차이는 어디에서 기인하는가? 그리고 어떻게 해석할 수 있는가? 이러한 연구 과정을 통해 특정 담화표지의 의미를 파악하게 되는 것이다. 즉, 하나의 학술적 대화로 묶일 수 있는 영역에 대해서도 변이를 나누어 비교하게 되면 그 의미적, 기능적 차이를 드러내게 되는 것이다.

3. 사용역 변이 연구 정리

이 장에서는 구어를 대상으로 한 사용역 변이 연구를 정리하고자 한다. 정리하는 순서는 어떤 유형의 말뭉치를 이용하였는가에 따라 원시 말뭉치, 형태 분석 말뭉치, 음성 말뭉치 등으로 나누고, 각각의 말뭉치를 이용하여 어떤 사용역 연구가 진행되었는가에 따라 구분하여 기술하고자 한다.[13]

---

11) 음악시험, 물리학강의, 철학토론, 면역학 실험실 모임, 관리자 모임으로 나누어져 있다.

12) 5가지 기능이란 "축소(minimizing), 강조(emphasizing), 특수화(particularizing), 상세화(specificatory), 애매한 것(ambiguous)"을 말한다.

13) 3장의 전개를 말뭉치에 따른 구분보다 사용역에 따른 구분으로 나누는 것이 더 적절할 것 같다는 심사위원의 지적이 있었다. 그러나 사용역의 차이보다는 원시 말뭉치, 주석 말뭉치,

3.1. 원시 말뭉치의 이용

원시 말뭉치를 이용한 연구에는 담화표지에 대한 연구가 가장 많다. 담화표지란 의미적 차원에서 불필요한 표현이나, 기존의 의미, 기능에 새로운 기능을 획득하게 된 표현이 언어 사용이나 담화 차원에서 새롭게 존재 이유를 갖고 사용되는 것을 말한다. 흔히 담화표지의 의미는 고정된 것이 아니라 변화하는 과정에 있는 떠다니는 의미로 파악한다.

구어 원시 말뭉치는 용례 추출을 통해 기본적으로 담화표지의 기능에 대한 질적, 양적 연구를 가능하게 해 준다. 비록 장르 변이 연구는 아니지만 김명희(2005)에서는 세종계획 구어 원시 말뭉치 전체(420만 어절)에서 “뭐, 어떻게, 왜, 무슨, 어디, 언제, 누구”와 같은 의문사의 담화표지 기능을 확인하였다.

국외에서는 LLC 말뭉치를 이용하여 문법서에서 확대어(amplifiers)로 알려진 두 가지 부사적 범주를 기술하였는데, 극대어(maximizers)와 증폭어(booster)가 그것이다. 이 범주들이 어떻게 사용되고, 이와 함께 사용되는 연어의 분포를 살핌으로써 말뭉치 기반 담화 연구 방법론을 잘 보여주었다(안동환 역, 2010:222~228).

원시 말뭉치를 이용한 사용역 변이 연구에는 남길임, 차지현(2010)을 들 수 있다. 여기서는 담화표지 ‘뭐’를 대상으로 하여 출현 환경과 분포 분석을 시도하였다. 이 논문에서는 ‘사용 패턴(usage pattern)’의 개념으로 담화표지의 기능에 접근하였는데, 사용 패턴이란 “하나의 중심어와 그 중심어와 빈번하게 자주 나타나는 문법, 어휘, 담화 환경 등의 언어적 요소의 결합”을 말한다(남길임, 차지현, 2010:96). 여기서 사용한 말뭉치는 사적 독백과 사적 대화 각각 5만 어절이며 형태 분석 말뭉치가 아닌 원시 말뭉치를 이용하였다. 연구 결과는 크게 두 가지 인데, 첫째 전체의 ‘뭐’와 담화표지로서의 ‘뭐’ 모두 독백에서 더 많이 쓰인다는 것이다. 두 번째 결과는 절 경계 위치에서는 접속부사 패턴, 어휘적 담화표지 패턴, 양태어미 패턴의 순으로 많이 쓰이고, 그

음성 말뭉치와 같은 말뭉치의 유형이 더 큰 범주에 속하는 것으로 보아 이를 큰 제목으로 하고 각각의 내용 속에서 다시 사용역에 따른 정리를 하였다.

렇지 않은 자유 위치에서는 대용어 패턴과 나열 및 예시 패턴, 수사의문문 패턴의 순으로 많이 쓰인다는 것이다.

그런데 이러한 사용역 변이 연구가 일반적인 담화표지 연구와 다른 점은 무엇일까? 남길임, 차지현(2010)의 연구 결과의 첫 번째는 담화표지 '뭐'가 독백에서 더 자주 쓰인다는 점을 발견하였는데, 이는 청자 중심적이라기보다는 화자 중심적인 의미 기능으로 더 많이 쓰인다는 것으로 해석할 수 있다. 그리고 일반적인 담화표지 연구가 담화표지가 되기 이전의 기본 의미와 확장 의미의 파악에 초점을 맞추고 있는데 반해, 남길임, 차지현(2010)의 연구는 결합 관계를 함께 파악함으로써 다양한 사용 패턴으로 나누어 보았고, 이를 사적 독백과 사적 대화의 상황에 맞춰 해석하였다는 점이다.

Lam(2009, Aijmer(2013:8)에서 재인용)에 따르면 담화표지 연구에 있어서 누가, 무슨 목적으로 사용하는지에 대한 다양한 상상을 할 수는 있지만, 이렇게 다양한 사회적 상황이나 텍스트 유형에 따른 담화표지 연구는 드물다고 한다. 이는 다양한 변이가 주어진다면 변이마다 다른 연구를 할 수 있지만 현실적으로 다양한 구어 말뭉치가 존재하지 않아 이를 뒷받침하지 못하는 측면도 있는 것이다.

3.2. 형태 분석 말뭉치의 이용

형태 분석 말뭉치를 이용한 연구는 적용한 사용역에 따라 구어/문어, 공적/사적과 대화/독백, 남/녀로 나누어 볼 수 있다.

3.2.1. 구어/문어 사용역

구어와 관련된 사용역 연구에서 가장 기본이 되는 것은 어떤 어휘가 과연 구어 어휘인지 아닌지를 파악하는 작업일 것이다. 학습자 사전으로 유명한 <Longman Dictionary of Contemporary English>는 비록 그 구분 기준은 나와 있지 않지만, 말뭉치를 참고하여 표제어 중 일부를 다음의 예처럼 '*Spoken*'과 '*especially Spoken*'으로 사용역 정보를 나누어 표시하였다(안의정, 2009:50).

(3) *Spoken* : I mean, I see what you mean
especially Spoken : intend sb to do sth

그러나 국내 사전의 경우 어떤 어휘가 구어 어휘인지 아닌지를 파악하는 연구는 그리 많이 이루어지지 못하였고 사전에 반영되어 있지 못하다. 안의정(2009:49)에서는 국내의 유형별 사전 4종에 기술된 구어 어휘 표시 여부를 조사한 바 있다.

<표 2> 사전별 구어 표제어 수와 표현 방식

사전명	전체 표제어 수 (단어)	구어 표제어 수 (단어)	구어 표제어의 비율 (%)	표현 방식1 (뜻풀이 구획)	표현 방식2 (참고란)
<조선말대사전> (1992년)	293,073	1,286	0.43	(말체)	(없음)
<연세한국어사전> (1998년)	49,561	94	0.19	[입말에서], [입말로], [입말투로].. ~의 입말.	입말/입말투/입말체/입말에 쓰임 (입말에서 주로 쓰인다)
<표준국어대사전> (1999년)	509,076	66	0.01	구어적으로 이르는 말/구어체로/구어체(에서 많이 사용된다)	(없음)
<외국인을 위한 한국어학습사전> (2006년)	7,551	150	1.98	(없음)	주로 말할 때 쓴다

위 <표 3>을 통해 알 수 있듯이 전체 표제어 중 구어 표제어의 비율은 0.01~2% 정도로 매우 낮은 편이다. 이는 구어 표제어가 기술되지 않았다는 의미가 아니라, 해당 표제어가 구어의 사용역에서 많이 나타나는가를 판단하지 못하였거나 사전에 따라 사용역 표시를 중요하게 생각하지 않았기 때문으

로 판단된다. 따라서 안의정(2009)에서는 사전 기술을 목적으로 말뭉치 비교를 통한 어휘의 구어성 판별 작업이 진행되었다.

여기서 이용한 말뭉치는 모두 10m[14] 어절이며 문어 90%, 구어 10%로 구성되었고 Log-likelihood 계수(또는 G^2)를 이용하여 통계치를 계산하였다. 말뭉치의 구성에 있어서 하위 말뭉치의 크기가 다른 경우 단순 빈도로 직접 비교하는 것은 의미가 없다. 따라서 말뭉치간의 유의미성 검증을 위한 여러 통계 기법 중 British National Corpus를 대상으로 한 Leech, Rayson, Wilson(2001)의 연구에서 사용된 G^2를 이용한 것이다.[15] 그 결과 구어성이 높은 어휘들과 함께 문어성이 높은 어휘들이 판별되었다. 구어성이 높은 어휘에는 일반부사, 감탄사들, 그리고 비격식적 종결어미들이 많이 포함되어 있다.[16]

그런데 안의정(2009)의 G^2값에는 다음과 같은 문제점이 있다. 첫째 측정 과정에서 이형태 통합[17]이 이루어지지 않았다. 종결어미 '-어'는 '-아'나 '-여'

14) m은 백만을 의미한다.

15) G^2값의 측정과 해석에 대한 설명은 안의정(2009:30~38)을 참고할 것.

16) 구어성이 높은 어휘로 판별된 100개의 어형을 들면 다음과 같다.
거(의존명사), 어(감탄사), 아(감탄사), 뭐(감탄사), 음(감탄사), 근데(부사), -어(종결), 예(감탄사), -구(연결), 하다(동사), 그거(대명사), 막(부사), 이렇게(부사), 그러다(동사), 안(부사), -야(종결), 이거(대명사), -잖아(종결), 뭐(대명사), 네(감탄사), -죠(종결), 그런(관형사), 가지다(보조용언), 내(대명사), -지(종결), 인제(부사), 응(감탄사), 되게(부사), -어요(종결), -는데(연결), -여(종결), 그렇게(부사), -에요(종결), 되다(동사), 얘기(명사), 진짜(부사), 그래서(부사), 일(수사), 보다(동사), 두(보조사), -니까(연결), 지금(부사), 그냥(부사), -시-(선어말), 좀(부사), 제(대명사), 십(수사), 다(부사), 요(보조사), -거든요(종결), 다음(명사), 딱(부사), ㄴ(보조사), 에(감탄사), 가다(동사), 애(명사), -면(연결), 이제(부사), 삼(수사), -잖아요(종결), 아니(감탄사), 이(수사), 많이(부사), -아(종결), ㄹ(목적격), -냐(종결), 굉장히(부사), 개(명사), 이런(관형사), 저(감탄사), 이(감탄사), 부터(부사격), 야(감탄사), -거든(종결), 갖다(보조용언), 거기(대명사), -구(종결), 그(관형사), 한테(부사격), 그럼(부사), 그렇다(형용사), 랑(부사격), 이러다(동사), 오다(동사), 같다(형용사), 왜(부사), 아까(부사), -겠-(선어말), 아이(감탄사), 또(부사), 오빠(명사), -는데(종결), 못(부사), 나오다(동사), 저희(대명사), 오(수사), 이(보격), -ㄴ데(연결), 너무(부사), 애(명사)

17) 안의정(2012:244)에 따르면 구어에서는 한 형태소에 대해서 여러 가지 변이형태가 나타날 수 있는데, 이 경우 통합하여 대표의 형태로 빈도를 보이게 되면 구어의 특징을 밝혀내기가 힘들다고 한다. 따라서 이 경우 이형태의 빈도가 모두 확인이 가능하도록 해야 한다. 예를 들어 연결어미 '-려고(의도)'가 '-려고, -려구, -르려고, -르려구, -르라고, -르라구'로 나타나는 경우 다음과 같이 처리하는 것이 바람직하다.

예.	-려고	EC	123
	-르려구		48
	-르려고		32

와 이형태 통합이 된 후에 점수가 측정되어야 하고, 연결어미 '-구'는 '-고'의 구어적 이형태이므로 각각을 측정하는 것이 바람직한데 이러한 점이 고려되지 않았다. 둘째는 서상규(2013ㄱ:84)에서 지적한 것인데 구어성이 플러스로 판정되어 구어 올림말로 판정된 경우라도 실제로는 문어 말뭉치에서도 적지 않은 빈도로 나타나는데, 이러한 특성이 반영되지 않은 점이다. 마지막으로 구어 어휘가 아닌데도 점수가 높게 나온 경우가 있다. 예를 들어 숫자 부류와 '오빠'와 같은 친족어들이 그것이다. 이들을 어떻게 해석해야 하는가의 문제가 있다.

안의정(2009)에서는 이렇게 추출된 구어성 어휘를 사전에 반영하기 위한 작업이 진행되었고, 의미 변화가 일어난 유형(강조, 지시 영역의 축소, 감탄사, 어미/조사 등)과 의미 변화가 일어나지 않은 유형으로 나누어서 기술 방안을 제안하였다.

그 밖의 구어와 문어의 사용역을 이용한 연구를 정리하면 다음의 (4)와 같다.

(4) 구어/문어 사용역 변이 연구
- 남길임(2010), "'아니다'의 사용패턴과 부정의 의미"
- 전영옥, 남길임(2005), "구어와 문어의 접속 표현 비교 연구-'그런데, -는데'를 중심으로"
- 조영순(2010), "세종말뭉치에 기초한 양보와 대조어미의 사용 분석"
- 최준, 송현주, 남길임(2010), "한국어의 정형화된 표현 연구"
- 한영균, 고은아(2011), "유의적 정도부사의 빈도, 분포, 결합관계의 분석과 그 활용"

(4)의 연구 목록에서 구어 어휘 의미 연구 방법론을 살펴보기 위해 각각의 연구에 대하여 연구 대상과 자료의 구성, 정보 추출 방법, 결과 분석 등을 정리해 보면 다음의 <표 3>과 같다.

-르라구	19
-르라고	9
-려구	8
-려고	7

<표 3> 구어/문어 사용역 변이 연구

	남길임(2010)	전영옥, 남길임(2005)	조영순(2010)	최준, 송현주, 남길임(2010)	한영균, 고은아(2011)
연구대상	'아니다'의 사용패턴	접속표현 (그런데, -는데)	양보와 대조어미	정형화된 표현	유의적 정도부사
말뭉치구성	구어, 문어 각 10만 어절	구어 4.5만, 문어 4.9만 어절	구어(80만) 창작문(100만) 학술산문(400만) 뉴스(335만)	자유대화 29만, 학술구어 22만, 학술문어 100만	137m 어절 (문어 75%, 구어체 17%, 구어 8%)
정보추출방법	각 200개의 용례를 추출하여 분석	용례를 의미별로 분류	18개의 양보/대조 어미 중 88%를 차지하는 상위 4개 (-는데, -지만, -아도, -(으)나)의 어미 분석	고빈도 형태소 연쇄의 파악을 위해 N-gram (5-gram) 기법 이용	1. '아주/매우/너무/되게'의 전체 및 하위 영역에서 빈도 추출 2. N+1 위치에 나타나는 연어 분석
연구결과	1. 담화부정은 구어에서 더 많이 나타남. (36:8) 2. 문법적 패턴은 "X가 Y가 아니- > X('것'절)가 아니- > 담화부정의 서술형" 순의 빈도를 보임.	1. 구어는 접속부사가, 문어는 접속어미가 많이 쓰임. 2. 담화 화용적 기능을 밝혀냄(화제 전개, 발화 상황 통제, 발언권 유지), 이들이 구어에서 더 빈번하게 쓰임.	1. 각 어미의 양보/대조/기타의 의미 표현 분포 2. 4개 사용역별 어미 사용 분석	1. 형태 통사적 특성 파악 : 의존명사 '수, 것/거'를 포함한 구성이 많음 2. 사용역별 담화 기능 분포와 담화 기능적 특성 파악	'아주'와 '매우'의 빈도, 분포, 결합관계를 바탕으로 구체적 용법 기술

<표 3>의 연구를 정리해 보면 다음과 같다. 먼저 연구 대상에 관한 것이다. 연구 대상은 '아니다'(부정), 정도부사, 양보/대조 어미, 접속표현, 정형화된 표현 등으로 다양하다. 하지만 특정한 어휘 하나, 비교되는 표현 둘 정도를

대상으로 하고 있어 폭넓은 연구가 진행되지 못한 측면이 있다. 예를 들어 최준, 송현주, 남길임(2010)에서 이루어진 정형화된 표현에 대한 연구도 5개 형태소 연쇄 중 고빈도를 선정하여 연구하였지만, “할 수 있는, ~기 때문이다, 그런 거 같애” 등과 같이 의존명사 ‘수, 것/거’를 포함한 구성에 한정될 수밖에 없었다.

두 번째는 사용된 구어 말뭉치의 크기와 유형에 관한 것이다. 말뭉치의 크기는 10만 어절 내외부터 대용량에 이르기까지 다양하지만 대부분 세종 구어 말뭉치를 이용하였다. 한영균, 고은아(2011)에서는 구어, 문어 이외에도 이른바 준구어로 분류할 수 있는 드라마 대본, 시나리오, 뉴스 스크립트와 같은 구어체 자료를 함께 이용하기도 하였다. 하지만 세종 구어 말뭉치에는 설득의 텍스트, 학술적 텍스트는 있지만, 상업적인 텍스트나 다양한 채널[18]에 따른 텍스트 등이 부재한 상황이어서 이를 이용한 연구도 수행되지 못하였다.

마지막으로 방법론에 있어서는 대부분 용례 분석과 함께 빈도 비교 작업을 이용하였다. 그리고 일부 연구에서는 특정 태그셋이 추가되기도 하였다.

그렇다면 이들의 연구에서 의미 연구는 어떻게 진행이 되었는가? 먼저 남길임(2010)의 연구에서는 담화부정이 구어에서 더 많이 나타난다는 것을 밝혀내었다. 담화부정이란 문장 또는 절 내에서 부정의 대상이 되는 명제적 요소를 찾을 수 없는 것을 말한다. 이러한 요소는 문맥 내에서 복원하기 힘들거나 복원해서 쓰이지 않는다는 특징이 있다. 따라서 ‘아니다’의 의미 기술은 문장의 명제를 부정하는 것을 넘어서 전체 담화 맥락의 차원에서 논의되어야 할 필요가 있다(남길임, 2010:62). 그리고 전영옥, 남길임(2005)에서는 “화제 전개, 발화 상황 통제, 발언권 유지”와 같은 접속부사, 접속어미의 담화 화용적 기능을 밝혀내었고, 이들이 구어에서 더 빈번하게 쓰임을 알아내었다. 마찬가지로 최준, 송현주, 남길임(2010)에서도 담화 기능에 초점을 맞추었다. 이 연구에서

18) 다양한 채널의 텍스트란 동일 주제나 영역의 다양한 채널로 구성된 텍스트를 말한다. 예를 들어 Bowker(2013)에서는 상업적인 주제(고용인 모집)로 이루어진 프레젠테이션 발화(구어), 파워포인트 슬라이드(문어), 전자 뉴스레터(문어) 등을 비교하였다. 이와 관련하여 국내에서는 성연숙(2005)에서 동일한 내용의 텍스트를 대본과 이를 바탕으로 하여 실제로 발화된 내용을 전사한 구어를 비교하는 연구를 수행한 바 있다. 이러한 연구를 통해서 구어로 생산되는 발화의 특성을 잘 포착할 수 있다.

는 정형화된 표현의 사용역별 담화 기능의 분포와 담화 기능적 특성을 파악하였다. 여기서 말하는 담화 기능은 “가능성, 추측, 인식, 의무”와 같은 태도 표현 기능, “주제 도입/명세화, 대조, 이유, 순서”와 같은 담화 조직 기능, 그리고 지시 표현 기능을 말한다. 모든 사용역에서 태도 표현 > 담화 조직 > 지시 표현 순의 빈도를 보였다. 조영순(2010)의 연구는 기존의 양보와 대조의 의미 연구가 의미와 구조의 이론적 연구에 집중하여 진행된데 비해, 사용(use)이라는 관점으로 표현과 의미의 사용 양상을 조사하였다. 마지막으로 한영균, 고은아(2011)의 연구는 학습자 사전의 용법 기술의 관점에서 진행이 되었는데, 학습자들은 어휘의 의미를 아는 것만으로 용법을 익히기 힘들기 때문에 빈도가 높고 사용역의 분포가 넓은 정도부사의 경우 사전이 의미 기술뿐 아니라 빈도와 분포의 전형성을 보여주는 방향으로 기술되어야 함을 강조하였다.

3.2.2. 공적/사적과 대화/독백 사용역

공적/사적과 대화/독백 사용역을 이용한 연구는 다음의 <표 4>와 같이 두 개의 연구가 있다.

<표 4> 공적/사적과 대화/독백 사용역 변이 연구

	현영희, 남길임(2011)	한송화(2013)
연구 대상	삽입 표현	접속부사
말뭉치 구성	8만6천 어절(공적 독백과 공적 대화 각각 2만1천, 사적 독백 2만2천, 사적 대화 2만1천)	274만2천 어절(구어-격식 28만5천, 비격식60만6천, 문어-소설 42만9천, 신문 50만4천, 학술 50만3천, 기타 51만2천)
정보 추출 방법	주석 태그셋 설정 ; 단순 삽입(형식, 내용), 대치 삽입(형식, 내용), 도치 삽입(형식, 내용)	빈도 비교
연구 결과	1. 형식 삽입의 출현 빈도는 [-상호성], [+공식성]과 관련이 깊으며, 대치 삽입은 [+상호성]과 비례. 2. 도치 삽입은 [+상호성], [-공식성]의 특성과 관련, 부가어의	1. 구어에서 더 다양한 접속부사가 사용되고 빈도도 높음. 구어 중에서 사적 독백이나 토론, 회의에서 더 많이 사용됨. 2. 문어에서 신문은 접속부사의 사용 양상이 단순하고 빈도도 낮음. 학술

	삽입 빈도가 가장 높음. 특히 공적 자료보다 사적 자료에서 8배 정도 높게 나타남.	텍스트는 고르게 사용된 반면, 신문은 대립관계 접속부사가, 소설은 조건관계 접속부사가 많이 쓰임.

현영희, 남길임(2011)에서 연구 대상으로 삼은 삽입 표현이란 “단일 화자의 발화 내에서 의미적으로 또는 통사적으로 발화 맥락을 끊고 들어가는, 삭제 가능한 표현(현영희, 남길임, 2011:63)”을 말하며, 통사, 의미적 차원에서 불필요한 표현으로 여겨져 온 군말, 입버릇, 말실수와 문맥 내용 보충, 의사소통 흐름 조정 등을 목적으로 삽입되는 모든 담화 구성 요소를 포함하는 다소 큰 개념이다. 여기서는 형태 분석 외에 삽입의 유형을 구분하는 특별한 태그가 추가된 후 연구가 진행되었으며, 결과의 해석에 있어서 자료의 [상호성]과 [공식성] 자질이 이용되었다.

한송화(2013)에서는 말뭉치를 구어와 문어로 나눈 후 다시 격식과 비격식으로 구분하여 연구가 진행되었다. 여기서는 사용범위(range)의 관점에서 말뭉치 분석이 시도되었다. 사용범위란 “하나의 어휘가 얼마나 다양한 텍스트에 사용되는가를 측정한 것(신동광, 2008:27)”을 말하는 것인데, 이러한 분석을 통해 학습자가 말뭉치를 통해 다양한 텍스트와 담화에 대한 축적된 경험적 증거를 제공받을 수 있을 것으로 기대하였다.

3.2.3. 남/여 사용역

남/여의 사용역이 반영된 연구는 김혜영, 강범모(2010)가 있다. 이는 의미운율에 대한 연구로 성별, 상황별 구어 말뭉치 총 34만 4천 어절이 이용되었는데 그 구성은 다음과 같다.

(5) 사적 대화 : 남, 여 각각 9만1천 어절
공적 대화 : 남, 여 각각 8만 어절

이 연구에서는 강조적 정도부사인 “가장, 매우, 몹시, 무척, 아주, 너무, 되게, 상당히, 진짜, 정말”의 영역별 빈도 분석이 이루어졌다. 그 결과 문어보다

구어에서, 공적 대화보다 사적 대화에서, 남성보다 여성이 강조적 정도부사를 더 많이 사용한다는 것이 밝혀졌다. 문어보다 구어에서 더 많이 쓰이는 현상은 부사의 일반적인 특성이면서, 강조적 정도부사가 더 특징적인 것은 그것이 명제 내용에 대한 주관적 심리 상태에 놓인 화자의 주장을 포함하는 성격이 강하기 때문으로 해석할 수 있다. 또 사적 대화에서 많이 나타나는 것은 발화 상황의 친밀도와 관련이 있으며, 여성이 남성보다 더 많이 사용하는 것은 여성의 전반적인 대화 태도와 특성에 기인하는 것으로 언어 보편적인 현상으로 해석하였다.

3.3. 음성 말뭉치의 이용

음성 말뭉치를 이용한 연구는 2장에서 언급한 Lindemann & Mauranen (2001)의 연구를 들 수 있다. 국내의 연구에는 음성 말뭉치를 이용한 사용역 변이 연구가 없기 때문에 부득이하게 국외의 연구를 언급하게 되었다. 비록 변이 연구는 아니지만, 송인성, 신지영(2013)에서는 17만 어절의 음성 말뭉치를 이용하여 담화표지에 대한 연구를 시도하였다. 여기서는 담화표지 '좀'의 기능과 형태([좀], [쫌])에 따른 운율적 특성을 파악하였다. 말뭉치의 구성은 친밀도가 높은 3명의 화자의 자연스런 발화로 구성되어 있는데, 총 57명이 참여하였다. 그런데 만약 이 연구에서 친밀도를 변이로 두었다면 담화표지의 실현형과 기능, 의미와의 관계를 밝히는 변이 연구가 가능했을 것으로 보인다. 친밀도는 김유정(2011)의 담화의 형식 중 지위와 함께 중요한 청화자 관계를 표시하는 지표인데, 친밀한 관계인지의 여부와 상하의 관계인지에 따라 담화의 형식이 달라질 것으로 예측되기 때문이다. 이렇게 달라진 형식은 담화표지의 출현 빈도와 음성적 실현형에 영향을 미칠 수 있을 것이기 때문에 친밀도를 변이로 설정한 연구가 가능할 것이다.

Lindemann & Mauranen(2001)의 연구는 1.7m 어절의 Michigan Corpus of Academic Spoken English에서 5개의 하위 영역에서의 'just'의 담화 기능의 분포를 파악하고, 'just'의 축약형 발음의 음성학적 특징을 분석하였다.

그 결과 의미 기능에 따라 모음의 실현 여부와 같은 음성학적 특징이 달리

나타남을 알 수 있었다. 즉 음성 말뭉치의 발음 정보가 의미 판별과 관련됨을 밝혀내었다.

지금까지 원시 말뭉치와 형태 분석 말뭉치, 그리고 음성 말뭉치를 추가적으로 검토한 사용역 변이 연구의 경향을 살펴보았다. 그런데 안의정(2009)을 제외한다면 전체 구어 말뭉치를 분석 대상으로 삼은 연구는 부재한 상황이다. 대부분의 연구가 말뭉치를 특정 단어나 표현의 용례 비교로 사용하였기 때문이다. 담화표지의 연구에 있어서 이런 상황은 극명하게 드러난다. 영어의 경우 Altenberg(1990:183, 안동환 역(2010)에서 재인용)에서는 전체 5만 어절의 LLC 말뭉치를 분석하여 4516번의 담화 항목이 사용되었으며, 이들은 텍스트 속에서의 기능에 따라 16개의 범주로 구분된다는 것을 밝혀내었다. 그러나 국어의 경우에는 말뭉치 분석을 전제로 전체 담화표지의 목록과 기능에 따른 유형 분류를 시도한 연구는 아직까지 없다.[19][20] 따라서 담화표지의 경우 전체 말뭉치를 대상으로 한 폭넓은 연구가 이루어질 필요가 있다.

4. 의미주석 말뭉치의 활용

4.1. 의미주석 말뭉치의 개념과 구축 방법

의미주석 말뭉치란 문맥에 출현하는 각 어휘의 의미를 특정 사전의 세부 의미항목에 대응시켜 주석한 것으로, 의미장(semantic field)에 따른 의미 주석 말뭉치와 어휘별 의미 범주를 기준으로 주석한 말뭉치(sense tagged corpus)로

19) 세종 구어 말뭉치의 경우에도 담화표지를 따로 구분해 내기 위해 특정 시기에 구축된 형태 분석 말뭉치 중 일부에 담화표지를 'DM'으로 형태소 분석을 시도한 적이 있었다. 담화표지는 감탄사를 포함하는 개념이나 기존의 감탄사 목록은 '감탄사(IC)'로 태깅하고, 전사상에서 '~'를 붙여 표시된 어휘에 한하여 담화표지로 태깅하였는데, 그 예는 다음과 같다. "거, 그, 그래가지고(구), 그래갖고(구), 그런, 막, 말이지, 무슨, 뭐, 아, 어, 어떤, 으, 음, 이, 이런, 인제, 저, 저기, 저런"

20) 오승신(1995)에서는 담화표지가 아닌 간투사의 개념으로 유형 분류를 시도한 바 있다. 여기서는 청자가 수신자가 되느냐의 여부에 따라 의사전달적 간투사와 비의사전달적 간투사로 분류하고 있다.

구별할 수 있다(서상규, 김한샘, 2000:252~253). 보통 의미주석 말뭉치라 하면 전자를 가리키는 것으로 후자와 같이 다의어 구분까지 주석된 의미주석 말뭉치는 매우 드문 편이다.

의미주석 말뭉치를 구축한 목적은 어떤 단어의 의미와 텍스트의 유형을 함께 살펴봄으로써 그 단어의 어느 의미가 주로 어떤 텍스트에서 나타나는지, 그리고 텍스트의 성질에 따라서 의미 용법에 어떠한 특징이 드러나는가 하는 것들을 관찰하기 위한 것이다(서상규, 2013ㄴ:284). 실제로 직관에만 의존해서 단어의 여러 의미 중 어떤 의미가 가장 많이 쓰이고 두루 쓰이는지는 판단할 수가 없는 것이다.

본고에서 살펴볼 의미주석 말뭉치는 <한국어 기본어휘 의미 빈도 사전>(2014)을 위해 구축된 자료로 100만 어절 규모의 한국어교육 표준말뭉치로 아래 (6)과 같이 구성되어 있다.

(6) 문어(86.3%): 교양해설산문(13만2천), 예술산문(22만4천), 실용산문(3만5천), 사적저술산문(7만6천), 초등학교교과서(28만7천), 한국어교과서(9만6천)
구어(13.7%): 녹음전사(7만), 준구어(6만4천)

(6)의 자료는 모두 218개의 텍스트로 구성되었으며, 한 텍스트당 5천 어절이 넘지 않는 원칙을 가지고 구축되었다. <한국어 기본어휘 의미 빈도 사전>은 (6)과 같이 구성된 말뭉치에서 10개 이상의 텍스트에 나타나면서 그 빈도수가 15이상인 총 7203개를 대상으로 하였다. 이 어형들은 모두 고빈도 어형이기 때문에 이들이 전체 말뭉치에서 차지하는 비중은 76%에 해당된다.[21] 즉 모든 어형에 대해 의미주석을 한 것은 아니지만, 전체 말뭉치의 4분의 3에 해당하는 어형에 대해 주석 작업이 이루어진 것이다. 그리고 텍스트 출현 분포를 살폈다는 것은 이 작업이 문어/구어 사용역뿐 아니라 한송화(2013)에서와 같이 사용범위를 고려한 연구라는 점이다.

이 말뭉치의 특징에 대해 한 가지 더 언급할 것은 말뭉치 구성에서 가장

21) 자세한 구성 정보와 구축 방법은 서상규(2013ㄴ:228~245)를 참고할 것.

중요한 요소인 균형성을 갖추었다는 점이다. 균형성은 단순히 문어와 구어를 똑같은 비율로 구성함으로써 이루어지는 것은 아닌데(서상규, 한영균, 1999:34~35), 그 이유는 우리의 일상에서 구어와 문어의 비중이 같지 않기 때문이다.

이렇게 출현빈도와 텍스트 빈도를 고려한 말뭉치를 대상으로 하여 의미 주석 작업이 이루어졌다. 의미 주석 작업은 <연세한국어사전>(1998)의 의미 항목을 기준으로 하여 진행되었다. 이는 균형 말뭉치를 기반으로 하여 사전에 누락된 구어 의미(sense)를 찾기 위한 작업이라고 해석할 수 있다.

4.2. 의미주석 말뭉치의 활용

4.1에서 소개한 의미주석 말뭉치는 어떻게 활용할 수 있는가? 서상규(2013ㄴ)에서는 의미 빈도 사전을 활용한 어휘 연구로, 품사통용어와 다의어의 의미 분포, 유의어와 반의어의 의미 대조, 준말과 본딧말의 의미 대조, 파생어 간의 의미 대조에 대해 연구한 바 있다. 여기서는 구어 어휘와 관련하여 구어 사용역에서 두드러진 쓰임을 보여 추가된 의미에 대해 살펴보기로 한다.

의미 태깅 작업은 기존 사전을 중심으로 진행되었는데, 이 때 사전에 없는 의미가 발견이 되면 사전을 수정하면서 진행하였다. 구어 텍스트에서 의미가 발견되어 사전에 추가된 어휘에는 다음과 같은 것들이 있다.

(7) 구어 사용역에서 의미가 추가된 어휘
감탄사 : 그, 어, 뭐, 이제, 인제, 응, 아, 네, 아이, 글쎄, 아주, 어어, 정말
부사 : 막, 딱, 그래서, 지금, 근데, 그때, 그냥, 또, 한번
명사 : 사람, 애, 생각, 오빠, 정도, 말, 전, 엄마, 고, 시간, 필요
의존명사 : 때문, 거, 것, 번, 가지
대명사 : 이, 이거, 그거, 그, 우리, 저거
수사 : 둘
관형사 : 어떤, 어느, 그런, 한
동사 : 그러다, 하다, 오다, 그러다, 위하다, 나다, 못하다, 되다, 나오다, 갖다, 잘하다, 이러다, 살다, 모르다, 데리다, 알다, 찾다

형용사 : 있다, 그렇다, 좋다, 중요하다
지정사 : 이다, 아니다
보조용언 : 주다, 않다, 하다
연결어미 : -아서

즉 위의 (7)은 특정 갈래뜻(sense)이 구어에서 15번 이상 쓰였는데도 기존 사전에 기술되어 있지 않은 의미를 가진 단어들이다. 대체로 감탄사와 부사가 많고 다의어들이 많이 포함되어 있다. 이 중에서는 기존 사전에 없는 단어, 예를 들면 "그(감탄사), 이케(부사), 고(명사), 어어(감탄사)"과 같은 것들도 있다. 이러한 목록이 있다면 3장과 같은 연구에서 다룬 대상 어휘를 더 찾아낼 수 있을 것이다.

강상호(1989:64-65)에서는 구어체의 의미론적 특성으로 일부 단어들이 구어에서 특수한 쓰임을 보인다고 하였다. 예를 들어 "다(부사), 이것, 보다(동사), 놓다, 죽다, 가만있다, 뻗다, 그것" 등이다. 이러한 현상은 의미주석 말뭉치에서 확인해 볼 수 있으며, 구어에서 두드러진 쓰임을 보여 추가된 의미의 예를 보이면 다음과 같다.

(8) ㄱ. 다(부사) 예) **다**들 그 사람을 싫어해(12, Text=13, Freq2=21)[22]
예) 속이 **다** 시원하다(8, Text=7, Freq2=10)
ㄴ. 뭐(대명사) 예) 그건, **뭐**야, 싫단 그런 말이다(16, Text=6, Freq2=16)
ㄷ. 아니다(지정사) (관용) 아니면 예) 뭔가 잘못된 거야, **아니면**, 내 오해거나.
(25, Text=80, Freq2=140)
(관용) 아니(요) 예) **아니요**, 잘못 본 거요.
(52, Text=54, Freq2=202)

(8ㄴ)의 '뭐야'는 묻는 말이 아니라 마음에 들지 않은 대상을 가리키는 말로 사용된 것이고, (8ㄷ)의 '아니면'은 부정의 대상이 절(문장) 밖에 위치하는

22) 괄호 안의 첫 번째 숫자는 해당 의미의 구어에서의 출현 빈도를 의미하고, 두 번째 숫자는 텍스트 빈도, Freq2는 해당 의미의 전체 출현 빈도를 의미한다. 즉, "다들 그 사람을 싫어해"와 같은 '다'의 의미는 전체에서 21번 쓰였는데, 13개의 텍스트에 등장했고, 그 중 구어 텍스트에서는 12회 출현했음을 의미한다.

담화 부정의 경우이다. 마찬가지로 ‘아니요’도 응답의 ‘아니요’(감탄사)가 아니라 표면적으로 드러나지 않은 요소를 부정하기 위해 쓰인 것으로 기존 사전에 기술되지 않은 의미이다. ‘아니다’의 이러한 의미는 남길임(2010)에서도 일부 파악된 내용이지만, 의미주석 말뭉치를 이용하여 더 많은 후보를 찾을 수 있을 것이다.

의미주석 말뭉치를 이용하면 구어에서 두드러지게 많이 쓰인 의미의 분포도 쉽게 파악할 수 있다. 다음의 예를 보자.

<표 5> ‘진짜’와 ‘무슨’의 의미 분포

올림말	품사	의미항목	텍스트유형	빈도수	정규화(1m)	비율
진짜	nng	-II	교양해설산문	4	15.7	3%
진짜	nng	-II	구어	104	871.3	78%
진짜	nng	-II	사적저술산문	2	13.4	2%
진짜	nng	-II	예술산문	13	29.3	10%
진짜	nng	-II	준구어	7	63.0	5%
진짜	nng	-II	한국어교과서	3	16.8	2%
무슨	mm	-II	교과서	1	1.9	4%
무슨	mm	-II	구어	12	100.5	50%
무슨	mm	-II	예술산문	8	18.1	33%
무슨	mm	-II	한국어교과서	3	16.8	13%

위의 표를 통해서 ‘진짜’와 ‘무슨’ 모두 품사통용어로서의 의미인 의미항목 II에서 구어의 쓰임이 두드러졌음을 알 수 있다. ‘진짜’II의 의미는 “[부사적으로 쓰이어] 실지로. 참으로.”의 뜻을 말하고, ‘무슨’II의 의미는 “[감탄사적으로, 문장의 중간에서 못마땅하거나 반대하는 뜻을 나타내어] ‘당치않게, 괜히, 쓸데없이’의 뜻.”을 말한다.[23]

23) 그런데 ‘진짜’II의 의미에는 강조적 의미로 “화가 나거나 흥분해서 말할 때 입버릇처럼 쓰는 말.”로 풀이할 수 있는 다음의 예와 같은 쓰임이 있다. 이는 누락된 듯하여 추후 추가해야 할 필요가 있다.
예) 너 **진짜**, 엄마가 하지 말라고 했는데 계속 이럴 거야./ 나는 **진짜**, 그런 나쁜 사람은 상대도 하고 싶지 않아요. **진짜**.

5. 결론

지금까지 말뭉치 기반 사용역 변이 연구사 검토를 통해 구어 어휘의 의미 탐구가 어떤 방식으로 진행되었는지를 살펴보았다. 가장 많은 연구가 이루어진 말뭉치는 형태 분석 말뭉치로, 이를 이용하여 특정 담화표지와 부정, 접속 표현, 양보/대조의 어미, 정형화된 표현, 정도부사 등이 연구되었다. 여기서 밝히고자 하는 의미는 사전에 이미 기술된 명제적 의미보다는 담화·화용적 기능을 나타내는 의미들에 초점이 맞추어져 연구가 진행되었다. 4장에서는 의미주석 말뭉치를 활용한 의미 연구에 대해 검토해 보았다. 이 말뭉치를 구축한 결과 구어 어휘와 관련하여 구어 사용역에서 두드러진 쓰임을 보여 추가된 의미들을 많이 발견할 수 있었다.

이 글은 사용역 변이 연구사 정리를 통해 구어 어휘의 의미 연구 방법을 모색하는 것으로 다양한 사용역에 따른 구어 어휘의 특정한 의미 사용을 확인할 수 있었다. 방법론적인 면에서는 용례 분석이나 빈도 분석에 한정된 면이 있지만, 안의정(2009)와 같은 통계적 기법도 확인할 수 있었다.

그런데 본고에서 살펴본 결과 국내의 연구에 사용된 말뭉치는 대부분 세종 구어 말뭉치로, 다양한 주제와 변인을 다룬 말뭉치가 없으며 따라서 이를 이용한 연구도 부재한 상황이다. 향후 다양한 유형의 구어 말뭉치가 구축되어 담화적 변인에 따른 다양한 의미 연구가 이루어져야 할 것으로 보인다.

또 자료를 비교하여 보여주는 방법에 있어서도 현재까지는 단순한 기법이 많이 이루어졌다. 따라서 보다 고차원적인 통계적 기법이나 연어 분석 등이 이루어져야 할 것으로 보인다. 아울러 구어 전사 말뭉치의 여러 전사 요소를 다양하게 활용할 필요가 있다. 예를 들어 담화표지의 경우 쉼(pause)과의 연어적 분석이 중요할 것 같은데, 이러한 연구는 아직까지 수행된 바 없다.

말뭉치 기반 변이 연구는 언어 변화의 측면에서도 연구될 수 있다. 영어의 경우 본문에서도 언급한 LLC가 1960년대부터 구축되었기 때문에 최근에 구축한 말뭉치와 비교해서 연구한다면 가능한 작업이 될 수 있다. 국어의 경우도 지속적으로 말뭉치를 구축해 나간다면 향후 언어 변화를 측정하는 연구가 중요한 연구로 자리잡을 수 있을 것이다. 아울러 사전에 표시되는 사용역 표

지에 대한 연구가 관용구로 확대되어야 할 필요가 있다. 즉, 대규모의 구어 말뭉치를 대상으로 하여 사전의 관용구를 보완하고, 구어에서 많이 쓰이는 관용구를 표시하는 작업이 필요하다. 이는 향후의 과제로 남기고자 한다.

참고문헌

[논문류]
강범모·김흥규·허명희(2000), 한국어의 텍스트 장르, 문체, 유형 - 컴퓨터와 통계적 기법의 이용, 태학사.
강상호(1989), 조선어입말체연구, 평양 : 사회과학출판사.
강소영(2007), 구어와 문어 자료의 실제적 연구방법론, 한국문화사.
김명희(2005), “국어 의문사의 담화표지화”, 담화와 인지, 제12권 2호, 담화인지언어학회, 41-63.
김유정(2011), “언어사용역을 활용한 ‘죽다’류 유의어 의미 연구”, 인문연구, 제62호, 영남대학교 인문과학연구소, 85-122.
김진해(2006), “코퍼스언어학적 관점에서 본 의미의 본질”, 한국어의미학 제21호, 한국어의미학회, 75-104.
김혜영(2009), 구어에서 나타나는 정도부사의 사용의미, 고려대학교 대학원 석사학위논문.
김혜영·강범모(2010), “구어 속 강조적 정도부사의 사용과 의미”, 한국어학, 제48호, 101-129.
남길임(2010), “‘아니다’의 사용패턴과 부정의 의미”, 한국어의미학 제33호, 한국어의미학회, 41-65.
남길임·차지현(2010), “담화표지 ‘뭐’의 사용패턴과 기능”, 한글, 288호, 한글학회, 91-119.
노대규(1996), 한국어의 입말과 글말, 국학자료원.
민경모(2008), 한국어 지시사 연구, 연세대학교 대학원 박사학위논문.
박석준(2007), “담화표지화의 정도성에 대한 논의”, 한말연구 21, 한말연구학회, 87-106.
배진영(2012), “구어와 문어 사용역에 따른 정도부사의 분포와 사용 양상에 대한 연구”, 국제어문, 제54집, 국제어문학회.
배진영 외(2013), 말뭉치 기반 구어 문어 통합 문법 기술1-어휘 부류, 박이정.
서상규(2013ㄱ), “한국어의 구어와 말뭉치”, 한국어 교육 24-3, 국제한국어교육학회, 71-107.
서상규(2013ㄴ), 한국어 기본어휘 연구, 한국문화사.

서상규(2014), 한국어 기본어휘 의미 빈도 사전, 한국문화사.
서상규·김한샘(2001), “의미주석말뭉치와 전자사전의 의미기술정보”, 2001년도 제13회 한글 및 한국어 정보처리 학술대회 논문집, 252-259.
서상규 외(2013), 한국어 구어 말뭉치 연구, 한국문화사.
서상규·한영균(1999), 국어정보학 입문, 태학사.
성연숙(2005), “한국어 구어의 특징 -TV 토론 방송 대본과의 비교를 중심으로-”, 제34차 한국어학회 전국학술대회 논문집, 한국어학회.
송인성·신지영(2013), “담화표지 {좀}의 기능과 형태적, 운율적 특성”, 제65차 한국어학회 전국학술대회 논문집, 한국어학회.
신동광(2008), “영어과 기초어휘선정을 위한 제언”, 한국사전학회 제14회 학술대회 발표논문집, 한국사전학회.
안동환 역(2010), 코퍼스언어학개론(Kennedy, G., *An Introduction to Corpus Linguistics*, 1998), 한국문화사.
안의정(1998), 한국어 입말뭉치 전사 방법 연구, 연세대학교 대학원 석사학위논문.
안의정(2009), 국어사전에서의 구어 어휘 선정과 기술 방안 연구, 한국문화사.
안의정(2012), “한국어 빈도 사전 편찬을 위한 기초 연구”, 한국사전학, 제20호, 한국사전학회.
오승신(1995), 국어의 간투사 연구, 이화여자대학교 대학원 박사학위논문.
전영옥·남길임(2005), “구어와 문어의 접속 표현 비교 연구- ‘그런데, -는데’를 중심으로”, 한말연구 17, 한말연구학회.
조영순(2010), “세종말뭉치에 기초한 양보와 대조어미의 사용 분석”, 언어, 제35권 제4호, 서울대학교 언어연구소, 1127-1148.
최준·송현주·남길임(2010), “한국어의 정형화된 표현 연구”, 담화와 인지, 제17권 2호 담화인지언어학회, 163-190.
한송화(2013), “한국어 접속부사의 사용 양상 -텍스트 유형에 따른 사용 양상을 중심으로-”, 언어사실과 관점, 31권, 연세대학교 언어정보연구원, 139-170.
한영균·고은아(2011), “유의적 정도부사의 빈도, 분포, 결합관계의 분석과 그 활용 -학습자 사전의 용법 기술의 관점에서-“, 한국어 의미학, 35권, 한국어의미학회, 335-394.
현영희·남길임(2011), “담화 장르에 따른 삽입 표현의 사용 양상”, 한글 제292호, 한글학회, 55-85.
Aijmer, K.(2013), *Understanding Pragmatic Markers: A Variational Pragmatic Approach*, Edinburgh University Press.
Aijmer, K. & Stenström, A.(eds)(2004), *Discourse Patterns in Spoken and Written Corpora*, John Benjamins Publishing Company, Amsterdam/ Philadelphia.
Altenberg, B.(1994), On the functions of such in spoken and written English, in Oostdijk, N. & de Haan, P.(eds), *Corpus Based Research into Language*,

Rodopi, Amsterdam, 223-240.
Baker, P., Hardie, A., & McEnery, T.(2006), *A Glossary of Corpus Linguistics,* Edinburgh University Press.
Biber, D. & Conrad, S.(2009), *Register, Genre, and Style*, Cambridge University Press.
Bowker, J.(2013), Variation across spoken and written registers in internal corporate communication, in *Variation and Change in Spoken and Written Discourse*, Bamford, J., Cavalieri, S., and Diani, G.(eds), John Benjamins Publishing Company, Amsterdam/Philadelphia, 47-64.
Hoey, M.(2009), Corpus linguistics and word meaning, In Lüdeling, A. & Kytö, M.(eds), *Corpus Linguistics: An International Handbook* Vol.2, Walter de Gruyter, Berlin/New York, 972-987.
Leech G., Rayson P., Wilson A.(2001), *Word Frequencies in Written and Spoken English*, Longman, London/New York.
Lindemann, S. & Mauranen, A.(2001), "It's just real messy": the occurrence and function of just in a corpus of academic speech, in *English for Specific Purposes*, Volume 20-1, Elsevier, 459-475.
McEnery, T., Xiao, R., & Tono, Y.(2006), *Corpus-based Language Studies*, Routledge, London/New York.
Sinclair, J.(2007), Meaning in the framework of corpus linguistics, in Teubert, W. & Krishnamurthy, R.(eds), *Corpus Linguistics* Vol.1, Routledge, London/New York, 182-196.

[사전류]
『연세 한국어사전』(연세대학교 언어정보개발연구원 편, 두산동아, 1998년)
Longman Dictionary of Contemporary English(3th, 1995년)

서울시 서대문구 성산로 262
연세대학교 언어정보연구원(위당관 518호)
120-749
전화번호: 02-2123-4047
전자우편: snoopyjinu@gmail.com, FAX: 02-393-5001

원고 접수일: 2014년 02월 26일
원고 수정일: 2014년 03월 23일
게재 확정일: 2014년 03월 25일