영어 동사의 의미적 유사도와 논항 선택 사이의 연관성: ICE-GB와 WordNet을 이용한 통계적 검증

송상헌*
University of Washington

최재웅†
고려대학교

Sanghoun Song and Jae-Woong Choe. 2010. The Strength of the Relationship between Semantic Similarity and the Subcategorization Frames of the English Verbs: a Stochastic Test based on the ICE-GB and WordNet. *Language and Information 14.1*, 113–143. The primary goal of this paper is to find a feasible way to answer the question: Does the similarity in meaning between verbs relate to the similarity in their subcategorization? In order to answer this question in a rather concrete way on the basis of a large set of English verbs, this study made use of various language resources, tools, and statistical methodologies. We first compiled a list of 678 verbs that were selected from the most and second most frequent word lists from the Colins Cobuild English Dictionary, which also appeared in WordNet 3.0. We calculated similarity measures between all the pairs of the words based on the 'jcn' algorithm (Jiang and Conrath, 1997) implemented in the WordNet::Similarity module (Pedersen, Patwardhan, and Michelizzi, 2004). The clustering process followed, first building similarity matrices out of the similarity measure values, next drawing dendrograms on the basis of the matricies, then finally getting 177 meaningful clusters (covering 437 verbs) that passed a certain level set by *z-score*. The subcategorization frames and their frequency values were taken from the ICE-GB. In order to calculate the Selectional Preference Strength (SPS) of the relationship between a verb and its subcategorizations, we relied on the Kullback-Leibler Divergence model (Resnik, 1996). The SPS values of the verbs in the same cluster were compared with each other, which served to

* 주저자. Dept. of Linguistics, University of Washington, Box 354340 Seattle, WA 98195-4340, USA. E-mail: sanghoun@u.washington.edu

† 교신저자. 서울특별시 성북구 안암동 5가 고려대학교 언어학과 136-701. E-mail: jchoe@korea.ac.kr

감사의 글: 본 연구는 'The Relationship between Semantic Similarity and Subcategorization Frames in English: A Stochastic Test Using ICE-GB and WordNet'이라는 제목으로 The 22nd Pacific Asian Conference on Language, Information, and Computation (PACLIC 22, 2008년 11월, Cebu, The Philippines)에서 발표된 것을 발전시킨 것으로, 전 과정을 재 분석하였고, 핵심 자원과 알고리즘을 새롭게 선택하였으며, 특히 3.1절, 3.4절, 5절 등을 새롭게 작성하였다. 좋은 지적을 해주신 심사위원들께 깊은 감사를 드린다. 이 논문은 2008년 정부(교육인적자원부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (NRF-2008-327-A00315)

give the statistical values that indicate how much the SPS values overlap between the subcategorization frames of the verbs. Our final analysis shows that the degree of overlap, or the relationship between semantic similarity and the subcategorization frames of the verbs in English, is equally spread out from the 'very strongly related' to the 'very weakly related'. Some semantically similar verbs share a lot in terms of their subcategorization frames, and some others indicate an average degree of strength in the relationship, while the others, though still semantically similar, tend to share little in their subcategorization frames. (**University of Washington, Korea University**)

Key words: semantic similarity, subcategorization frames, ICE-GB, WordNet, statistical method, clustering, dendrogram, selectional preference strength

1. 서론

어떤 언어에서든 술어별로 고유한 의미가 있고, 또한 하위범주화라는 통사적 특성을 지닌다는 점은 잘 알려져 있다. 또한 술어별 하위범주화가 그 술어의 의미적 특성을 결정하는데 주요한 역할을 할 것이라는 점도 일반적으로 받아들여지고 있다. 예를 들어 어떤 술어가 '자동사'나 '타동사'로 분류되는 순간, 그 술어가 취할 수 있는 의미가 상당부분 제약이 된다. 그렇다면 술어별 하위범주화 (또는 포괄적으로 '논항구조')와 그 술어의 의미 사이에는 어느 정도의 연관성이 있는 것일까? 그러한 논제에 대하여 대규모의 자료를 통한 검증을 할 수 있는 것일까? 이 연구는 영어를 소재로 하여 바로 그러한 질문에 대한 한 가지 답을 찾는 과정이 될 것이다.

위의 질문에 대한 답을 하기 위해서는 몇 가지 전제가 충족되어야 한다. 첫째로 영어의 주요 술어 선정에 대한 기준이 필요하고 그러한 술어의 명단을 모두 확보해야 한다. 둘째로 그 모든 술어의 개별 논항구조에 대한 정보도 파악이 되어 있어야 한다. 셋째로, 그 모든 술어의 의미적 특성이 알려져 있어야 하고, 그러한 의미적 특성에 따라 선택된 술어들 사이의 의미적 유사성을 파악할 수 있어야 한다.

본 연구는 최대한 객관적인 방법론을 도입하여 위의 전제가 되는 정보를 파악하고, 그러한 정보들을 바탕으로 술어의 논항구조와 의미 사이의 연관성 정도를 검증하는 것을 목표로 한다. 본 연구에서 기본 데이터로 삼는 언어자원은 크게 세 가지이다. 첫째는 Collins Cobuild English Dictionary(이하 COBUILD)이다. COBUILD는 1억 어절 이상의 코퍼스 기반하여 구축된 사전으로서 그만큼 실제 영어의 쓰임을 반영하며 빈도에 따라 영어 단어를 구분하고 있다. 본 연구에서 대상으로 삼는 영어 동사성 어휘는 COBUILD에서 빈도를 기준으로 하여 추출한다. 두 번째는 전 세계의 다양한 영어를 비교 연구하기 위한 목적의 일환으로 구축된 Internatinal Corpus of

English-Great Britain (ICE-GB, (Nelson, Wallis, and Aarts, 2002)) 로서[1] 이는 약 백만 어휘로 구성된 구문분석 코퍼스(트리뱅크) 이다. 본 연구에서 이 자료는 영어 동사의 통사적 특성, 구체적으로는 하위범주화 정보를 추출하기 위한 목적으로 이용된다. 세 번째는 Princeton Univ.의 인지 과학 실험실에서 개발하여 대표적인 의미관련 자료로 활용되고 있는 WordNet이다. 이는 어휘 참조 체계로서 인간의 어휘 개념에 대해 심리언어학적인 이론을 적용하여 영어 단어의 체계와 위상을 정리한 것으로서 본 연구에서 어휘 사이의 의미적 유사성을 결정하기 위한 용도로 활용된다.

이러한 자료의 추출 및 추출 결과를 바탕으로 하여 연관성을 검증하고, 그러한 결과를 바탕으로 연관성의 분포적 특성을 파악하는 것이 본 연구의 주요 목적이 된다. 본 연구에서는 언어현상을 객관적으로 분류하고 비교하는 모형을 제시할 것이며 결과를 분석하는 통계적 방법론도 제시할 것이다. 언어를 분석하는 데 컴퓨터의 이점을 최대한 살리고 수학적, 통계적, 확률적 모형으로 이를 계량화한다는 점이 본 연구의 주요 특성이다.

본고의 전체 구성은 다음과 같다. 2절은 본고의 주된 방법론 두 가지를 나누어 소개하며, 연구를 위한 자료 선정에 대하여 논의한다. 3절은 WordNet을 기반으로 의미적 군집화를 구성하는 전체 절차와 주안점에 대해 논의한다. 4절에서는 ICE-GB를 토대로 하위 범주화를 추출하고 이를 통해 동사와 하위범주화 사이의 선택 선호도를 계산한다. 5절에서는 3절의 결과와 4절의 결과 사이의 관계성 포착을 시도하며, 그 의의에 대해 검토한다. 6절은 본고의 결론이다.

2. 연구 배경, 방법론 및 자료

Levin (1993)에 따르면 동작(motion), 접촉(contact), 영향(effect) 등을 어휘 속성에 내재하고 있는 'cut'류의 동사는 사역 교체가 잘 이루어지지 않는다.

(1) a. Carol cut the bread.

b. *The bread cut.

이러한 특성은 같은 부류의 다른 동사들도 공유하고 있는 특성으로 예를 들어 *chip*, *scrape*, *snip* 등 *cut*과 동의어라 할 수 있는 동사들(Synonym Collection v1.1, Copyright © 2008 by Lexico Publishing Group, LLC.)도 마찬가지 분포적 특성을 보인다. 이처럼 의미적으로 유사한 동사들이 논항구조적인 특성을 공유할 수 있다는 점은 어휘적 특성을 대규모로 연구한 Levin (1993) 등의 연구에서 이미 확인된 바 있다.

[1] ICE-GB는 영국영어를 소재로 구축한 100만 어휘 규모의 균형말뭉치로, 구어 대 문어의 비율이 6:4 정도다. 자세한 내용은 ICE-GB 홈페이지(http://www.ucl.ac.uk/english-usage/projects/ice-gb/) 참조

그렇다고 의미적으로 유사한 계열의 동사들이 반드시 하위범주적 특성도 공유하는 것은 아니라는 점, 즉 이른 바 통사적 특성이 의미적 특성으로 귀결될 수 없다는 주장 역시 언어학에서 새로운 것은 아니다. 같은 'question'계열의 동사인 *ask*와 *wonder*를 보자. 그 둘은 'question'의 동어어라는 점에서 의미적으로 유사하다(Synonym Collection v1.1). 그러나 그렇다고 해서 그 둘의 하위범주적 특성이 같은 것은 아니다 (Lasnik, Uriagereka, and Boeckx, 2005).

(2) a. She **asked** what time it was.

b. She **asked** the time.

c. She **wondered** what time it was.

d. *She **wondered** the time.

위에서 제시한 두 가지 서로 상충하는 듯이 보이는 주장 및 자료를 볼 때 당연히 떠 오르는 의문은 의미와 하위범주화 사이의 연관성은 있는가, 있다면 어느 정도인가, 또한 그러한 연관성은 어떠한 양상을 보이는가 등이 될 것이다. 또한 그에 앞서 과연 그러한 질문에 답을 하기 위해서는 어떤 연구가 필요할 것인가라는 의문도 제기될 것이다.

동사의 의미와 하위범주화 사이의 연관성에 대한 질문은 여러 가지 방식으로 제기될 수 있다. 예를 들어 하위범주화가 유사하면 의미도 유사한가라는 측면에서 접근할 수도 있고, 반대로 의미가 유사하면 그 유사한 동사들끼리 비슷한 하위범주화를 취하는가라는 질문을 던질 수도 있을 것이다. 아니면 모든 동사의 의미와 하위범주화를 동시에 고려하면서 그 둘 사이의 연관성을 찾아가는 방식도 가능은 할 것이다. 본고에서는, 적절한 도구의 유무라는 현실적 제약과, 또한 효율성의 차원에서, 두 번째 방식을 선택하였다. 즉, 동사들끼리 상호 의미적 유사도를 측정하여 일정 기준 이상의 유사도를 보이는 것들끼리 소규모로 묶은 유의어 명단(clusters)을 찾고 나서, 동일 묶음에 속하는 동사들의 하위범주화 정보를 모두 찾아 해당 동사와의 연관성을 수치화한 뒤에, 동사와 하위범주 사이의 밀접도가 어떤 분포를 보이는가를 기준으로 연관성 정도, 즉 의미와 하위범주화 사이의 일치도를 평가하였다.

동사와 논항 사이의 연관성이나 밀접도를 평가하는 한가지 방법이 Resnik (1996)에 의해 제안된 바 있다. '선택 선호도'(Selectional Preference Strength)로 통용되는 이 개념은 대규모 자료를 바탕으로 수치화를 가능하게 해 준다는 점에서 특히 유용하다. 다음 예를 보자.

(3) a. Experimenter: *Could a cow be green?*

b. Subject: *I think they're usually brown or white.*

(4) a. Experimenter: *Could an idea be green?*

b. Subject: *No, silly! They're only in your head.*

(3)의 *a green cow*는 비록 실제 세계에는 존재하지 않는다 하더라도 비교적 쉽게 상상해 볼 수 있는 대상이다. 반면 (4)의 *a green idea*는 그렇지 못하다. 예컨대, *a green cow*를 그림으로 표현하는 것은 어렵지 않지만, *a green idea* 경우에는 이를 어떠한 방식으로 개념화해야 할지 혼란스럽다. 이는 구체명사 *cow*와 추상명사 *idea* 사이의 차이를 넘어서는 문제로, *cow*의 속성으로 *color*를 취하는 것은 자연스러우나, *idea*와 *color*를 조합한다는 것은 혼란스러운 과정임에 틀림없다. 즉, (4a)는 인간의 인지능력에 비추어 볼 때 가능한 결합임에 비해, (4b)는 결합이 불가능해 보이는 어휘들의 연쇄라는 점에서 서로 다르다. (3)과 (4)의 차이는, 주어진 두 단어 또는 두 어휘 군이 서로 얼마나 연관성을 지닐 수 있는가의 차이와 관련된다. Resnik (1996)은 이러한 제약관계에 주안점을 두고 어떠한 술어가 어떠한 어휘군을 논항으로 유의미하게 취할 수 있는가에 대한 형식적인 모형을 제시하였다. 즉, 대규모 자료를 바탕으로 동사별 논항구조 선택 선호도를 수치화 할 수 있다는 점을 보인 바 있다. 본 연구에서도 동사의 논항구조, 구체적으로는 개별 동사의 하위범주화의 특성을 이와 같은 통계기반의 형식적인 모형에 바탕을 두어 추출한다.

2.1 방법론

본고에서 제안된 가설을 검증하기 위한 방법론은 크게 두 가지로 설명할 수 있다. 첫째, WordNet에서 의미적 유사성을 기준으로 두 어휘 사이의 거리를 측정하는 WordNet::Similarity 모듈에 기반하여 (Pedersen, Patwardhan, and Michelizzi, 2004), 연구의 대상이 되는 동사를 군집화하는 알고리즘을 제시할 것이다. 둘째, 동사의 하위범주를 기준으로 하여 그 선택 선호도를 측정하는 방법을 도입하여 그 결과를 의미적 유사도를 기준으로 군집화된 대상과 비교할 것이다. 각각을 구체적으로 살펴보면 아래와 같다.

2.1.1 군집화 알고리즘: WordNet의 활용. 첫 번째는, 의미적 유사성에 따른 동사의 군집화이다. 일관된 기준으로 비교적 효율적으로 결과를 도출해 내기 위해 본고에서는 WordNet (*ver.* 3.0)[2] 을 사용하였다. 이처럼 본고는 의미적 유사도를 판별하기 위한 자원은 WordNet으로 설정하고, 그 유사도를 구체적으로 측정하기 위한 환경으로는

[2] WordNet의 활용과 관련된 장점 중 하나는, WordNet에 망라된 개념들 사이의 의미적 유사도를 측정할 수 있는 다양한 알고리즘이 이미 개발되어 있고, 이를 패키지화하여 제공하는 소프트웨어 역시 공개적으로 이용이 가능하다는 점이다. 현재 영어 WordNet의 최신 버전은 3.0으로서 이는 리눅스 또는 유닉스 환경에서의 동작만을 지원한다. 윈도우 환경에서 WordNet을 구동하려면 2.1 버전을 사용하거나, 혹은 윈도우 상에 가상 리눅스 환경을 구축하고 이 환경 위에 3.0 버전을 구축해야 한다. 본고에서는 가급적 최신의 정제된 자료를 활용하기 위하여, 리눅스 환경에 설치된 3.0 버전을 이용하였다.

WordNet::Similarity를 이용한다.[3] 군집화는 이 결과를 바탕으로 이루어 지며, 구체적으로는 Manning and Schütze (2002) 등에서 제시된 계층-상향식 군집화 알고리즘을 통해 추출된 계층도(dendrogram)를 활용할 것이다.[4]

2.1.2 선택 선호도 측정: ICE-GB의 활용. 두 번째는 하위 범주화 틀의 선택 선호도를 측정하는 것이다. 선택 선호도란 어떠한 어휘와 관련된 항목 사이의 관계의 정도를 말하는 데, 경우에 따라 이는 이론 언어학에서의 선택 제약과 유사한 개념이다. 두 용어 사이의 차이로, '제약'의 경우 참 또는 거짓이라는 다소 엄격한 제한을 요구하는 반면, '선호도'는 'excellent', 'fine', 'bad', 'almost impossible' 등과 같은 표현으로 환원될 수 있을 만큼 '정도의 문제'로 귀결된다. 즉, 자료 주도 접근법의 맥락에서는 참 또는 거짓의 이진 값으로만 귀결되는 다소 엄격한 판별보다는, 양자 사이의 통계적 계량치가 중요하게 작용한다. 자료 주도 접근법을 지향하는 본고에서는 이후 '선택 선호도'라는 용어를 사용할 것이다. 즉, 본고에서는 어떠한 동사가 어떠한 하위 범주화 틀을 취하는가에 대한 계량 빈도를 기준으로 논의를 진행할 것이다. 이때 선택 선호도를 도출하기 위한 대상이 되는 일반화 코퍼스가 요구되는데, 통상 원시 코퍼스 또는 주석 코퍼스가 그 대상이 된다. 원시 코퍼스를 활용한 연구로서는 Manning (1993) 이나 Erk (2007)을 예로 들 수 있다. 전자가 완전한 원시 코퍼스만을 사용한 반면, 후자는 BNC를 대상으로 하되, 영어 구문분석기 Minipar (Lin, 1993)를 통해 1차 분석된 결과를 바탕으로 선택 선호도 연산을 하였다. 그러나 일반적으로는 원시 말뭉치를 사용하는 것보다 비록 규모는 작더라도 정밀하게 주석이 된 코퍼스를 사용하는 것이 보다 성능이 좋은 것으로 알려져 있으며, 특히 심층 분석된 코퍼스인 트리뱅크를 사용하는 것이 더 바람직하다. 이에 따라 본고에서는 영어 트리뱅크의 일종인 ICE-GB를 그 일반화 코퍼스로 상정한다.

2.2 기본 자료

서론에서 언급하였듯이, 본고에서 사용하고자 하는 영어 동사의 목록은 COBUILD를 기반으로 한다. COBUILD는 몇 억 어절 이상 규모의 실제 영어 코퍼스를 바탕으로 구축된 사전인 바, 실제 각 어휘의 빈도 정보를 포함하고 있다. 사전상에서 이러한 빈도 정보는 ★표지로 표시되는데, 이 표시의 개수가 많을수록 고빈도 어휘에 속한다. ★표

[3] Pedersen, Patwardhan, and Michelizzi (2004)에 의해 개발되어 공개된 이른바 WordNet::Similarity 모듈은 공개 소프트웨어 정신에 입각하여 현재 인터넷상에서 누구나 내려받을 수 있다 (Pedersen, 2008). WordNet::Similarity는 Perl로 구현되어 있으며, 자신의 컴퓨터에 설치된 WordNet의 버전에 따라 선택적인 설치가 가능하다. 본고에서는 WordNet *ver.* 3.0에 맞는 모듈을 활용하였다.

[4] 실제 군집화를 하는 방법에 관련하여서는 여러 알고리즘이 연구 및 개발되어 왔는데, 이들을 간략히 소개하면 다음과 같다. 군집의 구성방식은 크게 계층형 군집화 알고리즘과 비-계층형 군집화 알고리즘으로 구분된다. 계층형 알고리즘은 다시 그 방향에 따라 상향식(bottom-up) 알고리즘과 하향식(top-down) 알고리즘으로 양분된다 (Manning and Schütze, 2002). 반면 비-계층형 알고리즘의 대표적인 것으로는 K-means 알고리즘을 들 수 있는데, 이는 공간 상에서 각 분포의 중심(center)을 반복적으로 추정하는 산술과정에 의해 구축된다. 상세한 군집화 알고리즘에 대해서는 Kaufman and Rousseeuw (1990) 또는 Manning and Schütze (2002)에 제시되어 있으므로 참고 바란다.

지는 0개에서 많은 경우 다섯 개까지 표시가 되며, COBUILD의 설명에 따르면 실제 영어 자료에서 ★표지 4개 이상의 어휘가 전체 사용된 어휘의 약 75%를 점유한다고 한다. 현재, COBUILD에서는 최고빈도 (★ 5개 부착) 동사가 655개, 두 번째 고빈도 (★ 4개 부착) 동사가 1,026개 망라되어 있다. 다시 이들 어휘 총 1,681개 가운데, WordNet (*ver.* 3.0)에서 동사로 처리된 것만을 추출하면 총 678개의 동사 목록을 얻을 수 있다. 본고에서는 이 목록을 연구의 출발점으로 한다.

이 목록은 이후 두 개의 언어 자원에서 비교의 대상이 된다. 하나는 의미적 유사도를 측정하기 위한 대상으로 앞서 언급한 WordNet이다. 다른 하나는 동사의 하위 범주화 틀을 추출하기 위한 것으로 ICE-GB가 그 대상이다. 한편, 이들 각각을 처리하기 위한 도구 또한 요구되는데, 각기 WordNet::Similarity와 ICECUP 3.0을 활용한다. 여기에 추가하여 구축된 의미적 군집화의 적합성을 판별하기 위해, 비교 목적의 자원 역시 요구된다. 본고에서 그 교차 검증을 위해 사용하는 언어 자원은 시소러스로서, COBUILD 시소러스와 Roget 시소러스가 활용된다. 이상의 내용을 정리하면 [표 1]과 같다.

[표 1] 대상 언어 자원

목적	언어 자원	검색 환경	비고
어휘 집합	COBUILD		상위 고빈도 어휘 1,681개
동사 목록	COBUILD & WordNet		동사 가능 어휘 678개
의미 유사도 측정	WordNet (*ver.* 3.0)	WordNet::Similarity	
군집화 결과 검증	시소러스		Cobuild, Roget
하위 범주화 틀 추출	ICE-GB	ICECUP	

3. 의미적 유사도와 군집화

동사의 목록이 선정되고 난 후, 연구의 첫 번째 단계는 동사 678개를 군집화는 하는 것이다. 이 단계는 다시 세 단계로 세분화된다. 첫째는 WordNet::Similarity를 이용하여 동사 678개가 서로에 대해 가지는 의미적 유사도를 각 알고리즘의 처리 기준에 따라 빠짐없이 측정하는 것이다. 둘째는 추출된 유사도를 기준점으로 하여 의미적 속성이 가까운 것끼리 순차적 / 계층적으로 묶어 계층도를 그리는 것이다. 셋째는 구성된 계층도에서 통계적으로 유의미한 것들만을 정리하여 이를 군집으로 확정하는 것이다.

3.1 WordNet::Similarity

현재 WordNet::Similarity 모듈은 비교를 위한 'random' 알고리즘을 포함하여 총 10개의 알고리즘을 제공한다. [표 2][5]는 위 678개 동사에 10개 알고리즘을 각각 적용하여

[5] 실제 연구의 단계에서는 동사뿐만 아니라, 같은 원리로 COBUILD에 추출된 명사 (1,074개), 형용사 (422개), 그리고 부사 (189개) 역시 유사도 계산을 통해 각 알고리즘 별로 정리되었다. 동사를 비롯한 이들 전체 계산 결과는 아래의 주소에서 확인할 수 있다 ('공개의 당위성'과 관련한 논의는 Pedersen

[표 2] WordNet::Similarity를 이용한 유사도 측정 결과

알고리즘	결과 개수	최대값	최소값	평균	표준편차
hso	130,774	16	1	3.3169	1.5908
jcn	4,347,772	1	0.0436	0.0611	0.0235
lch	4,374,326	3.3322	0.1542	1.4398	0.3276
lesk	4,221,315	7,547	1	9.8423	19.3758
lin	147,756	1	0.1705	0.3962	0.1309
path	4,374,326	1	0.0417	0.1591	0.0549
random	4,374,326	1	0	0.5	0.2887
res	147,756	10.7783	1.9561	3.3558	0.9738
vector	4,361,816	1	0.0004	0.0832	0.0623
wup	4,374,326	1	0.08	0.2773	0.0895

얻은 결과의 일차적인 통계 수치를 보여준다.[6]

이들 알고리즘을 서로 비교해 볼 때, 주목할 점이 크게 두 가지가 있다. 하나는 각각의 알고리즘마다 척도가 상이하다는 점이다. 예컨대, 'lesk' 알고리즘은 각 수치가 정수이며 그 변화의 폭이 표준편차의 수치에 드러나는 바와 같이 10개의 알고리즘 가운데 가장 크다. 반면, 'jcn' 알고리즘은 척도가 소수 단위이며 그 변화의 폭, 다시 말해 표준편차가 가장 작다. 따라서 이들 수치 각각을 동일 선상에서 비교할 수는 없다. 둘째, WordNet의 각 어휘는 표면형이 아닌 'synset'이라는 단위를 기반으로 연결되어 있는데, 이 'synset'을 처리하는 기준 역시 각 알고리즘마다 차이가 있다. 예컨대, 'path', 'wup' 등의 알고리즘은 각 어휘의 거의 모든 'synset'을 대상으로 비교가 가능한 반면, 'lin', 'res' 등의 알고리즘에서는 유의미한 수치만을 대상으로 한다. [표 2]에서 각 알고리즘의 처리 결과 개수에 차이가 있는 것이 이러한 원인에 따른다. 각 알고리즘마다 이처럼 결과의 형태가 다르다는 점은 이후 분석의 과정에서 역시 검토되어야 할 부분이다.

본고에서는 위 알고리즘 가운데, 선행 연구를 통해 성능이 어느 정도 입증된 'lesk'와 'jcn' 두 알고리즘을 대상으로 논의를 계속 진행하기로 한다. Banerjee and Pedersen (2003)에서는 'lesk' 알고리즘을 주요 알고리즘으로 제시하였으며, 송상헌 · 전지은 · 최재웅 (2008)에서는 'lesk' 알고리즘을 기반으로 영어 형용사를 군집화하는 모델을 제시

(2008) 참고).
http://corpus.mireene.com/download/wn-sim.html

[6] 실제의 결과에서는 몇 군데의 이상 수치가 발견되었다. 예컨대, 'jcn' 알고리즘의 결과 가운데 일부는 그 유사도 값이 백만 단위로 나타나는 경우가 몇 회 출현하였다. 이는 Perl에서 연산하는 과정에서 나타나는 문제점으로 엄밀히 말해 실수 처리를 하는 수학식 모듈의 결함이다. 이러한 예가 발견되는 경우에는 자동으로 그 값을 최대값으로 치환하는 방식으로 결과를 보정하였다.

하였다. 한편, Budanitsky and Hirst (2006)에서는 5개의 알고리즘을 비교하여 'jcn' 알고리즘이 가장 우월한 성능을 보임을 입증하였다. Jurafsky and Martin (2009)에서도 WordNet을 기반으로 하여 어휘 사이의 유사도를 측정하는 여러 모델을 소개하고 있는데, 결론적으로 'lesk'와 'jcn' 두 알고리즘의 우수성을 경험적으로 입증하고 있다.

Lesk (1986)에서 제시된 'lesk' 알고리즘은 기본적으로 각 개념의 풀이말을 이용한다. 즉, 개념 A와 개념 B의 풀이말에서 서로 중첩되는 부분이 얼마나 되는가를 상계한다. 이에 더하여, A와 B에 직접 연계된 하의어와 상위어 등도 유사도 측정에 활용된다. 기본적 모형은 아래의 수식에 드러나 있으며, 여기에서 *hype*와 *hypo*는 각각 상위어와 하의어를 의미한다.

$$\begin{aligned} relatedness(A,B) = {} & score(gloss(A), gloss(B)) \\ & + score(hype(A), hype(B)) + score(hypo(A), hypo(B)) \\ & + score(hype(A), gloss(B)) + score(gloss(A), hype(B)) \end{aligned} \quad (1)$$

Lesk (1986)와 Jurafsky and Martin (2009)에서 제시된 예는 아래와 같다. 두 단어 'pine', 'cone'이 아래와 같은 풀이말을 지닌다고 할 때, cone[3]은 'evergreen' 및 'tree'라는 단어에 의해 pine[1]과 공통분모를 가져 상호 정보량을 지니게 된다.

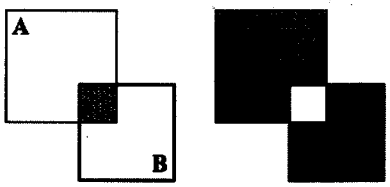
(5) a. pine[1] kinds of **evergreen tree** with needle-shaped leaves
pine[2] waste away through sorrow or illness

b. cone[1] solid body which narrows to a point
cone[2] something of this shape whether solid or hollow
cone[3] fruit of certain **evergreen tree**s

한편, 'jcn'은 Jiang and Conrath (1997)에 근간을 두고 있는데, 이는 기본적으로 Resnik (1995)에 제안된 알고리즘의 단점을 보완한 모델이라고 할 수 있다. Resnik에 의해 제시된 가장 기본적인 모형은 'IS-A' 계층에서 두 요소 C_1과 C_2 사이의 거리는 이 둘을 모두 포함하는 가장 낮은 상위 절점에 의해 결정된다는 것으로, 이때 그 절점을 LCS(Lowest Common Subsumer)라 정의한다. LCS의 기본적 모형은 이후 여러 알고리즘에서 발전적으로 활용되었는데, 그 가운데 가장 우수한 성능을 보이는 것으로 보고된 것이 아래 수식으로 연산되는 'jcn'이다 (Jurafsky and Martin, 2009).

$$dist_{JC}(C_1, C_2) = 2logP(LCS(C_1, C_2)) - (logP(C_1) + logP(C_2)) \quad (2)$$

다소 복잡하게 보이는 위 수식은 실제 개념적으로 보면 상당히 단순한 원리에 입각하고 있는데, [그림 1]의 면적을 통해 설명하기로 한다. 개념 A와 B가 [그림 1] 같은



[그림 1] Lowest Common Subsumer

분포를 보이고 있다고 할 때, A와 B가 서로 공통으로 취하는 분포는 왼쪽 그림의 가운데 교집합에 해당할 것이다. 이때 두 개념의 차이는 각각에서 둘의 교집합 부분을 제외한 여집합인 오른쪽 그림에 해당한다. 결과적으로 오른쪽 그림에서 음영으로 표시된 부분의 면적은 위 수식을 통해 나오는 결과와 같아진다. 예컨대, A의 면적이 30, B의 면적이 20, 그리고 왼쪽 그림 가운데 교집합 부분의 면적이 10이라면, A와 B를 합한 영역의 넓이가 40이 될 것이고, 따라서 음영으로 표시된 영역의 넓이는 '(2×40) - (30+20)'에 의해 30으로 결정된다. 즉, 위 수식은 두 분포의 차이를 구하는 개념적 모델과 동일하다고 할 수 있다.

이 두 알고리즘은 여타의 알고리즘에 비해 좋은 성능을 보인다는 공통점에도 불구하고 기본 전제가 다른 탓에 서로 분명한 차이점가 있다. 가장 뚜렷한 차이는 [표 2]에 드러난 바와 같이 표준편차의 차이이다. 즉, 'lesk'는 편차의 값이 가장 큰 반면, 'jcn'은 그 값이 가장 작아 둘은 서로 극과 극을 이루고 있다. 이 차이가 뒤에서 살펴볼 군집화 결과에도 크고 작은 영향을 주고 있다

3.2 계층도 (Dendrogram)

다음 단계는 위에서 언급된 각 유사도를 기반으로 하여 아래와 같은 유사도 행렬을 구성하는 것이다. 이때, `sim(A,B)`와 `sim(B,A)`는 재귀적 관계로서 동일한 값을 지닌다.

[표 3] 유사도 행렬

	v_2	v_3	v_4	...	v_n
v_1	$sim(v_1,v_2)$	$sim(v_1,v_3)$	$sim(v_1,v_4)$	...	$sim(v_1,v_n)$
v_2		$sim(v_2,v_3)$	$sim(v_2,v_4)$	...	$sim(v_2,v_n)$
v_3			$sim(v_3,v_4)$	...	$sim(v_3,v_n)$
...				...	$sim(...,v_n)$
v_{n-1}					$sim(v_{n-1},v_n)$

본고에서 사용하고자 하는 계층도는 제시된 유사도 행렬 3의 순차적인 조작으로 구성된다. 본고의 계층도는 이른바 파스-트리 구조의 일종으로서 각각의 절점은 부모

절점, 좌측 딸 절점, 우측 딸 절점의 세 요소로 구성된다. 이때 각 절점은 연산자와 피연산자의 관계에 의해 서로 연결되며, 하나의 절점은 다른 절점의 연산자인 동시에 또 다른 절점의 피연산자가 될 수 있다. 구체적인 예로서 집합 S는 관계 ○에 대하여 (6)과 같은 상호 정보량을 지닌다고 가정하자.

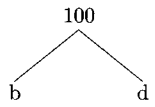
(6) S = { a○b=7, a○c=18, a○d=15, b○c=12, b○d=100, c○d=3 }

이 집합 S에서 계층도를 도출해 내는 과정은 아래와 같다.

[표 4] 계층도 추출 과정

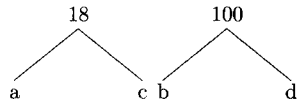
STATE 1

	a	b	c	d
a		7	18	15
b			12	**100**
c				3
d				



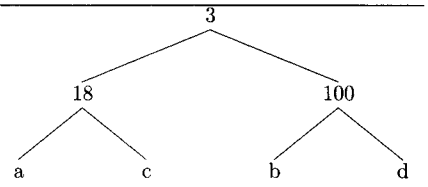
STATE 2

	b, d	a	c
b, d		7	3
a			**18**
c			



STATE 3

	a, c	b, d
a, c		**3**
b, d		



STATE 1에 나타난 행렬은 초기값이다. 이 행렬에서는 최대값 100이 선택된다. 이 값은 b와 d 관계에서 나타난 것이므로 b와 d가 최소의 파스-트리를 형성한다. 다음 STATE 2에서는 이전 STATE에서 결합을 이룬 'b, d'가 행렬의 맨 처음에 자리를 하고 나머지 요소가 차례로 뒤따른다. 각 값은 각각의 요소들의 관계에서 최소값을 택한다.[7] 즉, 처음 칸은 'a○b=7', 'a○d=15'의 관계에서 7이 선택된다. 모든 칸에 각 관계들의 최소값이 채워지고 나면, 전체에서 최대값 18이 선택된다. 이는 a와 c의 관계이기 때문에 파스-트리에 a와 c의 연결 절점이 형성된다. STATE 3에서는 'a, c'가 처음 위치에 자리를 잡게 되고 'a, c', 'b, d'의 관계에서 최소값 3이 결정된다. 이들 사이에 연결이 이루어지면 전체 요소에 대한 파스-트리가 완성된다. 이후, 더 이상 처리할

[7] 다른 방식으로 최대값을 택하는 경우와 평균값을 택하는 경우, 모두 실험을 해보았으나 최소값을 선택하는 것이 세 방식 가운데 가장 결과가 좋다는 점을 경험적으로 확인하였다. 이는 각 어휘가 여러 의미를 가질 수 있다는 점에 기인한다. 예컨대, 'tell'의 경우 통상 '말하다'의 의미를 가지지만, 자동사로 사용될 경우 '영향을 주다'와 같은 의미를 취할 수도 있다. 이처럼 의미가 분화되는 경우, 최대값 또는 평균값을 취해 그 의미적 상관성을 넓게 보면 자칫 문제가 발생할 수 있다. 'tell'이 'say, 'talk' 등의 동사보다 그 의미적 연관 빈도가 훨씬 적은 'impact' 등과 같은 어휘와 더 먼저 군집을 형성할 수 있기 때문이다.

행렬이 남아 있지 않기 때문에 알고리즘은 여기에서 종료된다. 전체 알고리즘 구조는 다음과 같다.[8]

```
1: V_i = {v_i}
2: M = {v_1, v_2, ..., v_{n-1}, v_n}
3: while n > 1:
4:         V_1 = V_{pos_i} ∪ V_{pos_j}
5:         for i = 1 to n-1:
6:             for j = i+1 to n:
7:                 sim(V_i,V_j) = min(V_i,V_j)
8:                 if max < sim(V_i,V_j):
9:                     max = sim(V_i,V_j)
10:                    pos_i = i
11:                    pos_j = j
12:        parse-tree(pos_i, pos_j)
13:        n = n - 1
```

이러한 알고리즘에서는 [표 2]에서 제시된 각각의 결과를 입력 자료로 활용하는 것이기 때문에 서로 척도가 상이하다는 점이 결과에 별다른 영향을 주지 않는다. 각 알고리즘의 내부 수치만을 대상으로 하여, 그 값만을 비교의 대상으로 삼기 때문이다.

그러나 다른 측면으로는 최소한 두 가지 사항에 대한 고려가 필요하다. 하나는 앞서 언급한 바와 같이 synset에 대한 고려이다. 예컨대, WordNet *ver.* 3.0에서 동사 *accept*는 7개의 synset을 가지며, 동사 *admit*는 4개의 synset을 가지는 관계로 이 두 개의 동사는 실질적으로 총 28개의 관계값을 가진다. 이 경우 각 표면형을 기준으로 한다면 추출된 28개의 관계값의 집합에서 최대값, 최소값, 또는 평균값을 순차적으로 택하여 보고 가장 성능이 좋은 값을 택하는 방법이 요구된다. 반면, 각 synset을 그대로 인정할 경우, 위 28개의 대응쌍은 각기 독립적인 결과로서 인정되어야 한다.

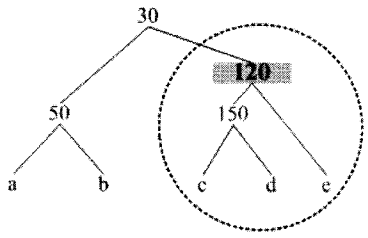
두 번째는 [표 2]에서 제시된 바와 같이 10개의 알고리즘이 동일한 차수의 유사도 집합을 만들어 내는 것이 아니라는 점이다. 이 경우 표면형만을 택하는 경우에는 위와 동일한 방식으로 처리되는 것이 바람직할 것이나, synset을 기준으로 하는 경우에는 차수의 차이는 분포의 차이를 낳는 것이기 때문에 문제가 될 수 있다. 동사의 synset 개수를 모두 확인하여 비어있는 관계값을 0으로 설정하는 방법도 고려될 수 있으나 선행 실험을 통해 이것이 결과를 왜곡할 수 있다는 점이 확인되었다.

[8] 여기에서 제안한 방식은 송상헌 · 전지은 · 최재웅 (2008)에서 이미 제시된 바 있다. 그러나 그 방식과는 몇 가지 측면에서 차이를 보이는 데, 하나는 synset에 대한 고려 사항이 추가되었다는 점이다. 반면 송상헌 · 전지은 · 최재웅 (2008)은 synset 처리에 독립적일 수 있는 'lesk' 알고리즘만을 사용하였다. 다른 하나는 각 WordNet::Similarity 알고리즘마다 결과치가 달랐다는 점에 대한 고려가 추가되었으며, 또한 송상헌 · 전지은 · 최재웅 (2008)에서는 WordNet *ver.* 2.0을 기준으로 하였다는 점에서 자료상으로도 차이가 있다.

이러한 차원을 종합하여 위 알고리즘의 대상이 되는 집합 M의 선정은 다음의 방식을 따르기로 한다. 첫째 각 synset을 독립적으로 인정하는 방식은 본 연구에서 택하지 않기로 한다. 우선 결과의 왜곡 없이 일관되게 비교 수치를 도출하는 구현이 요원할뿐더러, 더 결정적으로 일반화 코퍼스인 ICE-GB 결과가 표면형을 중심으로 구성되어 있다는 이유 때문이다. 즉, ICE-GB에서 추출된 각 동사의 논항 관계가 어떠한 synset과 연계된 것인지를 객관적인 방법으로 검증할 수 없기 때문이다. 두 번째로 두 동사의 synset 정보에 따른 복수의 관계값에서 어떠한 값을 두 동사의 대표값으로 선택할 것인가의 문제에 있어서는 최대값, 최소값, 그리고 평균값을 기준으로 하여 각각 결과를 추출한 뒤, 이 가운데 가장 성능이 좋은 것을 채택할 것이다. 그에 대한 검증을 하기 위해서는 비교의 대상이 있어야 한다. 본고에서는 시소러스를 그 비교 목적으로 활용할 것이다. 이를 통해 결과의 왜곡을 최소화하며, WordNet::Similarity 알고리즘 각각이 지니는 상이함이 비교에 영향을 줄 가능성 역시 최소화하였다.

3.3 군집화

계층도가 도출되고 나면, 그 다음 단계는 군집화이다. 즉, 그 계층도에 속한 절점들 가운데 통계적으로 유의미한 것들만을 추출하여야 하여야 한다. 통계적 유의미성을 판별하는 기준으로는 2.1.1 절에서 소개된 바와 같이 여러 방법론이 존재하는데, 여기서는 임계치를 이용하여 계층도에서 어떠한 절점의 값이 그 임계치를 넘으면 그 절점 이하의 요소를 군집으로 판별하는 기법을 사용하고자 한다 (송상헌 · 전지은 · 최재웅, 2008). 예컨대 [그림 2]에서 임계치를 100으로 설정하면 상위 단계의 절점에서부터 비교하여 해당 절점의 값이 100이 넘는 120이 선택될 것이다. 그러면 그 이하 절점에 속하는 요소인 c, d, e가 유의미한 군집으로 정리된다.

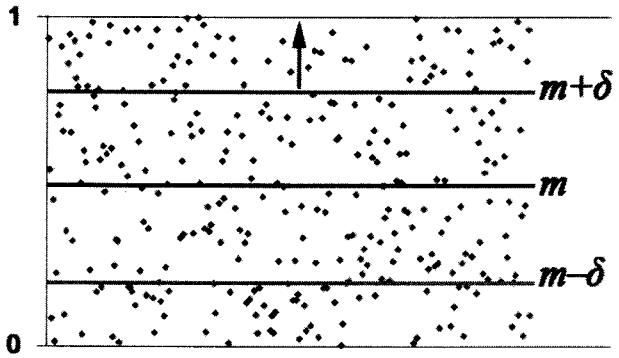


[그림 2] 임계치를 통한 군집화 결정

본고에서는 그 임계치로서 아래에서 제시된 *z-score*를 사용할 것이다. 어떠한 절점의 값을 x라 하자. 또한 전체 절점 값의 평균을 m, 표준편차를 δ라 하자. 표준편차는 각 값이 평균에서 멀어진 정도의 평균을 말하므로 값 x가 평균 m에서 표준편차 δ를 합한 값보다 크다면 그 x는 유의미한 수치라고 추정할 수 있다. 이는 다시 말해 다음 수식을 참으로 하는 x의 값을 구하는 일과 같다.

$$\frac{x-m}{\delta} > 1 \tag{3}$$

구체적인 예를 들어 *z-score*를 이용하는 방법에 대해 논의하면 다음과 같다. n개의 요소들이 0에서 1까지의 실수 값을 그 분포로 취한다고 가정하자. 이를 임의의 값들을 요소로 가지는 배열로 만들어 2차원 평면상에 도식화하면 [그림 3]을 얻을 수 있다.



[그림 3] 분포표 (n = 300)

[그림 3]에서 m으로 표시된 실선은 전체 분포의 평균에 해당하며, 그 상단과 하단의 실선은 그 평균값에서 표준편차를 더하고 뺀 값을 나타낸다. 위 수식을 참으로 하는 x값의 집합은 평균 m과 표준편차 δ를 합한 값보다 큰 값의 집합과 동치이기 때문에, 이는 [그림 3]에서 화살표로 표시된 지역의 점의 분포와 같다. 즉, 평균에서 표준편차보다 먼 거리에 있는 점의 분포는 전체 분포 가운데 유의미하게 상단에 놓여 있다고 가정할 수 있는 것이다.

이와 같은 결정과정을 통해, 678개의 동사 가운데 유의미성을 지니는 각 값들을 검출하였으며, 이들은 다음 절에서 소개할 검증의 대상이 된다.[9]

[9] 이와 관련하여 두 분의 심사자께서는 WordNet의 동계어 및 유의어 추출을 이용하여 군집을 구성하면 마찬가지 방식의 연구를 수행할 수 있음을 지적하였다. 실제로 WordNet은 각 synset과 동일 계층(sister)에 있는 다른 synset들의 집합 및 유의 관계(synonym)에 있는 집합을 검색하는 명령을 지원한다. 이들 각각은 WordNet이 설치된 환경의 프롬프트상에서 아래와 같은 명령행으로 추출 가능하다.

`$> wn verb_entry -s -coorv >output_filename`

`$> wn verb_entry -s -simsv >output_filename`

이러한 방법 역시 일정 정도의 성과를 이룰 수 있음은 분명하다. 그러나 본고에서는 두 가지 이유에서 이러한 방식을 택하지 않았다.

첫째는 WordNet::Similarity에서 제시되는 각 값은 엄밀히 말해 유의도만을 뜻하지 않는다는 점이다. 대표적인 예가 명사 'space'와 'star'의 관계이다. 이들 어휘는 상호 유의어도 아니며, 동계어도 아니지만, 개념구조 상에서는 서로 밀접한 관계를 가진다. 즉, WordNet::Similarity의 각 값은 엄격한 의미에서 연관성(relevance)의 정도를 뜻한다. 이는 (5)에서 제시된 'pine'과 'cone'의 관계에서도 마찬가지로 확인될 수 있는 부분이다.

두 번째는 WordNet::Similarity는 각 어휘들 사이의 관계성을 수치화하여 그 정도를 표시함에 비

3.4 군집화 검증

앞서 설명한 바와 같이, 본고에서는 선행 연구를 통해 실효성이 입증된 'lesk', 'jcn' 두 알고리즘을 대상으로 한다. 3.2 절에서 제시된 계층도에서 3.3 절의 *z-score* 를 임계치로 하여 추출된 군집화의 결과는 [표 5]와 같다. 3.2 절에서 기술된 바와 같이, 계층도는 각 synset 의 대응 관계의 최대값, 최소값, 평균값을 기준으로 하여 각각 추출되었기 때문에, 그 결과 총 6 개의 임의 결과가 얻어진다. 한편 여기에서는 비교를 위하여 하나의 결과치가 더 추가되는데, 바로 'random'알고리즘에서 얻은 결과이다. 이는 두 어휘 사이의 유사도에 무작위 값을 배당한 것으로, 다시 말해 해당 어휘 목록에 어떠한 유사도도 없다고 가정할 때 얻어질 수 있는 결과이다. 이때 'random'에 속하는 값은 모두 임의의 수치이기 때문에 이 경우에는 가장 일반적인 수치인 평균값만을 비교로 하여도 상관이 없다. 결과적으로 이후의 군집화 검증은 아래 7 개를 대상으로 한다 ('lesk' 알고리즘 기반 최대, 최소, 평균값 기준 추출 / 'jcn' 알고리즘 기반 최대, 최소, 평균값 기준 추출 / 'random' 추출 결과).

[표 5] 군집화 추출 결과

	lesk		**jcn**		**random**	
처리 기준	군집 개수	동사 개수	군집 개수	동사 개수	군집 개수	동사 개수
최대값	66	158	177	437	n/a	
최소값	28	70	36	97		
평균값	85	203	129	276	193	462

[표 5]에서 눈에 띄는 점은 'jcn'의 결과가 'lesk'의 결과보다 포괄하는 군집과 동사의 개수가 많다는 점이다. 이는 앞 [표 2]에서 본 바와 같이 'jcn' 알고리즘으로 추출한 결과의 표준편차가 매우 작다는 점에 기인한다.

이들 1 차적 결과는 두 시소러스와의 교차 비교를 통해 검증된다. 검증의 방법은 이른바 'F-measure'로서 이는 'precision'과 'recall'의 값의 조합을 통해 얻어진다. 'precision'은 찾는 자료가 '정확하게' 찾아졌는지를 평가하는 지표이며, 'recall'은 찾아낸 자료가 누락된 정보가 없이 '빠짐없이' 구성되었는가를 평가하는 지표이다. 우선 [표 6]을 보도록 하자.

[표 6] 2×2 결정 분포

	대상에 속함	대상에 속하지 않음
추출됨	*tp*	*fp*
추출되지 않음	*fn*	*tn*

교하여, 위 명령행에 의해 추출된 결과는 단순한 목록에 불과하다. 즉, 추가적인 수치연산이 쉽지 않다. 서론에서 언급한 바와 같이 본고는 각 값의 연관성 정도에 근거하여 결과를 도출함을 목표로 하기 때문에, 위 방식은 결론적으로 대상이 되지 못하였다.

예컨대, 추출된 군집화 목록에 어떠한 어휘 A가 포함되었으나, 이 A가 시소러스에 해당하는 유의어 또는 반의어로서 출현하지 않는다면, *fp* (false positive) 값이 늘어나 결과적으로 'precision'의 값이 떨어지게 된다. 반면, 어떤 어휘 B가 시소러스에는 출현하지만, [표 5]의 군집화에서 걸러지지 않는다면, 이때는 *fn* (false negative) 값이 늘어나며, 그 결과 'recall'에 부정적 영향을 준다. 한편 'precision' 및 'recall' 양자를 종합적으로 살필 수 있는 것이 F-measure로서, 이들 각각은 아래와 같은 수식에 의해 계산된다.

$$precision = \frac{tp}{tp + fp} \tag{4}$$

$$recall = \frac{tp}{tp + fn} \tag{5}$$

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \tag{6}$$

이들 수식의 연산 과정을 예를 들어 설명을 하면 다음과 같다. 어떠한 군집 CL_i가 (7a)와 같은 원소로 구성되어 있다고 가정하자. 또한 어떠한 시소러스에서 어휘 a가 (7b)에서 제시된 어휘들을 그 대상으로 취한다고 가정하자.

(7) a. $CL_i = \{a, b, c\}$

b. ***a***: *b*, *d*, *e*

이때, *tp*는 양자가 공히 취하는 값 *b*에 의거, 1이 된다. 한편, *fp*은 CL_i에는 나타나지만, (7b)에 제시되지 않은 값 *c*에 따라 역시 1이 된다. *fn*는 거꾸로 (7b)에 포함되나, CL_i의 원소가 아닌 *d*, *e*에 의해 2가 된다. 각 값을 계산하면, 'precision'은 (1/(1+1))에 따라 0.5, 'recall'은 (1/(1+2))에 따라 약 0.3333이 된다. 끝으로 F-measure는 α 값을 0.5로 취하였을 때 (1/(0.5×(1/0.5) + (1–0.5)×(1/0.3333)))으로 연산되어 0.4가 된다.

아래의 표들은 위 수식들에 의해 연산된 군집화 검증값을 보여준다. 먼저, 각 결과의 'precision'을 살펴보기로 하자. [표 7]에서 각 시소러스와 비교하여 가장 우수한 성능을 보이는 항목은 굵은 글씨로 표시하였다. 이는 최대값이 최소값과 평균값을 기준으로 한 경우보다 상회한다는 점을 보여준다. 그 경우에 있어서는 'lesk' 알고리즘으로부터 얻어진 결과가 10% 정도의 결과를 더 포함하고 있음을 알 수 있다. 아울러, 비교의 대상인 'random'의 값은 1% 미만으로 다른 값들에 비해 매우 낮은 수치를 보인다. 이는 WordNet::Similarity에서 추출된 결과가 실제 시소러스와 비교하여도 상당 부분 합치하고 있다는 점을 입증한다.

[표 7] precision

	알고리즘	Cobuild	Roget	평균
최대값	jcn	42.43%	44.83%	43.63%
	lesk	**51.25%**	**53.82%**	**52.53%**
최소값	jcn	6.25%	3.81%	5.03%
	lesk	12.24%	9.30%	10.77%
평균값	jcn	40.48%	37.15%	38.81%
	lesk	32.52%	31.01%	31.77%
비교	random	0.29%	1.61%	0.95%

다음으로 'recall'의 경우를 보도록 하자. 'recall'에 있어서는 최소값의 경우가 상대적으로 낙후한 반면, 최대값과 평균값은 큰 차이를 보이지 않는다. [표 7]과의 두드러진 차이는 'jcn'이 'recall'에 관해서는 'lesk' 보다 우수하다는 점이다. [표 7]과 마찬가지로 비교의 대상이 되는 'random'은 이 경우에도 매우 낮은 수치를 보인다.

[표 8] recall

	알고리즘	Cobuild	Roget	평균
최대값	jcn	11.25%	**6.41%**	**8.83%**
	lesk	10.08%	5.82%	7.95%
최소값	jcn	4.72%	1.47%	3.09%
	lesk	8.05%	2.39%	5.22%
평균값	jcn	**11.28%**	5.09%	8.19%
	lesk	10.99%	5.06%	8.03%
비교	random	0.13%	0.36%	0.25%

끝으로 위 두 수치를 조합한 F-measure를 보기로 하자. [표 9]는 두 시소러스 모두 최대값을 기준으로 한 'jcn' 알고리즘 추출 결과와 가장 합치하고 있음을 보이고 있다. 이상을 종합해 보면 최소값을 기준으로 한 결과는 상대적으로 성능이 낙후하며, 'lesk' 알고리즘이 'precision' 측면에서는 'jcn'보다 약간 우수하지만, 'recall'에 있어서는 'jcn'이 더 우수하다. 여기서 'recall'의 값이 'precision' 값보다 크게 못 미치는 경향이 있는데, 이는 군집화로 구성된 결과와 분류를 통해 이루어진 결과의 차이로 인한 것이다. 시소러스가 군집화 결과보다 포괄하는 대상이 더 크기 때문이다. 아울러 비교의 대상이 되는 'random'은 어느 항목에서나 큰 차이를 보이는데, 이는 군집화 결과의 유의미성을 입증하는 부분이다. 이상의 결과를 종합하여, 본고는 최대값 기준의 'jcn' 알고리즘 추출 결과를 인정하여, 이후 과정에 활용할 것이다.

[표 9] F-measure ($\alpha = 0.5$)

	알고리즘	Cobuild	Roget	평균
최대값	jcn	**17.78%**	**11.21%**	**14.50%**
	lesk	16.85%	10.50%	13.68%
최소값	jcn	5.38%	2.12%	3.75%
	lesk	9.72%	3.80%	6.76%
평균값	jcn	17.64%	8.96%	13.30%
	lesk	16.43%	8.71%	12.57%
비교	random	0.18%	0.59%	0.38%

다음 절에서 이루어지는 하위 범주화 틀의 추출은 이 결과에 속한 437개 동사를 대상으로 하며, 마찬가지로 그 다음 절의 관계성 검토 역시 이 결과에 속한 177개 군집을 바탕으로 이루어 진다.

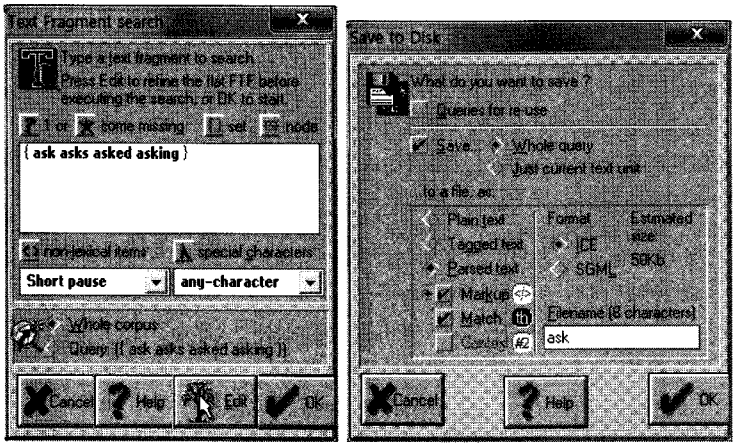
4. 하위 범주화 틀과 선택 선호도

대규모로 영어 동사의 하위범주화 틀에 대한 정보를 얻기 위해서는 사전에 명시된 정보를 취하거나 아니면 코퍼스로부터 추출해야 할 것이다. 그런데 둘 다 현재는 문제가 있다. 사전의 경우 동사별 하위 범주화 틀은 잘 제시되어 있는 편이나 어휘별로 각 틀이 사용된 빈도정보를 얻을 방법이 없다. 반면 원시 코퍼스에서는 동사별 분포 및 빈도 정보는 얻을 수 있으나 코퍼스 내 개별적인 예(token)별로 하위 범주화 틀에 대한 정보를 정확히 추출하는 것이 어렵다. 그 두 가지 요구조건을 고려할 때 가장 바람직한 대상은 구문분석이 철저히 이루어진 코퍼스로, 본고에서는 현재 가용한 대표적인 구문분석 말뭉치 중 하나인 ICE-GB를 택하여 하위 범주화 틀을 추출하였다.[10]

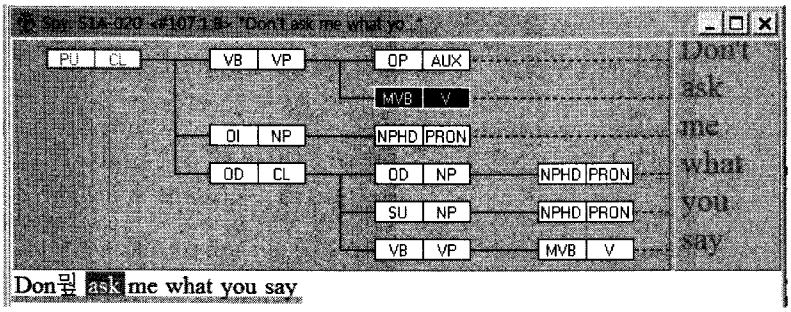
4.1 추출방법 및 절차

ICE-GB를 검색하기 위해서는 우선 검색 대상을 구체화할 필요가 있다. ICE-GB가 굴절 원형에 대한 정보를 담고 있지 않으므로 앞에서 선정된 동사 명단에서 개별 동사별로 굴절형을 모두 검색하는 절차를 거쳤다. 즉, 우선 동사별로 굴절형 명단 (예. *ask*, *asks*, *asked*, *asking*)을 만든 뒤에 ICECUP을 이용하여 해당 용례를 모두 추출하였고, 동사별 검색 결과를 별도의 트리뱅크 형태의 개별 파일로 (예. ask.tre) 저장하였다. 예컨대 [그림 4]와 같이 저장할 경우, [그림 5]에 제시된 문장 구조 정보가 모두 저장된다. [그림 5]의 수형도는 동사 *ask*가 NP인 *me*를 간접목적어(IO)로 취하고, CL(절)

[10] 한 심사자께서는 WordNet의 문형 정보를 이용하여 하위범주화 틀을 가져올 수도 있음을 지적한 바 있다. 물론 이 역시 가능한 방법 가운데 하나이다. 그러나 본고의 목적은 실제의 언어 자원을 일반화 코퍼스로 상정하고 그 빈도 정보를 이용하여 선택 선호도를 측정함을 목적으로 하기 때문에, 이 방식을 사용하지 않았다.



[그림 4] ICECUP III - Text searching / saving



[그림 5] ICE-GB 수형도

인 *what you say*를 직접목적어(OD)로 취하고 있다는 것을 보여준다. 이 수형도가 저장되는 형태는 아래와 같다.

```
<ICE-GB:S1A-020 #107:1:B>
PU,CL(main,imp,ditr,infin,-su)
 VB,VP(ditr,infin,do,neg)
OP,AUX(do,infin,neg) {Don't}
*  MVB,V(ditr,infin) {ask} **
 OI,NP
  NPHD,PRON(pers,sing) {me}
 OD,CL(depend,indrel,montr,pres,preod)
  OD,NP
   NPHD,PRON(nom) {what}
  SU,NP
```

```
   NPHD,PRON(pers) {you}
  VB,VP(montr,pres)
   MVB,V(montr,pres) {say}
...
```

수형도에서 하위 범주화 틀 정보를 추출하기 위해서는, 문장 수형도 별로 해당 동사에서 출발하여 관련 정보를 거두는 방법을 사용하였다. 예를 들어 위에 주어진 수형도에서 우선 다섯 번째 줄의 *ask* 를 출발점으로 한다. 일단 그 것의 최대투사인 (VB,VP)로 올라간 뒤에 그 최대투사와 같은 수준의 절점, 즉 자매절점((OI,NP), (OD,CL)) 중에서 논항이 될 수 있는 범주를 모두 찾아 모으는 방식으로 진행하는 알고리즘을 활용하였다. 아래는 그 알고리즘이다. 결과적으로 각 수형도 및 동사(파일) 별로 모든 논항 분포 와 그 빈도정보를 추출할 수 있었다.

```
1: ARGS = ['OD', 'OI', 'CO', 'CS', 'CI', 'CT', 'PROI', 'NOOD']
2: N_i = {PARSED_TEXT_i, DEPTH_i}
3: T = {N_1, N_2, ...}
4: num = 1
5: for nd in T:
6:     if nd has *:
7:         tmp = num - 1
8:         do until N_tmp is the maximal projection of nd
9:             if PARSED_TEXT_tmp is relevant to ARGS and
                                  DEPTH_tmp is equal to DEPTH_tmp - 1
10:                SF += PARSED_TEXT_tmp
11:            tmp = cnt + 1
12:        do until N_tmp is the maximal projection of nd
13:            if PARSED_TEXT_tmp is relevant to ARGS and
                                  DEPTH_tmp is equal to DEPTH_tmp - 1
14:                SF += PARSED_TEXT_tmp
15:            tmp = num - 1
16:    num = num + 1:
```

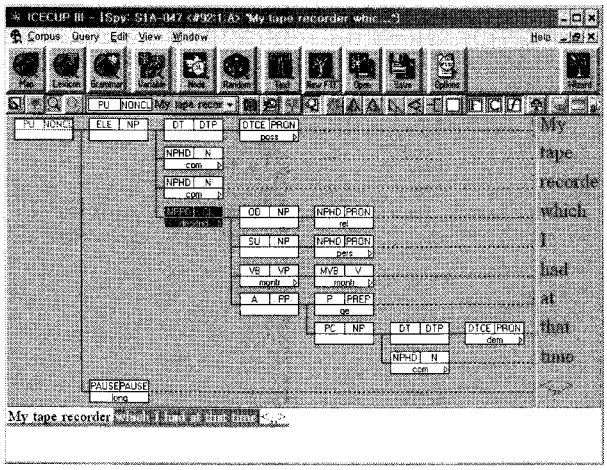
그런데 ICE-GB 자체의 특성 및 일부 한계로 인하여, 하위 범주화 틀 추출 과정에 제기된 몇 가지 문제점이 있었던 바, 다음 소절에서는 그러한 점을 논의해 본다.

4.2 하위 범주화 틀 추출상의 한계 및 제약

4.2.1 주어. 영어 문장에서 주어는 거의 필수적으로 존재한다는 점에서 동사별 하위 범주화 논의에서는 변수라기보다는 상수처럼 간주될 수 있다. 따라서 본고에서는 일단 주어를 포함시키지 않았다. 특히 [그림 5]의 *ask* 경우에서와 같이 명령문 등에서는 주

어가 표면에 등장하지 않더라도 주어가 없는 것으로 간주할 수는 없다는 점을 감안하였다.

4.2.2 관계절 및 의문사 의문문. 관계절의 경우 논항 중 하나가 표면에서는 생략된 구조이므로 일반적으로 코퍼스에서 논항을 추출하는데 있어 적지 않은 문제가 있다. 그러나 ICE-GB에서는 이동한 논항의 선행사도 다른 논항과 마찬가지로 자매절점 형태로 표시되어 있기 때문에 역시 일괄 추출이 가능하다. 한편 의문사 의문문의 경우도 ICE-GB상 자동으로 의문사를 찾는 것이 가능하므로 역시 자료 추출 범위에 포함 시켰다. 다만 관계사나 의문사가 생략된 구문은 배제하였다.



[그림 6] My tape recorder which I had at that time. (S1A-047 #92)

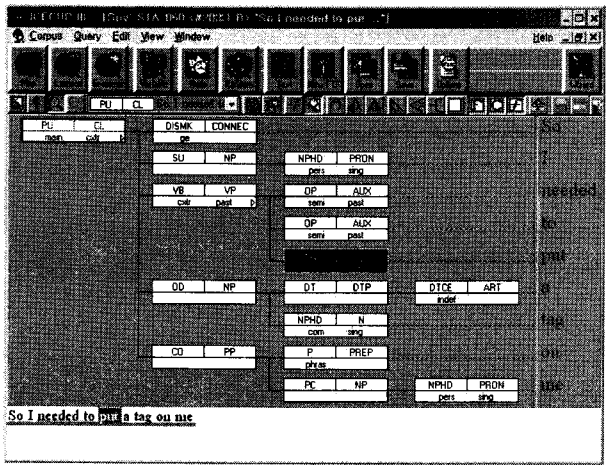
4.2.3 사격보어. 잘 알려져 있듯이 영어에서는 명사구나 절 이외의 요소가 필수 요소로 요구되는 경우들이 있다. 예컨대 *put*의 경우 아래에서 보듯 목적어 명사구는 물론 장소보어도 요구된다 (Levin, 1993).

(8) a. John put the book on the desk.

b. *John put on the desk.

c. *John put the book.

이 점 역시 ICE-GB 수형도상 구분이 가능하도록 되어 있어서 일괄 처리가 가능하다. 위의 장소보어는 다른 부사구와는 달리 (CO,PP)로 표기되어 있으므로, 필수 논항으로 추출하는데 아무런 문제가 없다. 실제 ICE-GB에서 가져온 예는 [그림 7]과 같다.



[그림 7] So I needed to put a tag on me. (S1A-060 #203)

4.2.4 수동태 구문. 수동태 구문의 경우 주어자리의 논항은 해당 구문을 능동태로 변형하였을 경우의 목적어 중 하나와 서로 상응하는 것으로 간주되고 있다. 그런데 어느 목적어와 상응하는지를 ICE-GB에서의 수동태 구문의 수형도 만으로는 파악이 가능하지 않다. 즉, 수동태 구문에서는 자동으로 일괄 논항 요소들을 추출하는 것이 쉽지 않은 듯 하다. 그렇다고 표면형에 의존해서만 추출할 경우 전체적으로 수치상 왜곡이 커질 것으로 판단되었다. 따라서 수동태 구문은 본고에서의 논의에서 모두 배제하였다. ICE-GB의 통계에 따르면 전체 VP중에서 수동태인 경우가 10%를 약간 상회한다.

4.3 선택 선호도

이 소절에서는 ICE-GB에서 추출된 수형도의 하위범주화 틀을 대상으로 각 동사가 어떠한 하위범주화를 가지는가에 대한 계량적 검토를 시도한다. 앞 3절에서 최종적으로 선택된 군집 177개에 속한 동사는 총 437개이다. 이들 동사들을 대상으로 하여 앞 4.1절과 4.2절에서 검토된 방법론에 따라 ICE-GB에서 해당 수형도와 관계된 하위범주화를 모두 추출하였다. 선택 선호도는 이들을 대상으로 하여 계산된다.

선택 선호도를 측정하는 가장 일반적인 모형은 Resnik (1996)에서 제시된 이른바 Kullback-Leibler Divergence 모델이다 (Manning and Schütze, 2002). 이 모델의 연산은 다음 수식에 의하여 이루어 진다. 아래 수식에서 S는 선호 강도('strength')를 v는 해당 동사('verb'), 그리고 sf는 하위 범주화 틀 ('subcategorization frame')을 각각 의미한다. 한편, $P(A|B)$는 B가 일어났다는 조건 하에 A가 일어나는 확률을 말한다.

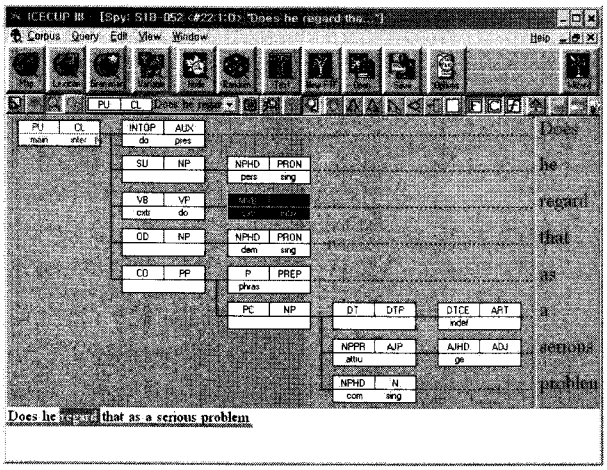
$$S(v, sf_i) = \frac{P(sf_i|v)log\frac{P(sf_i|v)+1}{q(sf_i)}}{\Sigma P(sf_i|v)log\frac{P(sf_i|v)+1}{q(sf_i)}} \quad (7)$$

위 수식은 어떠한 동사에 속한 각각의 하위범주화가 어느 정도의 비율로 출현하는가에 대한 선호 관계를 드러내는데, 아래는 그 예시로서 'regard'의 선택 선호도에서 상위 3개를 보여준다.

[표 10] 'regard'의 선택 선호도

	하위범주화	선택 선호도	Σ
1st	CO,PP / OD,NP	0.4583	0.4583
2nd	CO,PP	0.3311	0.7894
3rd	CO,PP / NOOD,CL	0.0716	0.861

상위 3개의 선택 선호 강도에서 모두 일관되게 'CO,PP'가 출현하고 있으며, 이는 동사 'regard'가 취하는 가장 일반적인 형태의 논항 구조인 'regard NP1 as NP2' 구성과 일치한다. [그림 8]은 실제로 ICECUP에서 검색한 해당 하위범주화 틀의 예시이다.



[그림 8] Does he regard that as a serious problem? (S1B-052 #22)

이러한 방식을 따라 검출된 하위범주화의 선택 선호도를 437개 동사를 대상으로 모두 계산한 다음, 그 결과를 주어진 177개 군집에 따라 재정렬하였다. 아래는 그 일부를 보여준다. 서론에서 살핀 'ask'와 'wonder' 두 동사가 의미적으로는 같은 군집에 속해 있으나, 그 하위 범주화의 선택 선호도 양상이 서로 상이함을 알 수 있다.

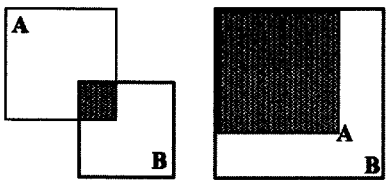
[표 11] 선택 선호도 예시

군집번호	동사	하위범주화 1		하위범주화 2		하위범주화 3	
124	launch	OD,NP	0.7983	-	0.2017	NILL	0
	open	OD,NP	0.6141	-	0.3303	CO,AJP/OD,NP	0.03
175	address	OD,NP	0.6881	-	0.3119	NILL	0
	ask	CT,CL	0.2673	OI,NP	0.2409	OD,CL/OI,NP	0.1577
	question	OD,NP	0.8026	-	0.1425	OD,CL	0.0549
	wonder	OD,CL	0.8442	-	0.1366	OD,NP	0.0192
68	couple	-	0.853	OD,NP	0.147	NILL	0
	pair	NILL	0	NILL	0	NILL	0

5. 연관성 정도 검토

여기에서는 앞 3절과 4절에서 논의되고 추출된 결과를 조합하여 그 연관성을 검토하고자 한다. 양자의 연관성은 6가지로 구분이 되는데, 각각은 VS ('very strong'), S ('strong') A ('average'), W ('weak'), VW ('very weak') 그리고 ?로 표지 되었다. S의 경우에는 군집을 이룬 동사의 하위범주화 선택 선호도가 서로 강한 연관성을 보이는 경우이며, A은 중간 정도의 연관성을, 그리고 W은 약한 연관성을 보이는 경우에 해당한다. 끝으로 ?는 어떤 어휘가 ICE-GB에서 동사로 사용된 용례가 없음을 말하는데, 예컨대 위 [표 11]에서 'pair'가 이에 해당한다.

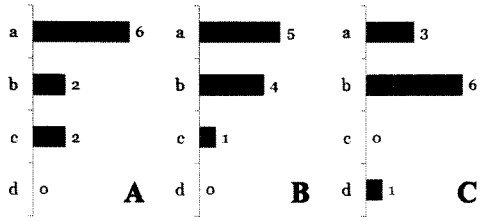
문제는 이들 각각을 판별하는 기준을 마련하는 것이다. [그림 9] 및 [그림 10]을 통해 본고에서 선택한 일치도 측정 기준을 비유적으로 설명하기로 한다. [그림 9] 그리고 [그림 10]은 일치도 정보를 포함하는 가상의 틀에 해당한다.



[그림 9] A와 B의 분포

어떠한 공간 안에 A와 B의 분포가 [그림 9]에서 나타난 바 같이 정사각형의 형태로 존재한다고 가정하자. 이때 양자가 얼마만큼의 분포를 공유하는가의 문제는 [그림 9]에서 패턴으로 표시된 부분처럼 둘의 공통 부분, 다시 말해 교집합에 해당하는 영역의 면적을 구하는 문제와 같다. 이때, 오른쪽의 그림처럼 A가 B에 포함되는 형태로 존재한다면, 이 경우 교집합은 A의 분포와 같고, 따라서 둘 가운데 작은 값을 취하는 A의

면적이 곧 공유되는 부분에 해당한다. 이를 본 소절의 연관성 포착에 적용하여 보기로 하자.



[그림 10] 공유 분포 추정

[그림 10]에서 A, B, C 각각은 a, b, c, d 가운데 세 항목을 대상으로 값을 취하며 그 각각의 분포는 위 그림과 같다. 이때, a, b, c, d 각각은 각 하위범주화 틀에 해당한다고 가정하자. 이 때, A, B, C가 공통으로 취하는 영역의 값을 취한다면, 이는 [그림 10]과 마찬가지로 a, b, c, d 각각에서 최소의 값을 보이는 항목을 선택하여, 그 값들의 합을 구하는 것과 동일하다. 위 [그림 10]에서는 a는 C에서 3을 취하고, b는 A에서 2를 취하며, c와 d의 경우에는 공통의 영역이 없어 0의 값을 가진다. 결과적으로 공통 부분의 면적은 5가 될 것이다.

논항 관계의 전체 일치도 계산은 이러한 방식에 의거하여 계산된다. 5단계로 구분된 각 연관성 정도는 0.2 단위로 분할되는데, 예컨대, VS는 0.8 이상의 일치도를 보이는 경우를 말한다. 이는 다시 말해, 어떤 어휘군이 서로 VS의 관계성을 가질 때, 그 해당 어휘를 같은 군집에 속하는 다른 어휘로 치환하였을 경우, 그 바뀌어진 문장이 의미적으로도 상통하고 또한 통사적으로도 정문일 확률이 80% 이상임을 의미한다. 앞의 [표 11]을 그 일치도의 여부에 따라 관계성을 기술하면 [표 12]와 같다. 이때, 'pair'는 동사로서의 쓰임이 없기 때문에, 군집 68의 경우 관계성은 ?로 명시되고 당연히 그 일치도 역시 0으로 결정된다.

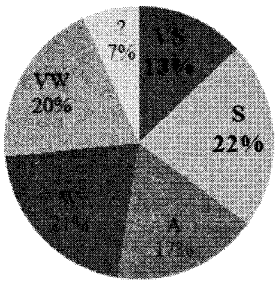
[표 12] 선택 선호도 예시

번호	일치도	관계성	동사	하위범주화 1		하위범주화 2		하위범주화 3	
124	0.8158	VS	launch	OD,NP	0.7983	-	0.2017	NILL	0
			open	OD,NP	0.6141	-	0.3303	CO,AJP/OD,NP	0.03
175	0.0000	VW	address	OD,NP	0.6881	-	0.3119	NILL	0
			ask	CT,CL	0.2673	OI,NP	0.2409	OD,CL/OI,NP	0.1577
			question	OD,NP	0.8026	-	0.1425	OD,CL	0.0549
			wonder	OD,CL	0.8442	-	0.1366	OD,NP	0.0192
68	0.0000	?	couple	-	0.853	OD,NP	0.147	NILL	0
			pair	NILL	0	NILL	0	NILL	0

이에 따라 수집된 관계성에 대한 분포는 아래 표와 같다. 우선 군집을 기준으로 하여 그 연관성을 포착하여 보면 아래와 같이 VS 및 S의 관계, 다시 말해 유사한 의미로 묶인 동사군이 하위범주화 틀을 공유할 확률이 큰 경우는 전체의 35%에 속하며,

[표 13] 군집별 관계성

	(A) 군집개수	%	Σ
VS	23	12.99%	12.99%
S	39	22.03%	35.03%
A	31	17.51%	52.54%
W	37	20.90%	73.45%
VW	35	19.77%	93.22%
?	12	6.78%	100%

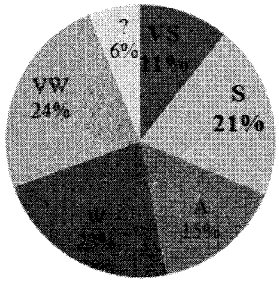


다시 W 이하 및 ?의 관계를 제외하고 나면 약 52%의 합치를 보인다.

동사를 기준으로 정렬하였을 경우에는 [표 13]보다 약간 낮은 확률을 보이고 있는데, 이는 아래 표에 나타난 바와 같이 W에 속하는 군집이 대체로 소속된 동사의 개수가 비교적 많기 때문이다.

[표 14] 동사별 관계성

	(B) 동사개수	(B)/(A)	%	Σ
VS	47	2.04	10.76%	10.76%
S	90	2.31	20.59%	31.35%
A	68	2.19	15.56%	46.91%
W	99	2.68	22.65%	69.57%
VW	106	3.03	24.26%	93.82%
?	27	2.25	6.18%	100%



이번에는 구체적인 자료를 검토해 본다는 차원에서, [표 12]에 나온 예를 일부 살펴보도록 하자. 우선 VS에 속한 그룹의 'launch', 'open'이 실제로 출현한 예문은 아래 (9), (10)과 같다.

(9) a. Well after the preamble we can now **launch** into the evidence of Egyptian art for this period (S2A-052 #61)

b. Top scientist will **launch** country's space age project. (W2C-017 #91)

(10) a. How late's Wombles **open** up. (S1A-011 #185)

b. **Open** the wine. (S1A-065 #317)

위에서 (9a, 10a) 및 (9b, 10b) 각각은 서로 동일한 논항 구조로 실현되어 있다. 각각의 실현 양상은 '-', 'OD,NP'에 해당한다.

반면 VW에 속한 (2)의 'ask', 'wonder'는 위 [표 11]에 드러난 바와 같이 그 논항 실현 양상이 판이하다. 'ask'가 아래 (11)에서 차례로 예시된 바와 같이 'CT,CL (0.2673)', 'OI,NP (0.2409)', 'OD,CL/OI,NP (0.1577)' 등의 비교적 고른 분포의 다양한 논항 실현 양상을 보인다면, 'wonder'는 대개의 경우 (12a)와 같이 'OD,CL (0.8442)'로 고정되어 있음을 알 수 있다.

(11) a. They **asked** me to cover for them. (W2F-006 #143)

b. Tomaso had **asked** his wife. (W2F-016 #39)

c. And don't **ask** me what it's called because I couldn't tell you. (S1A-016 #144)

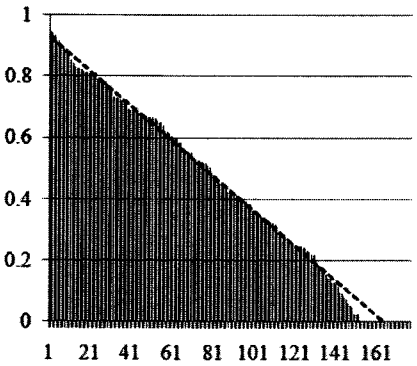
(12) a. I **wonder** what it is. (S1A-054 #93)

b. I was **wondering** that. (S1A-023 #280)

기존 이론적 논의에 나온 예문 (2)와 본 연구의 결과로 정리된 [표 11] 및 예문 (11)-(12)를 비교해 볼 때, 본고에서의 연구가 어떤 차별적 기여를 할 수 있는지를 'ask'와 'wonder'를 통해 좀 더 구체적으로 설명할 수 있다. (2)가 동사별 하위범주화 유형(type)을 보여주는데 그치고 있는 반면, 본고의 결과는 유형뿐만 아니라 그 유형이 개별적으로 어떤 빈도로 쓰이고 있는지, 그리고 상대적으로 어떤 분포적 특질을 지니는지를 상세하게 보여주고 있다. 두 번째로, (2d)에서는 'wonder'가 명사성 어휘를 직접목적어로 취하는 경우를 비문으로 파악하고 있으나, (12b)와 같은 구문이 코퍼스에서는 출현하고 있다는 점은, 그리고 그것이 상대적인 비중이 작지 않다는 점은, 연구자 자신의 직관에만 의존하는 연구의 한계의 한 면을 보여주고 있다. 거기에 더해서 'wonder'가 명사구를 직접목적으로 취할 수 있다는 것을 받아 들인다면, (2)에 제시된 'ask'와 'wonder' 사이의 분포적 차이는 사라지게 되고 마치 'ask'와 'wonder'가 동일한 하위범주화를 취하는 것으로 인식될 것이다. 그러나 본고의 연구 결과에 따르면 'wonder'가 명사구를 직접목적어로 취한다 해서, 그것이 곧 'ask'와 'wonder'가 유사한 하위범주화적 특성을 띤다는 결론을 내릴 수는 없다는 점을 명시적으로 제시하고 있다. 즉, 하위범주화 틀에다가 선택 선호도를 결합할 경우, 이론적 연구에서처럼 하위범주화만 고려한 경우와는 전혀 다른 결론이 나올 수 있다는 점을 보여주고 있다. 이러한 점들은 본 논문에서의 연구가 기존 이론적 연구에 비해 한 차원 추가된 측면을 고려하고 있다는 점과, 아울러 기존 연구에 비해 훨씬 더 세밀한 분류를 가능케 해 준다는 점을 입증해 준다.

5.1 일치도의 점진성

[표 13] 및 [표 14]의 원형 도표를 보면 범주별로 분포가 고른 편이다. 일치도의 전체 분포를 시각적으로 보다 잘 보여주기 위해 의미 유사도와 하위범주화 선택 선호도 사이의 일치 정도를 막대 도표로 변환해 볼 수 있다. [그림 11]은 177개 군집에 대한 각 일치도를 역순으로 정렬하여 이를 도표로 나타낸 것이다. 가운데 굵은 점선으로 표시된 것은 전체 값의 추세선을 나타낸다. 추세선이 드러내는 분포적 특징은, 각 일치도 값이 매우 점진적이라는 점과 아울러, 경계절점을 명확히 규정하기가 쉽지 않다는 점이다.



[그림 11] 일치도의 점진적 분포

실제로 몇몇 섬 제약 역시 점진적인 (squish) 분포를 드러낸다는 점은 Fodor (1983) 등에서도 논의된 바 있으며, Kluender (1998)를 위시한 연구에서도 본고의 구분법과 유사하게 강한 ('strong') 제약과 약한 ('weak') 제약이라는 다소 상대적인 개념을 통해 제약의 분포 특성을 파악하려 한 바 있다. 나아가, 최근의 Hofmeister and Sag (In Press) 등에서는 주요 통사적 제약들이 가지는 넓은 범위의 문법성 스펙트럼 설명하기 위해 처리율 (Processing Account)과 같은 개념을 도입하고 있다. 이를 본 연구에 적용하자면, 위에서 말하는 일치도 역시 참 또는 거짓의 개념이라기보다는 처리율상의 정도성 차이로 파악하는 것이 타당하다 하겠다. 즉, 'launch', 'open' 등의 이른바 강한 일치를 보이는 군집은 논항구조의 동일성이 쉽게 발견되는 반면, 'ask', 'wonder' 등의 이른바 약한 일치를 보이는 군집은 그렇지 못하다고 정리할 수 있다.

6. 결론

본고에서는 영어 동사의 논항 구조와 의미 사이의 연관성 정도에 대하여, 유사 의미 동사가 하위범주화적 특성도 공유하는가라는 관점에서 검토를 하였다. 구체적으로는 의미적으로 유사한 동사 묶음을 최대한 찾아낸 뒤에 그 묶음 별로 동사 사이의 하위범주화 일치도가 어느 정도인지를 판정하였고, 그러한 결과는 매우 상세한 표로 제시되었

다[11]. 그 표에 근거해 추가 분석을 해 본 결과, 일치도의 분포가 점진적인 성격을 띠고 있으며, 의미적 유사성이 하위범주화 상의 유사성으로 모두 반영되는 것은 아니라는 점을 확인할 수 있었다. 어떤 동의어 동사 묶음의 경우엔 하위범주화상의 유사성이 큰 반면, 반대로 의미적으로 유사한 동사들인데도 하위범주화상 특성이 공유되지 않는 경우들도 있었다. 그러한 분포가 점진적인 특성을 지니고 있다는 점은, 문법의 주요 현상들이 점진적인 분포를 보인다는 최근의 언어학적 연구결과들과도 부합된다. 특히 본 연구는 하위범주화에 대한 유형뿐만 아니라 각 유형이 얼마나 잦은 빈도로 쓰이고 있는지, 그리고 그러한 빈도가 상대적으로 어떻게 평가되는지에 따라 동사별 하위범주화 특성을 고려해 보아야 한다는 점을 보이고 있다.

본 연구에서는 전산언어학적인 방법론을 최대한 반영하였다.[12] 직관에 의존하는 전통적인 언어학적 연구 방법론을 취할 경우엔, 개별 어휘에 대한 꾸준한 분석 작업을 통해 본고가 출발점으로 삼은 의문에 답을 할 수 있을 것이다. 그러나 개인의 직관에 의존한 연구 및 분류는 뚜렷한 한계가 있다. 특히 소수의 연구자가 대규모 자료를 처리하는 데는 그것에 드는 시간과 자원을 감당하기 어렵고, 무엇보다도 객관적인 검증이 쉽지 않다는 문제점이 있다.[13] 따라서 이러한 개별적, 예시적, 직관적 연구 방법론을 대체할 수 있는 객관적인 방법론, 아니면 적어도 최소한 그러한 기존 방법을 보완할 수 있는 방법을 이용한 연구가 필요하다. 잘 알려져 있다시피, 이미 구축되어 공개된 코퍼스 및 언어자원을 이용한 최근의 전산언어학적 방법론이 이에 대한 한 가지 답이 될 수 있다. 이러한 방법론을 취할 경우, 1 차 자료의 수집과 정리가 연구자와 분리된다. 그 결과, 연구자 자신의 직관에는 극히 제한적으로만 의존하게 되고, 따라서 혹시라도 연구자의 이론적 성향에 따라 자료 수집과 판정이 좌우될 가능성을 원천적으로 피할 수 있다. 본 연구 역시 이러한 전산언어학적인 방법론을 취하여 의미와 하위범주화 사이의 연관성에 대한 한 가지 탐구 방식 및 그 결과를 제시하고 있다. 특히 의미 유사도나 하위범주화 선택 선호도 등을 모두 수치화하여 세밀한 계산 및 비교, 해석 등이 가능하도록 하였다는 점이 본 연구의 특징이라고 할 수 있다.

이러한 연구 결과 및 특징에도 불구하고, 본 연구는 문제를 해결한 것 이상으로 더 많은 의문을 제기하고 있다. 그 중 하나는, 어휘별 의미적 분화, 즉 어휘가 여러 어의(sense)를 가질 수 있다는 점을 충분히 고려할 수 없었고, 따라서 그러한 면에서의 한

[11] 전체 표는 아래의 주소에서 다운로드 받을 수 있다.
`http://corpus.mireene.com/download/wn-sim.html`

[12] 본 연구의 각 단계에서는 해당 목적에 따라 ANSI C++, Perl, Python 등의 프로그래밍 언어가 활용되었다. 우선 의미적 유사도 측정과 관련하여서는 WordNet::Similarity 를 바로 이용하기 위해 Perl 을 사용하였으며, 군집화 알고리즘은 산술 연산이 매우 많은 관계로 실행 속도가 빠른 ANSI C++로 구현하였다. 끝으로 선택 선호도 및 일치도 계산은 텍스트 처리에 능동적인 Python 으로 구현하였다. 이 가운데, 주어진 데이터에 독립적일 수 있는 모듈은 의미 유사도 측정 프로그램과 군집화 프로그램이다. 이들 역시 앞서 언급한 고빈도 어휘 유사도 목록과 함께 아래의 주소에 소스 전체가 공개되어있다
`http://corpus.mireene.com/download/wn-sim.html`

[13] 이는 'annotate automatically, correct manually'라는 전산언어학 일반에서 입증된 최적 방법론과 궤를 같이 한다 (Marcus, Marcinkiewicz, and Santorini, 1993).

계가 본 연구에 어떤 형태로 영향을 끼쳤는지가 확인되어야 한다(§3.2). 다른 하나는 선택 선호도를 바탕으로 한 하위범주화 일치의 정도를 계산하는 방식의 타당성이다. 거기에 더해 의미적 유사성을 결정하는데 사용한 척도(*z-score*)가 과연 타당한 것인지에 대한 추가 검증도 필요하다. 그 밖에도, 본 연구가 매우 세밀한 절차와 다량의 계산을 바탕으로 한 것이기 때문에, 단계별, 계산식별 대안의 여지는 충분하다. 또한 근본적으로 본고가 활용한 언어자원, 즉 ICE-GB, WordNet, COBUILD 등이 지니는 내적 한계를 벗어날 수 없다는 점도 본고의 한계 중 하나라 할 수 있다. 이러한 문제점과 의문점들은 그 하나 하나가 흥미로운 쟁점이면서 추가 연구가 필요한 주제라 할 수 있다. 또한 그러한 한계에도 불구하고 본 연구에서 타당성이 일부 객관적으로 입증되는 연구 결과가 나왔다는 점은, 본 연구와 같은 방식의 추후 연구 필요성을 뒷받침하고 있다고 본다.

본 연구 성과의 응용분야는 이론언어학적인 연구와 전산언어학적 연구로 구분된다. 우선 이론언어학적으로는 그 동안 연구자의 직관을 통해 구축된 언어자료 (예컨대, (Levin, 1993))의 실효성에 대해서 검증을 하고, 이를 통해 보다 실제 자원에 기반한 결과물을 얻어낼 수 있을 것이다. 뿐만 아니라, 앞서 5.1 절에서 살펴본 바와 같이 처리율에 따른 문법성 스펙트럼을 경험적으로 확인할 수 있는 바탕을 제공한다 (Hofmeister and Sag, In Press). 반면 전산언어학적 연구로는 본고에서 정리된 동사의 의미별 군집 및 선택 선호도 관계표를 바탕으로 하여 어의 중의성 해소 (Word Sense Disambiguation), 의미역 자동 부착 (Semantic Role Labeling), 및 기계 번역 등의 성능 향상을 도모할 수 있다 (Erk, 2007; Chan, Ng, and Chiang, 2007).[14]

< 참고문헌>

Banerjee, Satanjeev and Ted Pedersen. 2003. Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, Acapulco, Mexico.

Budanitsky, Alexander and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics* 31, 13–47.

Chan, Yee Seng, Hwee Tou Ng, and David Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic.

Erk, Katrin. 2007. A Simple, Similarity-based Model for Selectional Preferences. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic.

[14] 현재 필자들이 다음단계로 구상하고 있는 연구 대상은 크게 두 가지이다. 하나는 마찬가지의 연구를 한국어에 적용하는 것이며, 다른 하나는 한국어 및 영어의 결과물을 기반으로 해서 구문분석기 및 기계번역의 성능을 개선하는 것이다. 전자의 경우 WordNet은 KorLex(`http://korlex.cs.pusan.ac.kr`)를, 일반화 코퍼스는 세종 구문분석 말뭉치를 대상으로 한다. 후자의 경우에는 규칙기반과 통계기반으로 나누어 구문분석기와 영-한/한-영 기계번역 환경에 각각 결과를 적용해 볼 것이다.

Fodor, Janet D. 1983. Phrase Structure Parsing and the Island Constraints. *Linguistics and Philosophy* 6, 163–223.

Hofmeister, Philip and Ivan A. Sag. In Press. Cognitive Constraints and Island Effects. *Language* 86.

Jiang, Jay J. and David W. Conrath. 1997. Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, Taiwan.

Jurafsky, Daniel and James H. Martin. 2009. *Speech and Language Processing (second edition)*. Pearson Prentice Hall, Upper Saddle River, NJ.

Kaufman, Leonard and Peter J. Rousseeuw. 1990. *Finding Groups in Data : An Introduction to Cluster Analysis*. Wiley, New York.

Kluender, Robert. 1998. On the Distinction between Strong and Weak Islands: a Processing Perspective. In Peter Culicover and Louise McNally (eds.), *Syntax and Semantics 29: The Limits of Syntax*. Academic Press, San Diego, CA, pp. 241–279.

Lasnik, Howard, Juan Uriagereka, and Cedric Boeckx. 2005. *A Course in Minimalist Syntax: Foundations and Prospects*. Wiley-Blackwell, Malden, MA.

Lesk, Michael. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: how to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the the 5th Annual International Conference on Systems Documentation*, Toronto, Ontario, Canada.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press, Chicago, IL.

Lin, Dekang. 1993. Principle Based Parsing without Overgeneration. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio.

Manning, Christopher D. 1993. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio.

Manning, Christopher D. and Hinrich Schütze. 2002. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marcus, Mitchell P., Mary A. Marcinkiewicz, and Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: the Penn Treebank. *Journal of Computational Linguistics* 19, 313–330.

Nelson, Gerald C., Sean Wallis, and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English (Varieties of English Around the World)*. John Benjamins Publishing Co., Amsterdam, The Netherlands.

Pedersen, Ted. 2008. Empiricism Is Not a Matter of Faith. *Computational Linguistics* 34, 465–470.

Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. In *Proceedings of the 19th National Conference on Artificial Intelligence*, Boston, MA.

Resnik, Philip. 1995. Using Information Content to Evaluate Semantic Similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Quebec, Canada.

Resnik, Philip. 1996. Selectional Constraints: An Information-Theoretic Model and its Computational Realization. *Cognition* 61, 127–159.

송상헌 · 전지은 · 최재웅. 2008. 영어 '형용사+전치사구' 구문의 의미적 제약: ICE-GB와 WordNet을 활용한 통계적 검증. *언어와 언어학* 41, 75–103.

접수 일자: 2010년 5월 7일
게재 결정: 2010년 5월 24일