언어 자료를 활용한 한국어 복합명사 구조 분석

김동성
(고려대학교)

Kim, Dong-Sung. 2011. Analysis of Korean Complex Nominals, Using Linguistic Data. *The Linguistic Association of Korea Journal. 19*(3). 129-150. This paper introduces an analysis on the structure of Korean complex nominals. The analysis attracts both theoretical linguistics and language processing related areas, such as information retrieval and speech synthesis. Our approach has three stages. First, we identify endocentric data and exocentric data, using human intuition. Since exocentric data does not have the internal structure, we do not consider exocentric data. Second, we do the bracketing experiment on the endocentric data for representing a hierarchical structure of constituent parts, using statistical collocation measurements based on the 10 million Sejong corpus. The last stage is composed of several processes to figure out head-modifier or predicate-argument relations, using argument structure and selection restriction specified in the Sejong electronic dictionary. Our method is based on not only the corpus-based materials but also linguistic knowledge (with intuition-based judgement). The importance of our approach is to show how to use language resources to utilize linguistic knowledges in analyzing linguistic data.

Key Words: Morphology, Collocation, Complex Nominals, Computational Linguistics, Corpus Statistics, Computational Morphology, Endocentricity/Exocentricity, Bracketing

1. 머리말

Spencer (1993)의 분석에 따르면 영어에서 세 개의 명사가 연쇄적으로 만들어내는 복합명사 구조의 경우에 (1)은 괄호매김(Bracketing)을 어떻게 하느냐에 따라서 구조적으로 (2a)나 (2b)로 분석될 수 있다.

(1) student film society

(2) a. [student [film society]] (=film society for students)
b. [[student film] society] (=society for student films)

한국어 복합명사 구조의 경우도 (2a, b)와 같은 구조적 중의성이 발견된다. 예를 들어서 '한국 교회사 연구소'라는 복합명사 구조는 (3a)나 (3b)의 구조 분석 중 하나가 가능하다.

(3) a. [[한국 교회사] 연구소]
(=한국 교회들의 역사를 연구하는 연구소)
b. [한국 [교회사 연구소]]
(=한국을 포함한 세계의 교회사를 연구하는 연구소로
한국에 위치함)

복합명사 구조는 먼저 내심구조(endocentric) 합성 방식인지, 외심구조(exocentric) 합성 방식인지 구분한다. 외심구조는 동격 합성어를 포함하는데, 내심구조의 경우에는 각 구성요소들 간 관계성이 문장성분들 사이에서 관찰되는 관계와의 유사성을 분석할 수 있지만, 외심구조 합성 방식은 이러한 구조 분석이 불가능하다. Spencer (1993)에 따르면 이러한 내심구조 복합 구조 분석에는 형태론적 정보 이외에도 통사론적 정보도 필요한데, 핵-수식 관계와 술어-논항 관계가 분석된다.

복합명사 구조 분석의 중요성을 살펴보면 다음과 같다. 첫째로는 명사구 합성의 생산성(Levi 1978)과 연관된 측면이다. 명사구 합성은 생산성이 높은 형태론적 과정으로 만약 언어학적으로 예측이 가능하지 않다면 새로운 구조가 생성이 될 때마다 새로운 구조에 대한 분석이 있어야 하므로 효용적이지 않다. 따라서 합성구조 생산성에 대한 언어학적 분석에 대한 예측이 가능해야 하며, 이러한 명사 합성구조에 대한 연구는 언어 이론적 측면에서 형태론적 합성어 분석의 측면에서 주요하다(Spencer 1993; Selkirk 1982; Lieber 1983; Di Scullo and Williams 1987; Sproat 1985; Roeper 1988).

둘째로는 정보검색을 위한 목록인 색인어를 작성할 때, 목록화 작업에서 복합명사구의 분석은 중요하다. 예를 들어서 '강남 성심 병원'과 '강북 성심 병원'은 '성심 병원'이라는 색인어로 통일할 수가 있으므로, [강남 (강북) [성심 병원]]의 구조로 분석할 수가 있다.[1] 이와 같이 복합명사 구조의 분석은 정보 처리의 관점에서 필수적이다.

마지막으로 음성합성이나 음운론 연구의 경우에 띄어 읽기는 운율의 단위를 결정하는 주요 요소이다. 띄어 읽기의 경우에 '리들리 스콧 감독'의 구조는 [[리들리 스콧] 감독]이지 [리

1) 남지순·최기선(1997)에서는 정보검색을 위한 어휘사전을 구성을 제안하였다. 정보검색 색인기법과 관련된 부분은 원형석 외 (2000) 참조.

들리 [스콧 감독]]이 아니다. 어떠한 분석을 하는가에 따라서 다른 유형의 운율 단위가 결정된다(Selkirk 1982). Liberman and Sproat (1992)에서는 영어의 경우에 복합명사구의 구조적 차이가 강세를 결정하는데 매우 중요한 요소라고 주장하였다.[2] 전산적 언어처리 분야인 정보검색, 음성합성과 같은 분야에서 단위 설정과 관련되어서 중요하게 취급된다(남지순·최기선 1997; 윤보현 외 1997).

본 연구에서는 언어 자료를 활용해서 [명사1 명사2 명사3]의 복합명사 구조를 분석하는데, 활용하는 언어 자료는 '21세기 세종계획 1,000만 어절 형태소 분석 코퍼스(이하 세종코퍼스)'와 '21세기 세종계획 세종전자사전(이하 세종전자사전)'이다. 내심 및 외심구조를 구분하고 자료를 정련하기 위해서 화자 직관을 사용하였으며, 괄호매김을 위해서 세종코퍼스를 활용해서 통계적 연어 측정 기재를 사용하였다. 외심구조로 판정된 구성 성분들의 관계성을 파악하기 위해서 세종코퍼스와 세종전자사전을 사용하였으며, 구조적 관계성은 핵-수식, 술어-논항 구조로 분석하였다. 술어-논항의 경우에는 각각의 논항들을 구별하였다.

논문의 구성은 다음과 같다. 2절은 언어학적 논의와 언어 자료를 활용한 연구들을 개관하고 이론적 토대를 소개한다. 3절은 연구를 위한 자료추출 절차와 통계적 분석 기재를 보일 것이다. 4절에서는 합성의 유형을 분석하기 위한 연어 측정을 소개하고, 세종코퍼스를 활용한 연구를 제시할 것이다. 5절에서는 술어-논항과 핵-수식 분석을 하는데, 술어-논항 구조 분석을 위해서 세종전자사전의 논항구조를 활용한다. 6절은 이 논문의 결론이다.

2. 기존 연구 개관을 통한 이론적 토대 소개

앞서 논의한 바와 같이 복합명사 구조에 대한 분석은 크게 핵이 내부에 있는 내심구조와 내부에 없는 외심구조로 구분한다. 핵이 내부에 없는 외심구조의 경우는 고유 명사화된 경우가 많고 구조적 중의성이 없다. 내심구조의 경우에는 핵-수식, 술어-논항 관계와 같은 통사적 관계와 유사성도 분석해야 한다.

2.1. 내심 및 외심구조 분석

Levi (1978)에서는 여러 유형의 복합명사 구조를 분류하고, 구조를 변형문법의 틀에서 도출하였다. 서론에서 기술한 바와 같이 합성 방식은 핵이 내부에 있는가 없는가에 따라서 내심구조인지 외심구조인지 구분한다. 복합명사 구조는 핵이 내부에 있는 내심구조는 이분지 제약(Binary Branching Constraints)에 따라 두 개씩 짝지어지는 구조로 분석이 되지만

2) 운율 정보를 할당하기 위해서 Liberman and Sproat (1992)에서는 코퍼스에서 발견되는 통계적 정보를 측정하였다.

(Aronoff 1976; Scalise 1984), 핵이 내부에 없는 외심구조는 이분지 제약을 따르지 않는다. 따라서 (2a, b)의 내심구조는 두 개씩 짝지어진 계층 구조를 만들게 되지만, 'England-France'나 'mother-child'와 같은 외심구조는 내부에 핵이 없는 동격 관계를 이루게 된다(Spencer 1993).

내심구조는 괄호매김의 모호성으로 인해서 구조적 관계를 설정해야 하지만, 외심구조는 고유 명사화된 것처럼 행동하며 구조적 중의성이 없다. 따라서 외심구조와 내심구조를 먼저 분리하는 것이 중요하다. 예를 들어서 백악관을 가리키는 'White House'인 경우는 형용사와 명사의 복합구조로 '하얀색 집'을 가리키는 것이 아니라 '백악관'을 가리키는 고유 명사화된 구조이다. 이름이나 성이 호칭과 복합적으로 구성된 'Mr. Big'이나 'President Clinton' 또는 '오바마 대통령'과 같은 구조는 외심구조이다. 또한 '남녀 고용 평등법'과 같이 일반명사가 합쳐진 구조인 경우에도 고유명사화된 어휘화 항목으로 발전한 외심구조가 발견된다.

내심구조는 여러 어휘들과의 결합이 가능해서 생산성이 높은 반면에 외심구조는 내심구조보다 생산성이 낮다. 외심구조는 단일한 구조로 구조적 분석이 없이 어휘들이 접합한 것으로 분석할 수 있다. 반대로 내심구조는 유사한 유형의 어휘가 결합된 경우라도 다른 괄호매김이 가능하다. 예를 들어서, '홍콩 지국'과 같은 외심구조 유형은 '홍콩 지국 지국장' 또는 '홍콩 지국 개설'과 같이 활용되면 새로운 어휘가 접합된 (4a, b)의 구조만이 가능하다.

(4) a. [[홍콩 지국] 지국장]
b. [[홍콩 지국] 개설]

반면에, 내심구조는 괄호매김의 중의성으로 인해서 다른 구조 해석이 가능하다. '한국 방송 개발원'과 '한국 방송 문화원'일 경우에는 (5a, b)와 (6a, b)의 구조 해석이 가능하다.

(5) a. [[한국 방송] 개발원] (=한국 방송과 연관된 개발원)
b. [한국 [방송 개발원]] (=한국에 있는 방송 개발원)
(6) a. [[한국 방송] 문화원] (=한국 방송과 관련된 문화원)
b. [한국 [방송 문화원]] (=한국에 있는 방송 문화원 단체)

Levi (1978)에서는 등위적 합성구조도 외심구조와 유사하게 동격 관계의 합성구조로 분석하였다. 예를 들어서 '전자 정보 통신'이나 '문화 체육 예술'과 같은 구조는 [명사1 [명사2 명사3]]나 [[명사1 명사2] 명사3]의 구조로 분석되지 않고, 어휘의 나열인 [전자 정보 통신]과 [문화 체육 예술]과 같은 구조만 가능하다.

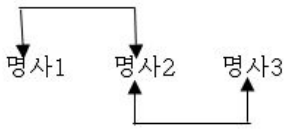
2.2. 핵-수식 및 술어-논항 구조 분석

내심구조의 구조적 관계는 핵-수식어 관계로 구성되었는지, 아니면 논항-술어 관계로 구성되었는지에 따라서 구분될 수 있다. 두 관계 모두 [명사1 명사2 명사3]의 구조에서는 괄호매김의 모호성이 있으므로, (7a, b)의 관계 중 하나를 선택해야 한다.

(7) a. [[명사1 명사2] 명사3]
b. [명사1 [명사2 명사3]]

다음 (8)을 고려해보자. 괄호매김은 (8)에서와 같이 단어 사이의 관계성을 나타낸다. (8)을 활용하면 '명사1 명사2'의 관계성과 '명사2 명사3'과의 관계성을 서로 비교할 수 있으며, (7a, b) 중 어느 괄호매김을 선택할 수 있을지를 알 수 있다.

(8)



외심구조로 묶여진 명사들이 하나의 단위로 행동하며, 묶여진 명사들 간에는 핵이 없다. 이름이나 성과 호칭이 결합한 구조는 고유명사화된 구조이다. '김연아 선수 발표'와 같은 구조에서 '김연아 선수'는 이전에 설명한 바와 같이 동격명사 구조로 고유명사화된 것처럼 행동한다. 이러한 고유명사가 포함된 구조분석에서 고유명사와 연이은 명사들의 유사성이 다른 명사보다 더 높게 나타난다. 따라서 (9)와 같은 구조 분석이 가능하다.

(9) a. [[김연아 선수] 발표]
b. [유엔 [갈리 총장]]
c. [[행정부 담당관] 처리]

그러나 일반명사가 연이은 대부분의 구조에는 괄호매김에 모호성이 있다. 이러한 구조 분석에서 각각의 일반명사 간의 유사성을 따져 보아야 한다. 언어 자료를 활용해서 유사성을 따져보기 위해서는 코퍼스 자료에서 나타나는 통계적인 차이를 확인한다(Lauer 1995). 여기서 통계적 차이는 (8)과 같이 '명사1 명사2'와 '명사2 명사3'의 연어(collocation) 정보의 강도를 측정을 통해서 나타난다. 연어 정보는 두 단어가 하나의 연어 단어로서의 통계적 측정이 된다. 이러한 연어 정보 특성은 두 단어가 통계적으로 얼마나 많이 또는 자주 공기

(cooccurrence)하는가의 통계적 차이를 말한다. Lauer (1995)는 이러한 통계적 차이를 발견하기 위해서 로제 시소러스(Roget's Thesaurus)를 활용하였는데, 본 연구에서는 코퍼스에서 발견도는 통계만을 활용하였다. 4절에서 이와 연관된 부분을 자세히 논의할 것이다.

복합명사 구조는 크게 수식 구조와 논항 구조로 구분될 수 있다. 수식 구조는 두 명사 간에 중심어와 수식어의 관계가 발견되는 핵-수식 구조이고, 논항 구조에서는 동사성 명사로 인하여 각각의 명사에 대한 논항 부여가 가능하다.

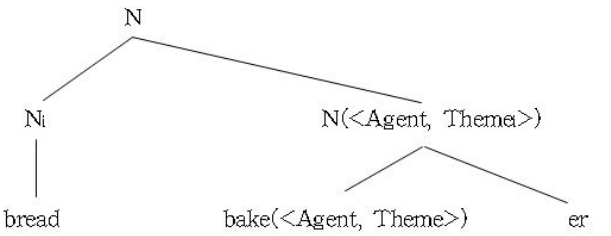
먼저 수식 관계를 살펴보면 다음과 같다. 예를 들어서 'student committee'의 경우에 'committee'가 중심어가 되고 'student'가 수식어가 되어서 중심어인 'committee'를 수식한다.

술어-논항 구조의 경우는 언어학적으로 많은 연구가 있었다. 그 중 Di Sciullo and Williams (1987)는 주어-동사가 있는 문장에서 서술어가 각 논항에게 부여하는 논항 구조가 합성어의 경우에도 가능하다고 주장하였다.[3] 술어-논항 구조의 경우에 'bake'동사는 [Agent(주어) ____ Theme(목적어)]를 취하는 구조로 (10a)와 같은 문장 구조를 생성한다. 이 경우에 명사구로 전환한 (10b)와 복합어 구조인 (10c)는 (10a)와 동일한 논항 구조를 공유한다.

(10) a. Tom baked the bread.
b. baker of bread
c. bread baker

(10c)는 (11)과 같은 구조를 갖는다. (11)에서 'baker'의 목적어인 'Theme'은 'bread'로 <Agent, Theme>의 논항구조와 'i'라는 같은 표지로 표시된다. 다시 말하면 'baker'의 논항 구조는 'bake'와 동일하며 논항 구조가 (11)과 같이 포착된다.

(11)



3) 이러한 논의는 Sproat (1985)과 Roeper (1988)을 거쳐서 사건 구조(Event Structure)와 의미역 일률성 가설(Uniform Theta Assignment Hypothesis)을 구조적으로 설명하면서 발전하여 왔다.

이와 같이 본 연구에서는 내심구조에서 핵-수식 경우가 아닌 경우에 술어-논항 구조를 가정하고 어떤 논항 구조가 실현되는지를 살펴보고자 한다. 이를 위해서 어휘 목록에 대한 논항 구조를 설정한 세종전자사전을 토대로 논항 구조에 따라서 구조적인 설명이 가능한지를 논의할 것이다.

3. 연구를 위한 자료 추출

연구를 위해서 추출한 자료는 형태소 분석이 된 세종코퍼스를 대상으로 [명사1 명사2 명사3]의 어휘 군집을 모두 추출하였다.[4] 모두 266,000여 개의 어휘 연쇄들이 추출되었다. 고유명사의 경우에는 외심구조가 될 가능성이 높으므로, 고유명사표지가 포함된 경우를 모두 제외하였다. 이를 통해서 모두 46,000여 개의 명사어휘 연쇄만이 추출되게 되었다.

이 중 생산성이 너무 낮은 경우는 고려하지 않기 위해서 출현 빈도가 모두 5회 이상이 경우로 제한을 하였으며 모두 1,000여 개가 수집되었다. 이를 토대로 화자 직관을 활용해서 복합명사 구조가 판별이 가능한 경우만을 고려하였다.

화자 직관을 활용하기 위해서 20, 30대 2명의 화자에게 동일한 목록의 [명사1 명사2 명사3] 어휘 군집을 보여주고 어느 괄호매김이 타당한지를 표시하게 하였다. 예를 들어서 [범죄 단체 조직]이라는 복합명사가 있으면, (12)와 같은 자료를 제시하고 이 중 가능한 분석의 경우를 하나만 명기하도록 하였다.

(12) a. ❑ [[범죄 단체] 조직]
b. ❑ [범죄 [단체 조직]]
c. ❑ [범죄 단체 조직]
(세 개의 명사들 사이에 구조가 없음)
d. ❑ 잘 모르겠음

추출된 자료 1,000여 개 중 30% 정도인 300여 개가 동격구조인 (12c)로 판단되었으며, 화자직관으로 판단하기 어려운 경우인 (12d)와 같은 경우도 40% 정도인 390여 개가 발견되었다. 또한 그 외에 310여 개 정도 중 200여 개가 화자들 간에 일치하였으며, 나머지 110여 개는 화자 간에 불일치하였다. 화자간 불일치는 크게 문맥적인 상황을 고려해야 하는 경우, 화자(들)의 실수와 같이 예측이 가능한 경우, 특이한 문제점이 없어서 예측이 되지 않는 경우들이 발견되었다. 문맥적 상황이 필요한 경우도 있는데, (13a, b)에서와 같은 구조적 모호성이 문맥적 상황에 의존되어 있는 경우이다.

4) 이러한 자료 추출 방식을 tri-gram이라고도 한다.

(13) a. [[자동 지급] 금지]
b. [자동 [지급 금지]]

화자 간 불일치 중 문맥적 모호성을 지닌 경우와 특이한 문제가 없이 예측이 되지 않는 경우를 제외하고 자료를 정리해서 270개의 자료를 최종적인 연구 자료로 선택하였다.

화자 직관 실험을 통해서 정확한 자료를 얻을 수 있었지만, 화자 실험과 같이 시간과 노력이 많이 소요된다. 또한 화자간의 불일치가 발견되었는데, 이러한 경우에도 후처리 작업이라는 노력이 필요하다. 언어 직관을 이용하는 것이 연구에 있어서 적절하다고 판단이 되나, 작업으로 인해서 시간과 노력, 재원들이 많이 소모된다. 따라서 이러한 것들을 줄일 수 있는 효용적 방안에 대한 연구가 필요하다. 만약 언어 자료를 활용해서 정확한 자료를 얻어 낼 수 있다면 이러한 시간과 노력을 절약할 수 있을 것이다. 본 연구에서는 연어 측정 방식을 적용해서 화자 직관을 통한 결과와 비교하였다.

4. 연어 측정을 통한 괄호매김 결정

연구의 대상은 [명사1 명사2 명사3]의 세 개 명사의 연속이며, 모두 일반명사들이다. 앞서 논의한 바와 같이 외심구조나 동격 명사구, 등위 합성구조와 같이 구조적 분해가 필요하지 않는 구조는 화자 직관을 활용해서 정련하였으며, 연구에서는 괄호매김의 모호성이 가능한지를 결정해야 하는 자료들을 활용하였다.

연구에서는 통계적 연어 측정 방식을 활용해서 괄호매김을 하는데, (8)과 같이 (14)나 (15) 중 하나로 통계량이 측정된다.

(14) a. [[명사1 명사2] 명사3]
b. (명사1 명사2)의 통계적 연어 측정량이 (명사2 명사3)과의 통계적 연어 측정량보다 더 크다.
(15) a. [명사1 [명사2 명사3]]
b. (명사2 명사3)의 통계적 연어 측정량이 (명사1 명사2)과의 통계적 연어 측정량보다 더 크다.

4.1. 연어 측정을 위한 통계적 기재들

본 연구에서는 여러 연어 측정 방식을 활용하고 서로 비교하였는데, 빈도, PMI, SCP, Dice, log-likelihood, chi-squared, z-test, t-test이다. 각각을 살펴보면 다음과 같다. 우선

빈도는 어떠한 단어의 연쇄 xy의 출현 횟수를 측정하여서 활용한다(Guiliano 1964). 이를 공식화하면 다음과 같다.

(16) $f_{xy} = count(x,y)$

PMI는 어휘들간의 정보량을 측정하기 위해서 제안되었으며, 두 어휘가 각각 독립적으로 출현하는 확률 중 공기하는 확률의 비율에 대한 엔트로피 정보량을 측정한다. 예를 들어서 '금융 세제'의 경우에 '금융'과 '세제'의 각각의 확률들이 독립적으로 출현할 확률 중에서 '금융 세제'가 공기하는 확률들의 비율을 나타내게 된다.

(17) $PMI(x,y) = \log_2 \frac{P(xy)}{P(x) \times P(y)}$ [5]

PMI는 엔트로피 정보량 중 두 어휘의 정보량만을 고려한 것으로서 Chuch and Hanks (1991)에서 사전의 표제어 추출을 위한 통계적 측정 방식으로 제안되었다.

SCP는 PMI와 다르게 두 어휘의 확률 정보에만 근거해서 어휘 결합 강도를 나타낸다.

(18) $SCP(x,y) = P(x|y) \times P(y|x) = \frac{P(xy)^2}{P(x) \times P(y)}$

(18)은 Ferreria et al. (1999)에서 제안된 방식으로 일정한 두 단어의 조건 확률을 모두 고려하고 일정한 출현 공간에서 결합의 강도가 어느 정도인지를 측정한다.

Dice는 Dice(1945)에서 주장된 확률 계산법으로 정보의 양을 계산하기 위한 방식이다. 이 방식은 두 어휘간의 정보량을 측정하기 위해서 제안되었다.

(19) $Dice(x,y) = \frac{2 \times f_{xy}}{f_x + f_y}$

(19)에서와 같이 정보량 계산이 0이 되는 것을 피하기 위해서 분자에 2를 곱하고, 이를 전체 개수로 나눈다. Dice를 활용하면 두 어휘간의 확률적 정보량을 계산하고 이를 정보검색에서 두 키워드간의 유사도를 측정하게 된다(Manning and Schütze 1999).

Log-likelihood 방식은 Dunning (1993)에서 제안된 방식으로 두 단어 연쇄 'xy'가 확

5) $P(x)$는 x의 확률을 말한다.

률적으로 연어로서 의미적으로 유사관계가 있는 통계 가설 H2와 의미적 유사성이 없고 우연한 관계로 구성된 통계 가설 H1의 확률적 비율을 측정하는 것이다. 이러한 의미적 유사성 가설 H1, H2와 xy 단어 연쇄의 확률적 분포는 표 1과 같은 통계적 분포성을 지닌다.

표 1. Log-likelihood의 가설과 확률의 통계적 분포

	H1	H2
$P(y\|x)$	xy 단어 연쇄가 의미적으로 유사하지 않은 경우	xy 단어 연쇄가 의미적으로 유사한 경우
$P(y\|\overline{x})$	x가 아닌 다른 단어와 y의 단어 연쇄가 의미적으로 유사하지 않은 경우	x가 아닌 다른 단어와 y의 단어 연쇄가 의미적으로 유사한 경우

따라서 H2 가설과 $P(y|x)$이 전체 통계 가설 공간에서 유의미하다면 단어 연쇄 xy는 의미 있는 연어가 될 것이다. 그리고, 표 1과 같은 전체 통계 공간의 연산을 (20)의 공식이 포착한다.[6]

$$(20)\quad \log\lambda(x,y) = \frac{H_1}{H_2}$$
$$= -2\times\log\frac{(P(x)\times P(y)\times P(\overline{x})\times P(\overline{y}))^{f_y}}{(P(xy)\times P(\overline{x}\overline{y}))^{f_{xy}}\times(P(x\overline{y})\times P(\overline{x}y)^{f_{\overline{x}y}}}$$

이 연산에서 단어 연쇄 xy가 의미적으로 유사하다면 x나 y가 출현하지 않는 확률 분포인 $\overline{x}$나 $\overline{y}$보다 더 통계적으로 유의미하게 측정될 것이다. 신효필 (2007)에서는 log-likelihood를 활용해서 연어를 측정하였고, 저빈도 어휘일 경우에도 유용하게 활용될 것이라고 주장하였다.

이 방식은 log-likelihood 방식과 마찬가지로 네 개의 서로 다른 의미 공간을 통해서 단어 연쇄 xy가 통계적으로 유의미한지를 검증한다.

6) 여기서는 Schone and Jurafsky (2001)에서 제시한 방식의 연산을 활용한다. Manning and Schütze (1999)에서 활용된 연산과 Schone and Jurafsky (2001)에서 활용한 연산의 방식은 동일하나 표기의 방식에 있어서 차이가 있다.

표 2. Chi-squared의 통계적 분포

	y	$\overline{y}$
x	xy 단어 연쇄의 통계적 분포	x와 y가 아닌 단어 연쇄와의 통계적 분포
$\overline{x}$	x가 아닌 단어와 y의 단어 연쇄의 통계적 분포	x도 아니고 y도 아닌 단어 연쇄의 통계적 분포

여기서 단어 연쇄 xy가 통계적으로 유의미하다면 xy단어 연쇄의 통계적 분포가 유의미해야 한다. 이러한 방식은 일반적인 chi-squared 통계 방식과 동일하며, (21)과 같이 기대 확률(E)에 대한 관찰 확률(O)과 기대 확률(E)의 차이의 비율로 연산된다.

$$(21)\ \chi^2(x,y) = \sum_{\substack{i \in [x,\overline{x}] \\ j \in [y,\overline{y}]}} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

이 방식은 Church and Gale (1991)에서 제시된 방식으로 하나의 코퍼스나 다른 여러 코퍼스에서에서 연어 관계의 성립여부를 측정하기 위해서 활용된다.

z-test는 Smadja (1993)에 의해서 연어를 추출하기 위한 방식으로서, 표본집단과 모집단의 평균의 차이가 통계적으로 얼마나 유의미한지를 측정하는 z-test 방식과 유사하다. (22)와 같이 관찰빈도(O)와 예상빈도(E)가 표준편차(σ)에 비율적으로 얼마나 차이가 나는지를 측정하게 된다(Barnbrook 1996; 강범모 2003).

$$(22)\ z-test(x,y) = \frac{O-E}{\sigma} = \frac{f_{xy} - E_{xy}}{\sqrt{E_{xy} \times (1 - \frac{E_{xy}}{N})}}$$

z-test는 모집단과 표본집단의 분포적 차이에 대한 특성을 측정하는 것이라면 t-test는 두 집단간 차이의 분포적 특성을 (23)와 같이 측정하게 된다.

$$(23)\ t-test(x,y) = \frac{O-E}{\sqrt{E}} = \frac{f_{xy} - E_{xy}}{\sqrt{f_{xy} \times (1 - \frac{f_{xy}}{N})}}$$

Church et al. (1991)에서 논의된 이 방식은 코퍼스에서 발견되는 통계를 활용해서 두 단어 연쇄 *xy*가 통계적으로 유의미한지를 검증하게 된다.

4.2. 실험 및 결과 분석

여러 통계적 연어 측정 방식 중에서 전반적으로 PMI가 가장 화자 직관과 유사하게 나타났다. 이 방식은 상호정보량 측정 방식으로 두 단어의 정보량이 얼마나 많은지를 측정한다. 그림 1은 4.1절에서 논의한 여러 통계적 연어 측정 방식이 화자 직관의 결과와 비교해서 어느 정도 정확한지에 대한 정확률이다.

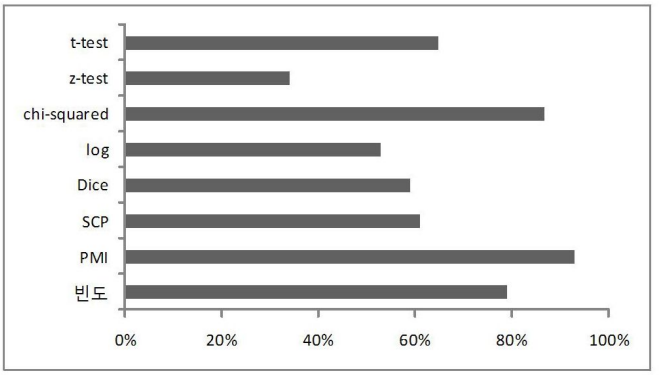


그림 1. 결과분석

본 연구는 통계적 방식을 적용했던 Lauer (1995)의 연구 결과 보다 더 향상된 방식으로 문제를 해결하였다. Lauer (1995)는 로제 시소러스를 활용하였는데, 로제 시소러스는 정련된 언어 사전 자료로서 효용성은 매우 높지만, 구축하는데 있어서 비용과 재원이 많이 소모된다.[7] 그러나 본 연구는 사전이 아닌 형태소 분석이 된 코퍼스만을 사용하였다. 따라서 효용적인 측면에서 로제 시소러스보다 비용과 재원이 더 적게 소모된 자원을 활용하였다는 측면에서 더 향상되었다고 할 수 있다.

실험 결과는 기존 연구인 원형석 외 (2000), 윤보현 외 (1999), Park et al. (1996), Yoon et al. (2001)의 주장과 다르다. 본 연구에서 활용한 연어 정보의 측정 방식은 크게 빈도 정보와 빈도 정보에 근거한 확률 정보를 활용하는 측정법과 통계량을 활용하는 측정법, 정보량을 활용하는 측정법들로 나뉜다. 빈도 정보와 빈도 정보에 근거한 확률 정보를 활용하

7) 영어의 경우에 전자사전인 워드넷이나 시소러스인 로제 시소러스는 무료로 배포된다. 한국어의 경우에 가용할 수 있는 언어 자원은 그 수가 많지 않다.

는 방식은 빈도, SCP, Dice가 있고, 통계량을 활용하는 측정법은 chi-squared, z-test, t-test가 있다.[8] 또한 정보량 측정법은 PMI, log-likelihood가 있다. 기존 연구인 원형석 외 (2000), 윤보현 외 (1999), Park et al. (1996), Yoon et al. (2001)에서는 복합명사 구조를 분석하면서 빈도에 근거한 확률 방식을 활용하였는데, 본 연구와 같이 여러 측정법을 비교해 보지는 않았다. 이와 다르게 본 연구는 여러 측정법을 비교하고, 이 중 PMI가 화자 직관과 가장 유사함을 살펴보았다. 그런데, PMI는 정보량 측정법에 해당한다. 따라서 기존 연구에서 주장한 빈도에 근거한 확률보다는 정보량이 복합명사 구조 분석에서 더 나은 결과를 보인다는 것이 본 연구에서 결과로 나타났다.

5. 핵-수식 및 술어-논항 구조 분석

복합명사의 내부 구조가 내심구조이면 핵-수식, 술어-논항 구조로 분석된다. 핵-수식 구조는 핵어와 수식어로 구성되며 술어-논항 구조는 서술적 성격의 명사와 논항으로 구성된다.

예를 들면, '전문 기술 모임'은 [[전문 기술] 모임]으로 구성되는데, '전문 기술'은 '전문적 기술'과 같이 한자어의 소유격화가 가능한 구조로 핵-수식 구조로 구성되고, 전체 구조는 '전문적 기술의 모임'과 같이 [전문적 기술]의 핵-수식 구조가 다시 수식어가 되어서 '모임'이라는 핵어에 수식어로 연결된 구조이다.

반면에 술어-논항 구조의 예는 '환경 영향 평가'로 [[환경 영향] 평가]로 분석이 되는데, '환경 영향'이 핵-수식 구조이고 다시 '환경 영향'이 '평가'의 논항으로 사용되어서 '[환경 영향]을 평가하다'와 같은 술어-논항 구조 분석이 가능하다.

두 구조는 통사적인 측면에서 다른데, 핵-수식 구조는 동사성 성격의 명사가 아닌 명사들이 수식 관계로 이어져 있는 반면에 술어-논항 구조는 동사성 성격의 명사가 자신의 논항으로 명사나 명사구를 선택하는 구조이다.

연구에서는 모든 명사 어휘들의 괄호매김을 분석하고, 각각의 관계성을 규정하도록 하였다. 전체적으로 (24)와 같은 분석이 가능하다.

(24) a. [[협동 조합] 중앙회]
'협동 조합' -> 외심구조
'협동 조합 중앙회' -> [협동조합]: 수식어,
중앙회: 핵심어

8) t-test는 빈도 통계를 활용한 확률 정보에 근거한다(Church et al. 1991). 반면에 z-test나 chi-squared는 통계량에 근거한 가설 통계에 기초하고 있다(Smadja 1993, Church et al. 1991).

b. [[무역 역조] 시정]
'무역 역조' -> 무역: 수식어, 역조: 핵심어
'무역 역조 시정' -> [무역 역조]: 수식어, 시정: 핵심어
c. [[상품 불매] 운동]
'상품 불매' -> 상품: Theme, 불매: 서술어
'상품 불매 운동' -> [상품 불매]: Theme, 운동: 서술어

270개의 자료가 각각 세 개의 명사로 이루어져 있고 두 개씩 짝지어지므로 전체 관계성은 모두 540개이다. 연구에서는 외심구조, 핵-수식, 술어-논항 구조로 구분하였는데, 전체적으로 구분된 구조를 살펴보면 외심구조는 매우 적고, 대부분이 핵-수식 구조로 분석되었다. 그림 2는 외심구조, 핵-수식, 술어-논항 구조들의 통계적 수치이다.

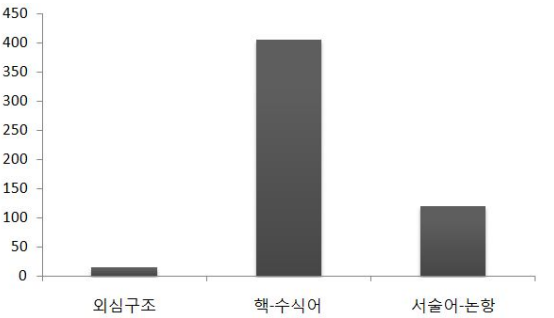


그림 2. 구조분석 결과

5.1. 외심구조 판별

이전 절에서 분석한 것은 구체적인 괄호매김을 통계적 연어정보를 활용해서 선택한 것이지만, 부가적으로 통계적 연어정보가 외심구조가 되는 구조를 분석하는 틀로도 활용될 수 있다. 높은 통계적 수치는 연어적 특성보다는 오히려 고유명사와 같이 하나의 특정 명칭으로 활용되는 것을 보여준다. 예를 들어서 빈도의 경우에 일정 수준의 빈도를 고유 명사화된 어휘 항목의 측정으로 활용할 수 있다. '특별 검사제 도입'에서 [특별 검사제]는 하나의 고유명사인데, 1,000만 어절로 구성된 세종코퍼스에서 빈도를 조사하면 [특별 검사제 도입]의 단어 연쇄는 7번 사용되었는데, [특별 검사제]의 단어 연쇄는 8번 사용되었다. 따라서 [특별 검사제]가 고유명사로 사용될 가능성이 높다고 할 수 있다. 만약 빈도 8회 이상을 고유 명사화된 어휘 항목의 측정 기준으로 활용한다면 [특별 검사제]는 하나의 고유 명사화된 어휘 항목으로 간주될 수도 있다.

이와 같이 통계적 장치는 판단의 근거로 활용할 수 있는데, 빈도는 통계적 검정이 아닌

통계적 기술이므로, 어떤 사실을 입증하거나 증명하지 못한다. 반면에 통계적 검정장치를 사용하면 더 객관화된 증명이 가능하다. 통계적 검정장치는 일정한 유의수준의 확률로 검정해야 할 통계적 사실들을 입증하게 된다. 통계적 검정장치를 활용하는 연어 측정방식은 chi-squared, z-test, t-test등이 있다. 이 방식들을 검정장치로 활용해서 유의수준 (0.005% 또는 0.05% 또는 기타 % 수준으로) 단어 연쇄 A가 고유 명사화된 어휘 항목인지를 입증할 수 있다. 여기서 적정한 수준의 유의수준을 설정하면 가능한데, 유의수준을 너무 높이면 너무 적은 숫자가 추출되고, 유의수준을 너무 낮추면 너무 많은 숫자가 추출되어서 불필요한 대상들도 포함되게 된다. 연구에서는 적정한 0.005% 수준으로 chi-squared, z-test, t-test를 활용해서 측정하였는데, 다음의 고유 명사화된 명사 연쇄를 추출하였다.

(25) 당좌 대월, 하드 디스크, 병목 현상, 신용 카드, 가계 수표, 도시 가스, 고등 교육, 위성 방송, 정기 간행물, 신문 방송, 정보 처리, 고속 도로, 실습 기자재, 정치 자금법, 병목 현상

(25)에 '신문 방송'이 외심구조로 판단되었다. 연구의 대상이 된 세 개의 명사 연쇄는 [신문 방송 편집인]인데, '신문 방송'의 연쇄가 '언론'의 동치의 의미로 사용되었다. [신문 방송]의 빈도는 76회인데, 모두 '신문 방송'으로 사용되었으며, '방송 신문'으로 사용된 경우는 한 번도 없다.

5.2. 핵-수식 구조

외심구조가 아닌 단어 연쇄의 경우 소유격 조사나 어미가 활용이 가능하다. 한자어인 경우에는 한자어 접미사인 '-적(的)'을 사용할 수 있다. 따라서 다음의 예에서 '-적'을 붙인 형태의 도출이 가능하다.

(26) a. [왕립 국제 문제]
[국제 문제] -> '국제적 문제'
b. [전문 기술인 모임]
[전문 기술인] -> '전문적 기술인'

또한 소유격 접미사 '-의'가 가능한 경우도 있다. '경기 안타'의 경우에 '경기의 안타'와 같이 바꾸어 쓸 수 있다.

연구에서는 술어-논항 구조가 아닌 모든 구조를 핵-수식 구조로 판단하였다. 이를 위해서 먼저 핵이 되는 어휘들이 동사성 명사가 가능한지를 살펴보고 아닌 경우를 핵-수식 구조로 고려하였다. 동사성 명사는 '-하다'의 조어법이 가능한 경우에 판정하였다. 예를 들어서 (27)

과 같은 명사는 동사성 명사로 간주하였다.

(27) 집행, 거래, 통치, 도입, 수사, 조사, 시정, 운동, 통제, 개시, 해석, 폐지, 방문, 시험, 처벌, 신고, 조치, 할당, 판매

핵-수식 관계를 나타내는 단어 연쇄의 경우에 해당 연쇄에 소유격화가 가능한지를 세종 코퍼스를 통해서 찾아보았다. (28a)과 같은 명사 연쇄의 소유격이 (28b)와 같이 세종코퍼스에서 실제로 발견되었다.

(28) a. [[주부 살림살이] 노하우]
-> '주부의 살림살이' (주부: 수식어, 살림살이: 핵어)
b. 여러분이 주부의 살림살이를 함께 체험하는 여행입니다.

(28)에서와 같이 핵-수식 구조로 판정할 수 있는 경우에 코퍼스상에서 소유격화된 용례가 출현하기도 하는데, 이러한 용례는 핵-수식 구조로 판정할 수 있는 근거를 제공한다. 그러나 이러한 구조가 모든 경우에 발견되는 것은 아니다. 예를 들어서 [[교내 폭력] 발생률]의 경우에 '교내의 폭력'이라는 구조가 가능하지만 실제적으로 '교내의 폭력'이라는 용례는 코퍼스에서 발견되지 않는다. 그러나 이러한 경우에도 직관적으로 수식관계가 성립하므로 핵-수식 관계로 정의하였다. 코퍼스에서 발견되는 자료보다 소유격 접미사 '-의'나 '-적'의 활용의 생산성이 더 높으므로, 코퍼스에서 발견되지 않는 자료의 경우에도 소유격화가 가능하다. 따라서, 언어 직관으로 핵-수식 관계를 포착하는 것이 분석에서 더 타당하므로, 가능한 모든 경우를 포괄하도록 하였다.

5.3. 술어-논항 구조

술어-논항의 경우는 '-하다'와 결합이 가능한 명사를 대상으로 하였다. 이 명사들은 동사적 성격이 있어서 논항 구조가 된다. (29a)와 같은 문장은 (29b)와 같은 명사구로 전환이 가능하며, 조사를 생략하면 (29c)와 같은 명사구 연쇄가 가능하다.

(29) a. 검찰이 참고인을 조사하였다.
b. 검찰의 참고인 조사[9)]
c. 검찰 참고인 조사

9) 이 구조는 '검찰이 참고인을 조사하다.'와 'X-가 검찰의 참고인을 조사하다.'의 의미에서 모호성을 갖는다. 여기에서는 (29a)의 해석에 따라서 분석하였다.

세종전자사전을 살펴보면 '조사하다'의 논항구조는 (30a)와 같고, 각 논항이 추하는 격조사는 (30b), 선택제약(selectional restriction)은 (30c)와 같다.

(30) a. <Agent, Theme>
b. Agent-이 Theme-을 (Theme-에 대해)
c. Agent = 인간 또는 인간집단
Theme = 전체

(29a, b, c)를 활용하면 (30)은 다음과 같이 분석된다. '검찰'은 Agent, '참고인'은 Theme이 된다. 여기서 주목할 점은 한국어에서 Agent의 역할을 하는 항목이 (29b)와 같이 명사구에서 소유격조사인 '-의'가 부착될 수 있다.

명사구 연쇄에서 술어성 명사의 전체 논항들이 나타난 경우는 많지 않다. 자료를 통해서 얻은 자료는 '주부 살림살이 체험'이나 '검찰 수사 착수'와 같이 신문기사의 헤드라인과 같이 축약된 구조들이다. 연구의 대상이 되는 자료에서는 Agent가 생략된 구조들이 발견되었다.

전자사전은 그림 3과 같은 구조로 작성되어 있는데, 선택제약을 발견할 수 있다. '거래하다'의 경우에 Agent(인간), Theme(구체물,....), Comitative(인간,...)가 논항으로 사용된다.

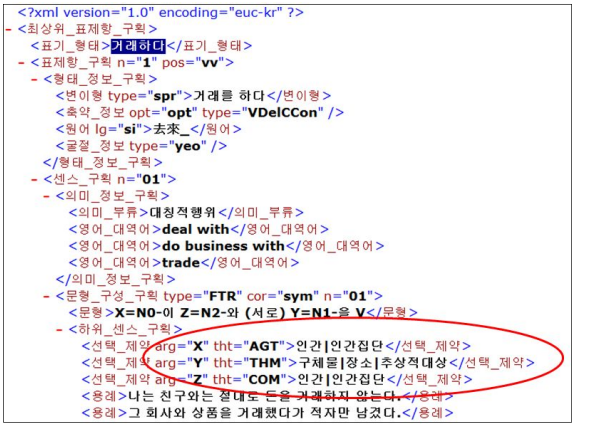


그림 3. 세종전자사전의 예

논항구조와 연관되어서 논항이 2개 이상인 경우에는 해당 논항이 어디에 해당하는지를 찾아내야 한다. '거래'의 경우는 세 개의 논항을 취하므로, 어떠한 논항인지를 구분할 필요성이 있다. 예를 들어서, [[당좌 대월] 거래] 복합구조의 경우 '당좌 대월'이 Agent인지, Theme인지, Comitative인지 구분해야 한다. 이러한 경우에 해당 논항이 어떠한 논항에 해

당하는지 구분해야 한다.

세종전자사전의 경우에는 각 명사들에 대한 의미부류 체계인 '대상부류'라는 작은 온톨로지 체계를 갖고 있다. 최상위부류를 구체물, 집단, 장소, 사태로 나누고 그 아래 580여 개의 작은 의미들로 구분하였다. 예를 들어서 '인간'은 '구체물->구체자연물->생물->인간'으로 구분되고 '인간' 아래에 '관계적 인간, 화시적 인간, 호칭'등등의 여러 의미부류를 구분하였다. 그림 4는 세종전자사전의 의미부류 체계를 보여준다.

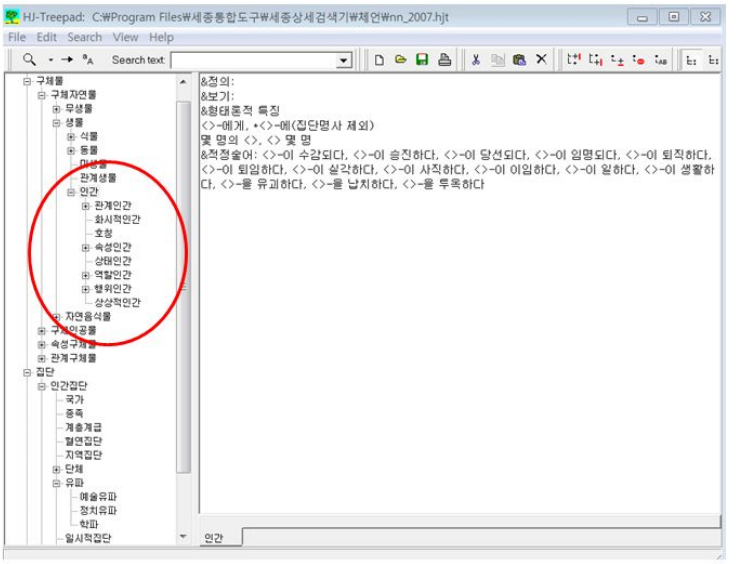


그림 4. 세종전자사전의 의미부류 체계

그림 4에서 제시한 바와 같은 의미부류 체계를 활용하면 [[당좌 대월] 거래]에서 [당좌 대월]은 '구체물'의 의미부류에 해당하므로 Theme으로 분류할 수 있다.

이러한 절차에 따라서 연구에서는 세종전자사전에 등록된 '-하다' 어휘의 경우를 서술성 명사로 간주하고, 논항구조를 찾아서 어떠한 논항으로 구성되어 있는지를 찾아보았다. 하나의 예를 통해서 자세히 설명하면 다음과 같다. '전면 전쟁 개시'와 같은 구조는 (31)의 구조로 분석이 된다.

(31) a. [[전면 전쟁] 개시]
b. '전면 전쟁' -> 전면: 수식어, 전쟁: 핵심어
c. '전면 전쟁 개시' -> [전면 전쟁]: Theme, 개시: 서술어

'개시'는 서술성 명사로 세종전자사전에 표시된 논항구조와 선택제약은 다음과 같다.

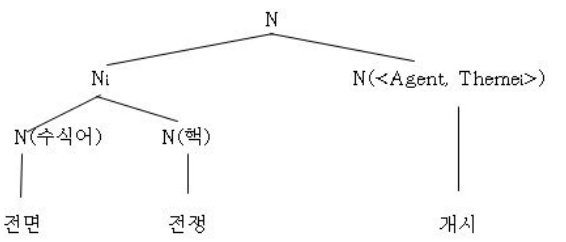
(32) a. <Agent, Theme>
b. Agent-이, Theme-을
c. Agent = 인간|인간집단
Theme = 행위(공격|작전|행동|작업)

'전면'은 '-적'이 붙어서 수식어가 될 수 있으므로, '전쟁'의 수식어 역할을 한다. 핵인 '전쟁'은 세종전자사전에 의미부류가 '물리적 충돌'로 표기되어 있으므로 '물리적 충돌'이 '행위'와 어떤 연관성이 있는지 검색해야 한다. 세종전자사전의 의미부류체계는 (33)과 같다.

(33) 구체물 ->
사태 ->
행위 ->
대칭적 행위 ->
물리적 충돌 행위

'전쟁'의 의미부류인 '물리적 충돌'은 '구체물' 아래의 '행위' 의미부류체계 아래에 '물리적 충돌 행위'에 위치하므로 '개시'의 논항이 된다. 따라서 (34)와 같이 논항구조가 완성이 된다.

(34)



6. 결론

본 연구는 언어 자료를 활용해서 복합명사 구조를 분석하였다. 1,000만 어절로 구성된 세종코퍼스와 세종전자사전을 활용해서 분석을 시도하였다. 분석은 크게 외심구조와 내심구조

로 구분해서, 내심구조의 경우에 분석을 계속한다. 외심구조의 경우에는 동격합성어나 병렬구조로 구조를 찾을 수 없는 경우로 하나의 고유 명사화된 어구로 처리한다. 내심구조의 경우에 괄호매김을 하여서 구조를 분석하는데, 세종코퍼스를 사용해서 통계적 연어측정을 하였다. 측정한 결과 통계적으로 유사성이 높은 이분지 구조로 분석한다. 내심구조는 다시 핵-수식 관계, 술어-논항 관계로 구분하는데, 술어-논항 구조가 아닌 모든 경우는 핵-수식 관계로 분류하는데, 소유격 조사인 '-의'와 한자어 수식어인 '-적'이 결합 가능한 경우를 코퍼스상에서 찾아보았다. 핵-수식의 경우에 소유격화의 생산성이 높아서 코퍼스에서 발견되지 않는 경우에도 직관을 적용해서 분석하였다. 술어-논항 구조의 경우에 세종전자사전을 활용해서 해당 논항구조에 따라서 구조를 분석하였다. 분석의 단계에서 선택제약을 활용하였으며, 의미부류에 따라서 논항을 부여하였다.

복합명사 구조 분석은 언어학적, 정보처리, 음성/음운 처리 부분에서 중요한데, 생산성이 높아서 언어학적 분석이 필요하다. 또한 연구는 관련된 여러 분야에 활용도가 높다.

현재 언어 자료는 세 개의 명사연쇄 구조분석에 집중하였는데 구조적 차이점을 분명히 하고, 이 논문을 필자들 읽는 독자들의 이해를 돕기 위함이었다. 그런데, 언어 자료는 이 외에도 많은 무수한 명사연쇄가 가능하다. 본 연구자도 향후 다양한 연쇄 구조도 분석할 것이며, 더 현실적 언어 자료를 이해하고 분석하기 위해서 매진할 것이다.

참고문헌

강범모. (2003) *언어, 컴퓨터, 코퍼스 언어학.* 고려대학교 출판부.

남기순 · 최기선. (1997) 검색 엔진의 '색인 모듈'의 문제와 합성어 사전 및 구문 정보 사전의 필요성. *한국정보관리학회 제4회 학술대회 논문집*, 5-15.

신효필. (2007) 연어의 통계적 접근을 통한 로그 우도비 중심의 연어 검증. *언어학*, 47, 107-138.

윤보현 · 조민정 · 임해창. (1997). 통계 정보와 선호 규칙을 이용한 한국어 복합명사의 분해. *정보과학회논문지*, 24, 900-909.

원형석 · 박미화 · 이근배.(2000) 복합명사 분할과 명사구 합성을 이용한 통합 색인 기법. *정보과학회논문지*, 27, 84-94.

Aronoff, M. (1976) *Word formation in generative grammar.* MIT Press.

Barnbrook G. (1996) *Language and computers.* Edinburg University Press.

Church, K., and Gale, W. (1991). Concordances for parallel text. In *Proceedings of*

the 7th Annual Conference of the UW Center for ITE New OED & Text Research, 40-62.

Church, K. and Hanks, P. (1991) Word association norms, mutual information and lexicography. *Computational Linguistics, 16*, 22-29.

Church, K, Gale, W., Hanks, P., and Hindle, D. (1991) Parsing, word associations and typical predicate-argument relations, In Tomita, M. (ed.) *Current issues in parsing technology*(pp. 103-111). Kluwer Academic Publishers.

Dice, L. (1945) Measures of the amount of ecologic associations between species. *Journal of Ecology, 26*, 297-302.

Di Scullo, A. and Williams, E. (1987) *On the definition of word.* MIT Press.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*, 61-74.

Ferreira, S., Pereira J., and Lopes, G. (1999) A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. *Sixth Meeting on Mathematics of Language*, 369-381.

Giuliano, V. E. (1964) The interpretation of word associations. In M.E. Stevens et al. (Eds.) *Statistical association methods for mechanized documentation* (pp. 25-32) National Bureau of Standards Miscellaneous Publication 269, Dec. 15, 1965.

Lauer, M. (1995) Corpus statistics meet the compound noun: Some empirical results. In *Proceedings of ACL-1995*, 47-54.

Levi, J. (1978) *The syntax and semantics of complex nominals.* New York: Academic Press.

Liberman, M. and Sproat, R. (1992) The stress and structure of modified noun phrases in English. In I. Sag (Ed.), *Lexical matters*(pp. 131-181). CSLI Publications, University of Chicago Press.

Lieber, R. (1983) Argument linking and compounding in English. *Linguistic Inquiry, 14*. 251-286.

Manning, C. and Schütze, H. (1999) *Foundations of statistical language processing.* MIT Press.

Park, H., Han, Y., Lee, K., and Choi, K. (1996) A probabilistic approach to compound noun indexing in Korean texts. In *Proceedings of the 16th conference on computational linguistics*, 514-518.

Roeper, T. (1988) Compound syntax and head movement. *Yearbook of Morphology,*

1, 187-228.

Scalise, S. (1984) *Generative morphology*. Dordrecht: Foris.

Selkirk, E. (1982) *The syntax of words*. MIT Press.

Spencer, A. (1993) *Mophology*. Blackwell Publishing.

Sproat, R. (1985) *On deriving the lexicon*. Ph.D. Dissertation. MIT.

Schone, P. and Jurafsky, D. (2001) Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of Empirical Methods in Natural Language Processing*, 100-108.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics, 19*, 143-177.

Yoon, J., Choi, K., and Song, M. (2001) A corpus-based approach for Korean nominal compound analysis based on linguistic and statistical information. *Natural Language Engineering, 7*, 251-270.

김동성

136-701 서울시 성북구 안암동 5가

고려대학교 언어정보연구소

전화: 02-921-4376

이메일: dsk202@korea.ac.kr

Received on 15 July, 2011

Revised version received on 7 September, 2011

Accepted on 7 September, 2011