The Journal of Studies in Language

The Journal of Studies in Language 37(3), 291-305 (2021)

로지스틱회귀모형을 통한 언어적 요인들의 유기적 상호작용 분석: 조건추론나무모형과의 비교

신근영

전남대학교

An Integrated Interaction of Multiple Linguistic Factors Logistic Regression Models: Comparison with Tree Models

Shin, Keun Young

Chonnam National University

ABSTRACT

The Journal of Studies in Language 37.3, 291-305. Both tree models and logistic regression models are widely used to analyze multifactorial data in recent corpus studies. Using my previous corpus study on relative clauses, this paper argues that tree models have difficulties dealing with the integrated effect of multiple linguistic factors, that is, a three-way interaction of non-syntactic factors that affect the preference of relative clause types. The integrated interaction effect cannot be captured by adding interaction terms in a logistic regression model but by suppressing an intercept and creating a single variable that is the combination of all three factors. A mixed-effects logistic regression analysis is ultimately implemented by adding the random effect of register, which has been ignored in the corpus linguistics literature on relative clauses. **(Chonnam National University)**

Keywords: corpus, interaction, logistic regression, multi-factorial data, tree model

OPEN ACCESS

https://doi.org/10.18627/jslg.37.3.202111.291

pISSN : 1225-4770
eISSN : 2671-6151

Received: October 11, 2021

Revised: November 07, 2021

Accepted: November 15, 2021

This is an Open-Access article distributed under the terms of the Creative Commons Attribution NonCommercial License which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright©2021 the Modern Linguistic Society of Korea

본인이 투고한 논문은 다른 학술지에 게재된 적이 없으며 타인의 논문을 표절하지 않았음을 서약합니다. 추후 중복게재 혹은 표절된 것으로 밝혀질 시에는 논문게재 취소와 일정 기간 논문 제출의 제한 조치를 받게 됨을 인지하고 있습니다.

1. 서 론

최근 코퍼스 언어학 분야의 통계분석은 단순한 기술통계(descriptive statistics)에서 벗어나 데이터 마이닝(data mining)과 기계 학습(machine learning)에서 널리 사용되는 나무모형(tree model)과 로지스틱회귀모형(logistic regression model)을 통해 이루어지고 있다. Eddington(2010)은 의사결정나무(decision tree)모형이 범주형 독립변수(categorical independent variables)들 간의 상호작용(interaction)을 더 효과적으로 분석할 수 있지만 로지스틱회귀모형은 더 정교한 통계 정보를

제공할 수 있기 때문에 코퍼스 연구에서 두 모델을 상호 보완적으로 사용할 필요가 있다고 하였다. 반면 Gries(2020)는 코퍼스의 다요인 데이터(multifactorial data)를 다룸에 있어 나무 모형 분석은 방법론적 문제가 있다고 주장하였다.[1] Gries가 지적한 문제들은 나무 모형의 설계 및 나무 모형의 부분 대칭에 대한 원인 분석을 통해 해결될 수 있지만(Shin, 2020a), 나무 모형의 장점이라고 알려진 이해하기 쉽고 간결한 분석 결과라는 것이 연구자의 재해석을 요구하는 복잡한 결과일 수 있음을 보여주었다.

나무모형과 로지스틱회귀분석은 코퍼스 연구 중에서도 특히 같은 의미를 가지며 서로 대체가 가능한 문법적 변이형들의 선택에 영향을 주는 요인들을 찾는 데 사용된다. 하지만 이들 변이형 연구들은 언어적 요인들 간에 나타날 수 있는 유기적 관계에 대해서는 크게 주목하지 않고 있다. 예를 들어 Jenset et al.(2018)는 give, send 등의 이중타동사(ditransitive verb)들이 목적어를 취하는 두 개의 변이형인 [V-NP-NP] 패턴과 [V-NP-PP] 패턴의 선택에 영향을 주는 요소들을 알아보기 위해 로지스틱회귀모형을 사용하였다.[2] 그 분석 결과에 따르면 변이형의 선택이 수여자(recipient)와 대상(theme)이라는 두 의미역이 대명사인지 여부, 이들 논항의 길이, 그리고 수여자 의미역이 사람 등의 생물 객체인지 여부에 영향을 받았다. 하지만 대명사의 길이가 일반 명사에 비해 그 길이가 짧고 사람을 지칭하는 경우가 많다는 점에서 이들 언어적 요인들 간에 생길 수 있는 상호작용에 대해서 고려해야 한다. 그럼에도 불구하고 Jenset et al.(2018)는 이에 대한 거의 논의하지 않았다. 또한 목적어들의 길이가 [V-NP-NP]와 [V-NP-PP] 선택에 영향을 준다는 주장은 목적어들의 상대적인 길이를 비교해야 가능하지만 이 또한 간과되었다.

본 논문은 이처럼 언어구조의 다층적 체계(multiple levels of linguistic structure)로 인해 발생하는 언어적 요인들 간의 상호작용이 나무모형과 로지스틱회귀분석에서 어떻게 분석될 수 있는지 살펴보고자 한다. 필자의 관계사절 유형 분포에 관한 연구(Shin, 2020b)에서 사용되었던 코퍼스 데이터를 활용하여, 구문의 통사적, 의미적, 화용론적 구조가 서로 유기적으로 연결되어 있기 때문에 발생하는 언어적 요인들의 통합적 효과를 설명할 수 있는 모형을 논의하고자 한다. 이중타동사의 패턴들과는 달리 주격 관계사절과 목적격 관계사절은 서로 교체가 가능한 변이형 구문이 아니다. 하지만 필자의 선행연구(Shin, 2020b)를 포함한 여러 코퍼스 연구들은 이 두 관계사절은 선행사의 유생성(head noun animacy), 종속 명사의 유생성(embedded noun animacy)과 종속명사구의 주제성(topicality)에 따라 매우 다른 빈도 분포를 보인다는 것을 보여주었다. 더욱이 조건추론나무(conditional inference tree)를 사용한 분석에 따르며 두 관계사절 사용에 유의미한 영향을 미치는 이 세 요인들 간의 상호작용이 있다(Shin, 2020b). 하지만 코퍼스 연구에서 보편적으로 사용되는 조건추론나무모형이나 상호작용항(interaction term)으로 상호작용 효과를 분석하는 회귀모형으로는 이러한 3차 상호작용(3 way-interaction)이 관계사절의 타동성(transitivity)을 반영하는 통사적, 의미적, 화용론적 구조의 유기성을 나타낸다는 해석을 도출해내기 어렵다. 본 논문에서는 이러한 문제에 대해 논의하고, 상수항(intercept)을 제거하고 단일 설명변수를 사용하는 로지스틱회귀분석을 통해 언어적 요인들 간의 통합적 상호작용 효과를 설명할 수 있음을 보여주고자 한다. 이 과정에서 범주형 코퍼스 자료를 로지스틱회귀모형으로 분석할 때 나타날 수 있는 준-완전분리(quasi-complete separation)의 문제에 대해 논의할 것이다. 또한 혼합모형을 이용하여 기존의 관계사절 빈도 연구에서 고려되지 않았던 비문법적 요소의 임의효과의 중요성에 대해 고찰하고자 한다.

1) 필자의 논문(Shin, 2020a)에서 Gries 주장의 문제에 대해서 자세히 논의하였다.
2) Jenset et al.(2018)은 그들의 코퍼스 자료가 일부 이중타동사만을 대상으로 한 자료이기 때문에 실제 논문에서는 동사어휘의 임의효과를 고려한 혼합모형을 사용하였다.

본 논문의 통계 분석은 오픈소스 프로그램인 R (R Core Team, 2021)을 사용하여 수행되었다. 기본적으로 조건추론나무모형은 partykit 패키지의 ctree 기능을 사용하였으며, 로지스틱 회귀 분석 및 혼합모형은 R의 기본함수인 glm와 lme4 패키지를 사용하였다.

2. 조건추론나무분석

2.1 관계사절 유형 분포에 관한 코퍼스 연구

본 논문에 실제 사용될 코퍼스 자료는 필자의 선행 연구(Shin, 2020b)에서 사용되었던 자료로 관계대명사 who, which 그리고 that으로 시작하는 주격 관계사절과 목적격 관계사절을 축출하고 분석한 자료이다. 본 자료는 실험 연구에서 나타난 주격 관계사절과 목적격 관계사절의 읽기 속도의 차이를 언어적 경험으로 설명할 수 있는지를 살펴보기 위한 자료였다. 최근까지도 선행사가 가지는 관계사절 내의 통사적 기능이 관계사절의 처리 속도 및 이해력에 영향을 주기 때문에 주격 관계사절이 목적격 관계사절보다 더 빠르게 읽힌다는 주장이 정설처럼 받아들여졌다(Gordon et al., 2001, 2004; Just and Carpenter, 1992; King and Just, 1991 등). 주격 관계사절과는 달리 목적격 관계사절에서는 관계사절 내에서 명사구가 동사보다 먼저 나온다. 점진적 언어처리 방식에 따르면 목적격 관계사절의 경우 선행사뿐만 아니라 관계사절 내의 종속 명사구(embedded NP)도 동사가 나올 때까지 그 통사적·의미적 역할이 정해지지 않아 우리의 작업 기억(working memory)에 부담을 가중시킬 수 있다. 따라서 통사적 구조 차이로 인해 주격 관계사절과 목적격 관계사절의 읽기 속도에 차이를 보인다고 주장하였다.

(1) a. The reporter that attacked the senator admitted the error.
b. The reporter that the senator attacked admitted the error.

하지만 이러한 주장의 근거는 (1)과 같이 선행사와 종속 명사(embedded noun)가 모두 사람을 지칭하는 일반 명사인 경우에 한하여 실행한 실험들의 결과들이었다. 선행사가 무생물 객체(inanimate entity)를 지칭하는 경우 목적격 관계사절의 읽기 속도가 주격 관계사절에 비해 결코 느리지 않으며, 관계사절 내의 명사구가 주제성(topicality)이 강한 대명사인 경우에는 목적격 관계사절이 더 빠르게 처리된다는 것을 최근 여러 실험 연구들 통해 밝혀졌다(Gennari and MacDonald, 2008, 2009; Roland et al., 2012; Reali and Christiansen, 2007; Traxler et al., 2002, 2005 등). 이러한 비통사적 요소들의 영향력은 화자들의 언어적 경험 때문이라는 주장이 나오면서 관계사절에 관한 코퍼스 연구가 활발히 이루어지고 있다. 즉, 우리가 더 자주 접하여 익숙한 구조의 구문을 다른 변이형 또는 다른 유사한 구조의 구문 보다 더 빨리 처리한다는 것이다. 여러 코퍼스 연구들은 비록 주격 관계사절이 목적격 관계사절보다 많이 사용되지만, 선행사가 무생물이고 종속 명사구가 대명사일 때는 목적격 관계사절이 더 많이 사용된다는 것을 보여주었다(Roland et al. 2007; Reali and Christiansen, 2007; Roland et al., 2012 등). 이는 화자들이 무생물의 선행사를 수식하는 대명사를 내포한 목적격 관계사절에 자주 노출된다는 것을 의미하며, 이러한 언어적 경험이 관계사절 처리 및 이해 속도에 영향을 준다고 주장되었다. 하지만 기존의 실험 연구나 코퍼스 연구들은 대부분 이러한 선행사의 유생성과 종속 명사구의 주제성의 영향력을 각

각 별개로 다루고 있으며 이들 사이에 나타날 수 있는 상호작용에 대해서는 고려하지 않았다.

본 논문에서 사용될 필자의 선행 코퍼스 연구(Shin, 2020b)의 자료 역시 실험연구 결과들을 언어적 경험으로 설명할 수 있는지를 살펴보기 위한 것으로 타동사를 포함하고 목적어가 명사구인 주격 관계사절과 목적격 관계사절의 빈도를 비교한 것이다. 선행연구(Shin, 2019)에서 선행사의 유생성과 종속 명사구의 주제성뿐만 아니라 종속 명사의 유생성 역시 관계사절 유형의 빈도 차이를 설명하는 유의미한 요인임을 밝혔기 때문에 필자의 2020년 코퍼스 연구는 이들 세 개의 요인 들 간의 상호 작용이 관계사절의 분포뿐만 아니라 처리 속도의 차이를 체계적으로 설명할 수 있는지를 살펴보고자 했다. 하지만 상호작용을 찾기 위해 필자가 사용하였던 조건추론나무 분석으로는 언어학적 요소들의 유기적인 상호작용의 원인을 명확히 보여주기 어렵다. 이 장에서는 이에 대한 논의를 하고자 한다.

2.2 나무분석의 문제점

우선 사용된 관계사절 코퍼스의 정보를 살펴보자. <표 1>과 같이 코퍼스 정보를 포함한 관계사절 유형, 그리고 명사의 유생성 및 주제성에 대한 정보를 포함한다. 본 자료는 5개의 코퍼스에서 축출한 자료로 구어체(Spoken) 코퍼스 3개와 문어체(Written) 코퍼스 2개로 이루어져 있다. 편의상 구어체는 S로 문어체는 W로 표기하고 각 레지스터에 번호를 달아 구분하였다. RC.type은 관계사절의 유형을 나타내는 변수로 주격 관계사절인 SRC와 목적격 관계사절인 ORC라는 두 개의 수준(level)을 갖는다. Head.noun과 Embedded.noun 변수는 각각의 관계사절에서 선행사와 종속 명사구가 지칭하는 대상이 생물(animate entity)인지 여부에 따라 ‘animate’과 ‘inanimate’를 나타내는 A 또는 IA의 수준을 갖는다. Topicality 변수는 종속 명사구의 주제성을 반영한 변수로 주제성이 강한 대명사나 고유명사는 Topic으로 그 외의 주제성이 약한 명사구는 Non-topic으로 분류하였다.[3)]

표 1. 관계사절 자료의 구조

Mode	Register	RC.type	Head.noun	Embedded.noun	Topicality
Spoken	S1	SRC	A	IA	Non-topic
Spoken	S2	ORC	IA	A	Topic
Spoken	S3	SRC	A	A	Topic
Written	W1	SRC	IA	IA	Non-topic
Written	W2	ORC	A	A	Topic

총 1,068개의 주격 관계사절과 853개의 목적격 관계사절을 분석하였으며 , 카이제곱 검정 결과 관계사절 유형의 사용은 세 언어적 요인에 각각 유의미한 영향을 받는다. <그림 1>은 필자의 선행 연구(Shin, 2020b)에서 수행된 조건추론나무모델을 통해 분석한 결과로 세 요인들 간의 상호작용이 있음을 알 수 있다.[4)]

3) 각 설명변수의 분류 기준 등 본 자료에 대한 자세한 설명은 필자의 선행 연구(Shin, 2020b)를 참조하기 바란다.
4) 조건추론나무모형은 별도의 가지치기가 필요치 않으며 rpart()기능을 이용한 의사결정나무모형의 과적합(overfitting) 문제 등을 야기하지 않는다.

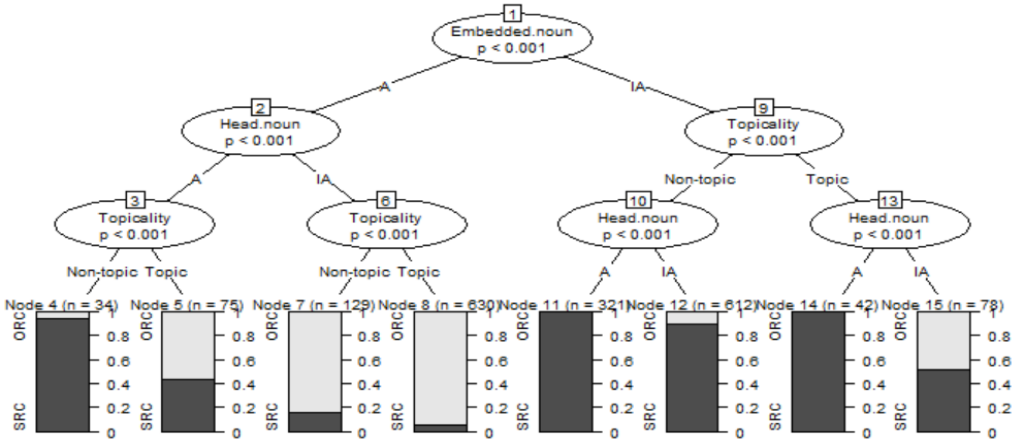

그림 1. 세 설명변수(종속명사의 유생성, 선행사의 유생성, 주제성)를 사용한 조건추론나무분석 결과

<그림 1>의 나무구조가 시작되는 뿌리마디(Root Node)가 종속 명사의 유생성을 나타내는 Embedded.noun으로 두 관계사절의 분류에 가장 큰 영향을 미치는 요인이다. 종속 명사가 생물이면 선행사의 유생성인 Head.noun이 그 다음으로 관계사절 분류에 영향을 주며 주제성이 가장 적은 영향을 미친다. 반면 종속 명사가 무생물이면 주제성이 선행사의 유생성보다 관계사절 유형 선택에 더 큰 영향을 준다.

하지만 필자의 선행 연구들(Shin, 2019, 2020b)에서도 논의하였듯이 설명 변수들은 유기적으로 연결되어 있다. 종속 명사구의 유생성과 주제성은 매우 밀접한 관계가 있다. 관계사절 내의 종속 명사구가 사람과 같은 생물이며 대부분 주제성이 강한 대명사나 고유명사로 나타나는 반면, 종속 명사구가 무생물이면 주제성이 상대적으로 낮은 일반 명사를 핵으로 하는 명사구로 나타났다. 따라서 주제성이 강한 명사구와 자주 쓰이는 목적격 관계사절인 경우 종속 명사가 대부분 생물이며, 주제성이 약한 명사구와 자주 쓰이는 주격 관계사절은 종속 명사가 대부분 무생물이였다. 아래의 <그림 2>처럼 74.8%의 목적격 관계사절은 생물이며 주제적인 명사구(A_Topic)를 논항으로 내포하고 있는 반면, 82.2%의 주격 관계사절은 무생물이며 비주제적인 명사구(IA_Non-topic)를 내포하고 있다.

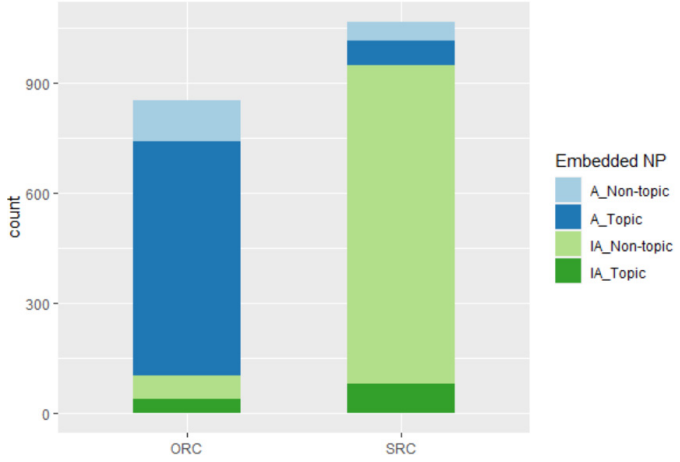

그림 2. 주격 관계사절(SRC)과 목적격 관계사절 내의 종속명사구의 분포

이러한 종속 명사구의 의미·화용론적 특징은 관계사절 내의 종속 명사구의 통사적 역할과 관련이 있다고 할 수 있다. 목적격 관계사절의 경우 종속 명사구가 주어의 역할을 하는 반면, 주격 관계사절 내에서는 명사구가 목적어의 역할을 한다. 여러 학자들의 연구에 따르면 주어와 목적어라는 통사적 기능은 언어 보편적으로 비통사적(non-syntactic) 자질인 유생성과 주제성과 연관되어 있다(Aissen, 1999, 2003; Comrie, 1989; Silverstein, 1976 등). 주어는 생물 특히 사람이면서 주제성이 강한 대상을 선호하고 반대로 목적어는 무생물이며 주제성이 약한 대상을 선호한다는 것이다. 따라서 종속 명사구의 유생성과 주제성은 주어성(subjecthood) 또는 목적어성(objecthood)을 나타내는 새로운 설명 변수로 통합될 수 있다. 하지만 만약 이 통합 변수를 종속 명사구(Embedded NP)로 설정하여 조건추론나무모형으로 분석을 할 경우 아래와 같은 결과가 나온다.

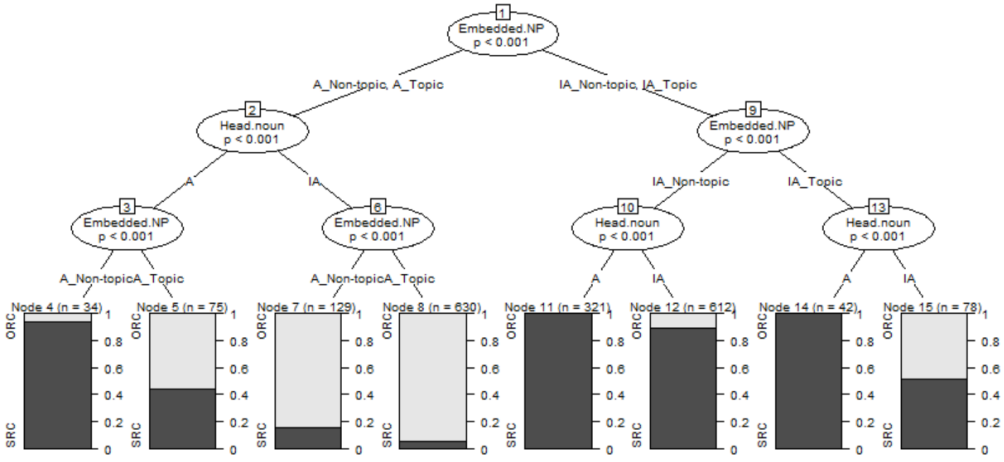

그림 3. 종속명사구의 자질을 통합한 설명 변수(Embedded.NP)를 사용한 조건추론나무 분석의 결과

<그림 1>과 <그림 3>의 끝마디를 비교해보면 두 결과가 같다는 것을 알 수 있다. 다시 말해 <그림 3>은 종속 명사구의 통합적 영향력을 파악할 수 있는 분석 결과를 주지 못하고 있다. 뿌리노드의 Embedded NP가 좌우 양측에서 그리고 다른 마디 위치에서 다시 분리됨으로써 오히려 <그림 1>에 비해 해석이 용이하지 않다. 이러한 결과의 주요 원인은 나무 모형이 가지고 있는 구조의 단순성 및 경직성 때문이다. 나무 모형은 한 번의 분기 때마다 좌우 두 개의 가지로만 분할하기 때문에 Embedded.NP처럼 이산형 변수가 아니라 네 개의 수준을 갖는 변수로 그 효과를 분석하는 데 효율적이지 않다.

종속 명사구의 자질들은 선행사의 유생성과도 밀접한 관계를 보인다. <그림 1>의 끝마디 7, 8, 11, 14를 비교해보면 유생성 배열이 A-IA(생물 선행사 − 무생물 종속 명사)로 이루어진 경우 상대한 수의 주격 관계사절이 쓰이지만 목적격 관계사절은 전혀 나오지 않는다. 이와 대조적으로 IA-A(무생물 선행사 − 유생물 종속명사)인 경우 목적격 관계사절이 압도적으로 많이 나온다. 이러한 유생성 배열의 영향력 역시 일반적인 주어와 목적어의 유생성 선호도로 설명될 수 있다. 주어가 생물이고 목적어가 무생물인 경우는 주격 관계사절과 목적격 관계사절이 각각 A-IA 배열과 IA-A 배열로 나타나기 때문에 이들 유생성 배열에서 두 관계사절 유형이 대조적인 분포를 보이는 것 역시 주어성과 목적어성의 영향으로 설명되어져야 한다. 한 가지 더 주목해야 할 사실은 A-A 배열과 IA-IA 배열의 경우 주제성에 따라 관계사절 유형의 상대 빈도가 매우 달라진다는 것이다.[5] 선행사와 종속 명사의 유생성이 같은 경우는 유생성 배열 자체로 주어와 목적어 기능을 반영하기 어렵다. 이 경우 관계서절 유형의 상대적 빈도가 종속 명사구의 문법적 기능을 반영하는 주제성에 더 크게 영향

을 받기 때문이라고 할 수 있다. 하지만 <그림 1>의 대칭구조를 보면 두 번째 상위노드가 양쪽이 서로 달라 이러한 선행사와 종속 명사의 유생성의 상호작용의 효과를 쉽게 도출하기 어렵고 최종 노드를 기반으로 그 유기적 관계를 재해석해야만 한다. 더욱이 나무모형은 각 설명변수가 독립된 변수임을 전제로 하기 때문에 언어적 변수들이 서로 유기적인 관련성을 보여주기 어렵다. 선행사의 유생성과 종속 명사의 유생성을 통합하는 하나의 변수로 처리하는 방법을 대안으로 제시할 수 있지만 이 변수의 수준이 네 개가 되어 두 개의 가지로 분리하는 나무모형에서는 위에서 언급한 종속 명사구(Embedded.NP) 변수를 사용했을 때와 같은 문제에 직면하게 된다. 이처럼 나무모형의 경우 관계사절의 빈도분포에 영향을 주는 언어적 요인들이 유기적으로 연결되어 통합적인 상호작용을 보인다는 사실을 보여주기가 어렵다.

3. 로지스틱회귀분석

3.1 단순로지스틱회귀분석

우선 R의 로지스틱회귀분석을 통해 각 설명변수가 주격관계사절과 목적격관계사절 사용에 어떤 영향을 미치고 있는지 알아보자.[6] 아래의 단순 로지스틱 회귀분석 결과에 따르면 세 개의 각 요인이 모두 주격 관계사절과 목적격 관계사절의 사용에 유의미하게 영향을 미치는 것을 알 수 있다.

표 2. 단순 로지스틱회귀분석의 결과[7]

	Estimates	S.E.	Z value	OR	95% CI		p-value
					low	high	
(Intercept)	1.847	0.249	7.42	6.341	3.926	10.424	0.00
Head.noun (IA)	-3.226	0.242	-13.307	0.040	0.024	0.063	0.00
Embedded.noun (IA)	3.462	0.188	18.443	31.891	22.234	46.452	0.00
Topicality (Topic)	-1.797	0.184	-9.791	0.166	0.115	0.237	0.00

<표 2>에서는 주격 관계사절과 목적격 관계사절의 사용에 미치는 영향의 정도를 승산비(OR, odds rate) 값으로 보여준다. 목적격 관계사절이 참조범주로 설정되어 승산비는 주격 관계사절이 나올 확률 0 과 목적격 관계사절이 나올 확률 1의 비율이다.[8] 승산비가 1보다 작으면 음의 영향을 준다는 의미로 목적격 관계사절이 사용될 가능성이 높아진다는 것을 뜻하며, 반대로 승산비가 1보다 커서 양의 영향을 준다는 것은 주격 관계사절이 사용될 가능성이 높아진다는 의미이다. 즉, 선행사가 무생물이거나 종속 명사구가 주제적일 때 목적격 관계사절이 나타날 확률이 커지는 것과 대조적으로 종속 명사가 무생물이면 주격 관계사절이 나타날 확률이 커진다. 구체적으로 살펴보면 선행사의 유생성(Head.noun (IA))의 승산비는 0.04로 선행사가 생물(A)일 때에 비해 무생물(IA)일 때 목적격 관계사절이 사용될 가능성이 96%가량 높다

5) 실제 카이제곱 검정을 해보면 선행사의 유생성과 종속 명사의 유생성은 유의미한 연관성이 있으며 그 효과 크기도 매우 크다.
6) 다음의 표는 glm() 명령어를 사용한 결과로 종속변수가 이항변수이므로 family=binominal로 설정하였다. 로지스틱회귀분석은 승산(odds)을 log-변환으로 모형화하기 때문에 회귀 계수(estimate)를 exp로 역변환하여 승산비와 승산비의 95% 신뢰구간을 계산하였다.
7) 모형의 카이제곱(자유도) 유의확률: 1603.10(3) <0.0001; R2= 0.758
8) R은 기본적으로 참조 수준을 알파벳 순서로 결정되기 때문에 목적격 관계사절(ORC)이 참조 범주로 지정된다.

는 것을 의미한다. 이와 대조적으로 종속 명사(Embedded.noun (IA))의 경우 승산비가 31.89 이며 무생물이 생물에 비해 주격 관계사절이 나타날 확률이 97% 증가한다. 종속 명사구의 주제성(Topicality (Topic))의 승산비가 0.166이라는 의미는 주제성이 강한 명사구일 때 목적격 관계사절에 비해 주격 관계사절이 사용될 확률이 83.4% 줄어든다는 것이다.

3.2 상호작용 효과를 보기위한 로지스틱회귀분석

3.2.1 상호작용항(interaction term)을 추가한 분석

하지만 위의 단순로지스틱회귀분석의 결과는 설명변수 간의 상호작용을 고려하지 않은 결과이다. 앞에서 살펴보았듯이 세 설명변수는 타동성의 성질을 나타낸다는 측면에서 서로 밀접한 관계를 가지고 있다. 코퍼스 연구에서 상호작용의 효과를 분석하는 가장 보편적인 방법은 회귀방정식에 상호작용항을 추가하는 것이다(Gries, 2015; Levy, 2012; Manning, 2007 등). 그러나 조건추론나무모형에서 결과를 토대로 세 설명변수의 3차 상호작용을 추가하여 로지스틱회귀분석을 실시하면 식이 수렴(convergence)되지 않는다.[9] 단계적(stepwise)선택법을 사용하여 최적의 회귀방정식을 찾으면 세 변수의 2차 상호작용을 추가한 아래의 식이 나온다.[10] [11]

$$\text{(2) Logit}(\Upsilon=1) = \beta_0 + \beta_1\text{Head.noun} + \beta_2\text{Embedded.noun} + \beta_3\text{Topicality} + \beta_4(\text{Head.noun}\times\text{Topicality}) + \beta_5(\text{Embedded.noun}\times\text{Head.noun}) + \beta_6(\text{Embedded.noun}\times\text{Topicality})$$

분산분석(ANOVA)을 통한 모형 비교 결과 단순 로지스틱회귀(simple logistic regression)모형과 상호작용항들이 추가된 모형 간에는 유의미한 차이가 있으며, 이는 상호작용항들이 유의미한 설명변수임을 의미한다.

<표 3>은 (2)의 회귀방정식의 결과이다. 우선 추정계수에 대해 매우 큰 표준오차를 보이는 경우들이 있다. 종속명사의 유생성(Embedded.noun (IA))의 승산비는 비정상적으로 크며, 종속명사와 선행사의 유생성의 상호작용항(Embedded.noun (IA): Head.noun (IA))의 승산비는 0이다. 두 변수의 표준오차 또한 매우 크다. 이는 본 모형에 문제가 있음을 의미한다.[12] 3.2.2절에서 자세히 논의하겠지만 두 변수의 표준오차가 큰 이유는 준-완전 분리(quasi-complete separation)라고 불리는 현상으로 인해 야기된 것이다.

9) glm()를 사용하며 식의 수렴되지 않는다는 경고와 함께 확률(fitted probabilities)이 0또는 1이 되는 경우가 있다는 경고문이 나온다. 이는 완전분리나 준-완전분리 문제가 있음을 암시한다.

10) 회귀식에서 가장 유의한 변수부터 하나씩 순차적으로 추가하고 새로운 변수가 추가될 때마다 기존 변수의 유의확률을 재평가하여 제거하는 방식으로 최적의 변수들을 선택하는 방식이다. 전진(forward)선택법이나 후진(backward)제거법을 사용해도 동일한 결과가 나온다.

11) R에서 개별 설명변수와 상호작용항을 모두 방정식에 표현하는 Head.noun + Embedded.noun + Head.noun:Embedded.noun은 Head.noun*Embedded.noun으로 축약해서 표현할 수 있다.

12) 로지스틱회귀분석에서 표준오차 값이 큰 경우 가장 먼저 고려되는 것은 다중공선성(multicollinearity)이다. 로지스틱회귀분석에서는 다중공선성 확인을 위해 설명변수에 대한 VIF값을 계산하면 위 세 변수의 VIF 값들이 1.5미만으로 상관성이 낮다고 할 수 있다. 더욱이 다중 공선성은 선형회귀분석 항목에서만 측정 가능하므로 설명 변수들 중 연속 변수가 존재하지 않기 때문에 다중공선성을 논하는 것은 적절치 않다.

표 3. 2차 상호작용항을 추가한 회귀분석(GLM)의 결과[13)]

	Estimates	S.E.	Z value	p-value
(Intercept)	2.773	0.729	3.804	0.000
Embedded.noun (IA)	18.440	498.864	0.037	0.971
Head.noun (IA)	-4.468	0.768	-5.815	0.000
Topicality (Topic)	-3.014	0.765	-3.939	0.000
Head.noun (IA): Topicality (Topic)	1.845	0.822	2.245	0.025
Embedded.noun (IA): Head.noun (IA)	-14.631	498.864	-0.029	0.977
Embedded.noun (IA): Topicality (Topic)	-0.893	0.398	-2.244	0.025

준-완전 분리 문제를 해결한다 하더라도 상호작용항을 사용하는 로지스틱회귀분석에는 또 다른 문제가 있다. 그것은 상호작용 효과의 해석이 용이하지 않다는 점이다. 선형(linear)분석과 달리 비선형(non-linear) 로지스틱회귀분석에서 상호작용은 매우 복잡한 개념으로 특히 상호작용항의 p값은 신뢰하기 어렵다(Ai and Norton, 2003; Norton et al., 2004 등). <표 3>에서 종속명사의 유생성과 선행사의 유생성의 상호작용항(Embedded.noun (IA): Head.noun (IA))의 유의확률을 보자. 0.977로 상호작용 변수가 유의하지 않다는 잘못된 추론을 할 수 있다.[14)]

< 표 3>에서 상호작용항의 해석을 어렵게 하는 또 다른 이유는 참조범주(reference category)를 설정하는 것과 관련되어 있다. 예를 들어 종속 명사의 유생성과 주제성의 상호작용 효과를 제대로 보기 위해서는 상수항을 포함하지 않는 4개의 경우, 즉 각 변수의 두 개의 수준을 결합한 조합에 대한 정보가 필요하다. 하지만 참조범주를 설정하고 상호작용항을 사용하는 현 모형에서는 이러한 상호작용의 효과를 분석할 수 없다.

3.2.2 하나의 통합 변수를 이용한 상호작용 효과

범주형 독립변수 간의 상호작용 효과를 살펴볼 수 있는 방법 중 하나는 실험설계의 요인배치법(factorial design)처럼 각 설명변수의 수준들을 모두 조합하는 방식이다. 위의 나무모형의 결과에 따르며 세 설명변수간의 유의미한 상호작용이 있었기 때문에 우선 세 설명 변수의 수준을 모두 결합한 새로운 변수 Config를 만들었다. Config는 8개의 수준으로 선행사의 유생성-종속 명사의 유생성-주제성 순으로 표기하였다. 예를 들어 Config의 한 수준인 IA-A (Non-topic)은 선행사는 무생물이며 종속 명사구는 생물이며 주제성이 약한 경우를 나타낸다. 이러한 조합 순서는 주격 관계사절과 목적격 관계사절의 분포적 차이가 주어성 또는 목적어성을 나타내는 비통사적 구조들 때문인지 볼 수 있다. 다시 말해 선행사와 종속명사의 유생성 구조 및 종속명사구의 비통사적 성질들이 유기적으로 각 관계사절 내의 선행사와 종속명사구의 문법적 역할을 나타내기 때문에 주격 관계사절과 목적격 관계사절이 서로 다른 분포를 보이는지를 확인할 수 있다.

다음 단계는 상수항을 제거하는 것이다. R를 포함한 대부분의 통계프로그램들은 기본적으로 *k*수준의 범주형 변수를 *k*-1수준의 변수로 처리하는 더미코딩(dummy coding) 방식을 사용하여 회귀분석을 한다. 따라서 변수의 각 수준을 기본 참조 수준과 비교한다. 만약 관계사절 데이터에서 상수항을 제거하지 않는다면 생략되는 기본 참조 범주 조건인 A-A

13) 모형의 카이제곱(자유도) 유의확률 1618.25(6) <0.0001; R2 = 0.749

14) drop1()기능을 사용하여 단일항 삭제(single term deletion)에 대한 정보를 확인하면 Embedded.noun:Head.noun를 포함한 모든 상호작용 변수가 유의하다는 것을 확인할 수 있다.

(Non-topic)에서 주격 관계사절이 나올 확률에 비해 각 수준에서 값의 변화를 계산하게 된다. 그러나 상수항을 제거하여 로지스틱회귀분석을 실시하면 각 수준에서 목적격 관계사절이 나올 확률이 0.5인 경우를 기준으로 승산비를 계산한 값이 나오게 된다. 세 변수들의 조합으로 이루어진 단일 변수 Config를 가지고 로지스틱회귀분석을 실시하기 때문에 상수항 제거로 인해 아래와 같이 Config의 모든 수준의 값이 나온다.[15)]

표 4. 상호작용을 보기 위해 통합된 단일변수를 사용한(절편 없는 모형) 로지스틱회귀분석 결과

	Estimates	S.E.	Z value	p-value
A-A (Non-topic)	2.773	0.729	3.804	0.000
A-A (Topic)	-0.241	0.233	-1.037	0.300
A-IA (Non-topic)	19.566	600.230	0.033	0.974
A-IA (Topic)	19.566	1659.380	0.012	0.991
IA-A (Non-topic)	-1.696	0.243	-6.970	0.000
IA-A (Topic)	-2.864	0.176	-16.242	0.000
IA-IA (Non-topic)	2.113	0.130	16.214	0.000
IA-IA (Topic)	0.051	0.227	0.226	0.821

그러나 위의 결과에서도 앞서 언급한 준-완전분리(quasi-complete separation)문제가 여전히 나타난다. 준-완전분리 문제는 설명변수의 조합이 종속변수를 완벽하게 예측하는 경우가 일부 있을 때 회귀모형에서 발생한다.[16)] A-IA (Non-topic)와 A-IA (Topic)의 경우 표준오차의 값이 비정상적으로 큰 이유가 두 조건에서 목적격 관계사절이 전혀 나타나지 않아 주격 관계사절이 나타날 확률이 1이기 때문이다. 준-완전분리 문제는 A-IA (Non-topic)와 A-IA (Topic)의 경우를 모형에서 제거함으로써 해결될 수 있다(Albert and Anderson, 1984). 즉, 선행사가 생물이고 종속 명사가 무생물인 경우는 주격 관계사절만 나오기 때문에 관계사절 유형의 승산비를 구하는 것은 의미가 없다. 하지만 이러한 독점적인 분포가 A-IA배열의 목적격 관계사절이 자연언어에서 나타나지 않는다는 것을 의미하지 않는 점에 주의해야 한다. 다시 말해 코퍼스 샘플 사이즈를 늘려서 목적격 관계사절이 A-IA의 조건에서 나타나는 경우들을 찾는다면 준-완전분리 문제는 해결될 수 있다. 5개의 모든 레지스터에서 A-IA의 조건에서 목적격 관계사절이 나오지 않았기 때문에 코퍼스 규모를 늘려 목적격 관계사절이 A-IA 구조로 쓰이는 경우를 찾는다 하더라도 그 숫자가 매우 미미하리라는 것을 예측할 수 있다. 따라서 데이터에서 A-IA 유형을 제거하여 준-완전분리 문제를 해결하였다. A-IA (Non-topic)과 A-IA (Topic)에 나타나는 주격 관계사절을 제외한 총 1,558개의 관계사절(목적격 관계사절 853개; 주격 관계사절 705개)을 대상으로 로지스틱 회귀분석을 재실시하였다. 참조범주를 설정하지 않는 방식을 사용했기 때문에 두 개의 수준을 제외해도 <표 4>의 다른 수준들의 통계값에 영향을 받지 않는다.[17)]

15) R에서 상수항을 0 또는 −1로 설정하여 절편없는 모형을 만들 수 있다. 따라서 사용한 명령문은 다음과 같다: glm (RC.type ~ 0+Config, family=binomial, data=RC.data)
16) 분리에 대한 자세한 내용은 Albert and Anderson(1984)을 참조하시오.
17) 모형의 카이제곱(자유도) 유의확률 1139.1(6), p <0.000; R2= 0.69

3.3 레지스터 임의효과를 추가한 혼합모형

Gries는 그의 2015년 논문에서 심리언어학 실험 분석에서는 널리 사용되고 있지만 코퍼스 언어학에서 가장 사용되고 있지 않는 혼합모형(mixed effects models)이 언어학적 요소뿐만 아니라 레지스터(register)등과 같은 비언어학적 요소까지 고려해야 하는 코퍼스 언어학 분석에서도 매우 유용하다는 것을 보여주었다. 그 이후 코퍼스 언어학 연구에서의 혼합모형(mixed effects model) 사용에 대한 논의가 활발히 이루어지고 있다(Barth and Kapatsinski 2018; Gries, 2021; Hörberg, 2018 등).

관계사절 관련 선행 코퍼스 연구들은 어떤 언어학적 요인들이 목적격 관계사절과 주격 관계사절의 빈도분포에 영향을 미치는지를 찾는 데 집중되어져 있었으며 레지스터의 효과에 대해서는 간과하였다. 즉, 대부분의 이들 연구들은 문어체 코퍼스 위주로 분석한 결과로 주격 관계사절이 목적격 관계사절보다 많이 사용되지만 선행사의 유생성이나 종속명사구의 주제성 때문에 목적격 관계사절이 더 많이 쓰이는 환경이 있다는 것을 보여주었다. 하지만 일부 코퍼스 연구들은 구어체인지 문어체인지에 따라 관계사절 유형의 분포에 차이가 있음을 지적하였다(Gordon and Hendrick, 2005; Roland et al., 2012). 예를 들어 Roland et al.(2012)은 Brown and Switchboard 코퍼스(Treebank-3버전)에서 목적격 관계사절과 타동사를 포함한 주격 관계사절을 축출한 결과 목적격 관계사절이 주격 관계사절보다 더 많이 출현하는 것을 보여주었다. 하지만 관계사절 분포 연구에서 이러한 비언어적 요소가 주는 효과에 대해서는 통계적으로 분석되지 못하였다.

본 논문에서 사용된 자료는 3개의 구어체 레지스터와 2개의 문어체 레지스터에서 축출한 것으로 모드(mode)에 따른 관계사절 분포를 살펴보았다. 문어체 코퍼스에서는 주격 관계사절이 더 많이 사용되지만 구어체 코퍼스에서는 목적격 관계사절이 1.24배 더 자주 출현하였다. 더욱이 <그림 4>의 관계사절 분포를 보면 모드뿐만 아니라 레지스터 간의 차이도 발견되었다. BNC 코퍼스의 구어체 강의 데이터인 S2는 학생들의 에세이로 이루어진 문어체 W1과 W2처럼 주격 관계사절이 더 많이 사용되었다.

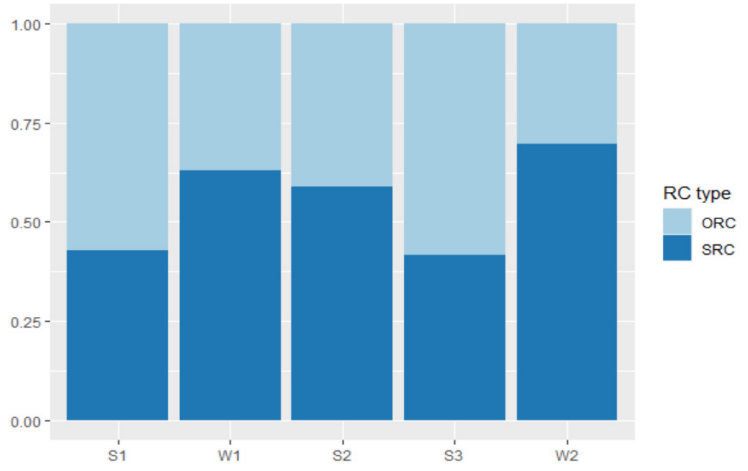

그림 4. 구어체(S)와 문어체(W) 레지스터별 관계사절 유형의 분포

이렇듯 레지스터에 따라 주격 관계사절과 목적격 관계사절의 상대적 빈도가 매우 다를 수 있기 때문에 레지스터의 임의효과를 추가한 혼합로지스틱회귀모형(mixed effects logistic regression model)을 사용하여 분석하였다. 이 혼합모형은 고정효과만을 고려한 모형과 유의미한 차이가 있었다.

표 5. 혼합로지스틱회귀분석 결과

	Estimates	S.E.	Z value	OR	95% CI		p-value
					low	high	
A-A (Non-topic)	2.837	0.748	3.792	17.059	4.918	107.828	0.000
A-A (Topic)	-0.267	0.280	-0.952	0.766	0.431	1.351	0.341
IA-A (Non-topic)	-1.720	0.292	-5.896	0.179	0.097	0.318	0.000
IA-A (Topic)	-2.778	0.233	-11.923	0.062	0.038	0.102	0.000
IA-IA (Non-topic)	2.164	0.208	10.397	8.703	5.587	14.014	0.000
IA-IA (Topic)	0.172	0.284	0.604	1.187	0.675	2.148	0.546

위 결과에 따르면 목적어-주어의 전형적인 유생성 구조인 IA-A배열과 종속명사구가 전형적인 주어인 A (Topic)인 경우에 목적격 관계사절이 사용될 확률이 더 높으며 그 외의 경우에는 주격 관계사절의 사용 확률이 높다. R의 effects 패키지(Fox et al., 2020)을 이용하여 설명변수의 효과를 아래와 같이 시각화할 수 있다.

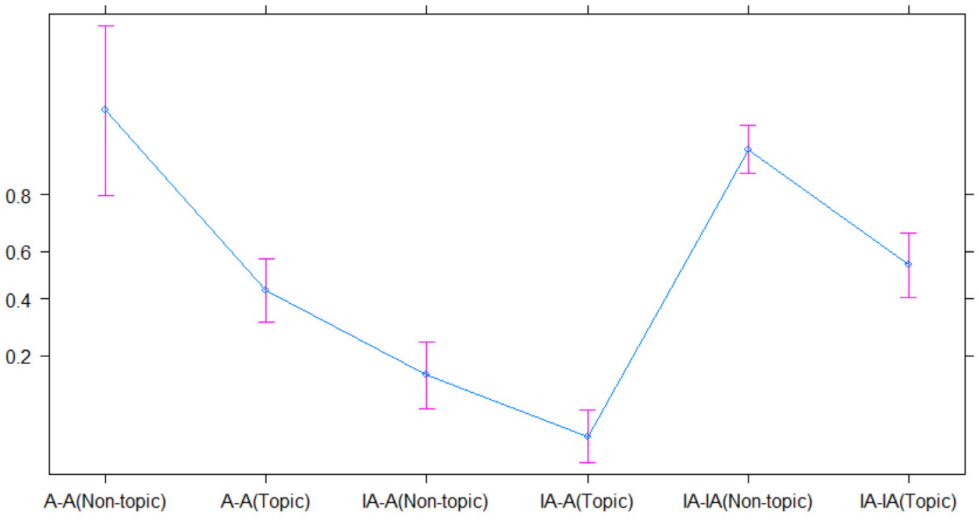

그림 5. <표 5>의 로지스틱 회귀분석 결과에 따른 Conf 변수의 효과

그림에서 유생성이 같은 쌍들을 비교해보자. 선행사와 종속명사구의 유생성이 같은 경우와 달리 IA-A의 경우 주격 관계사절에 비해 목적격 관계사절이 나타날 확률이 항상 유의미하게 크다. 따라서 A-IA의 경우 주격 관계사절만 출현한다는 것을 감안한다면 주격 관계사절과 목적격 관계사절이 A-IA와 IA-A의 경우 대조적인 분포를 보이며 이는 주격 관계사절과 목적격 관계사절 각각 주어가 생물이고 목적어가 무생물인 전형적인 타동사절의 유생성 구조를 선호한다는 것을 알 수 있다. A-A와 IA-IA의 경우를 비교해 보면 종속명사구의 주제성이 약한 Non-topic 일 때 주격 관계사절이 쓰일 확률이 유의미하게 커지지만 주제성이 강한 Topic일 때는 이러한 주격 관계사절 선호도는 사라지는 패턴을 보인다. 이처럼 선행사와 종속명사의 유생성이 같으면 유생성으로 관계사절의 타동성을 나타내지 못하지만 주제성이 약하면 목적어성이 커지기 때문에 종속명사구가 Non-topic인 경우 종속명사구가 목적어의 역할을 하는 주격 관계사절이 선호된다는 것을 보여준다. 비록 A-A (Topic)와 IA-IA (Topic)의 경우 종속명사구의 주제성이 주어성을 나타내지만 관계사절 선택에 영향을 크게 주지 않는다. 하지만 이 역시 A-A (Non-topic)과 IA-IA (Non-topic) 경우에서 보이던 주격 관계사절 선호

도는 사라진다는 점에서 주제성이 강한 종속 명사구(topic NP)로 인해 목적어성이 약해지는 동시에 주어성이 커져 관계사절 유형 선호도에 영향을 준다는 것을 보여준다. 결론적으로 혼합로지스틱회귀분석을 통해 주격 관계사절과 목적격 관계사절이 보이는 분포의 차이는 각 관계사절이 전형적인 타동사 구문의 의미·담화적 구조를 가지려는 경향이 있기 때문이라는 것을 효과적으로 보여줄 수 있다.

4. 결 론

언어 구조의 다양한 측면을 반영하는 요인들은 서로 유기적으로 연결되어 구문 유형 선택에 있어 상승효과(synergistic effect)를 일으킬 수 있다. 예를 들어 타동사 구문은 단순히 주어와 목적어를 취하는 통사적 구조뿐만 아니라 이 통사적 정보와 유기적으로 연결된 의미·담화적 정보 구조도 가지고 있다. 이러한 비통사적 구조 때문에 주어의 자리에는 주제성이 강하고 사람을 지칭하는 명사구가, 목적어 자리에는 반대로 주제성이 약하고 무생물을 지칭하는 명사구가 사용될 가능성이 크다. 선행사의 유생성, 그리고 종속 명사구의 유생성과 주제성은 명사구의 통사적 기능을 반영하기 때문에 이 요인들은 주격 관계사절과 목적격 관계사절의 사용 차이를 설명하는 데 있어 독립적인 역할을 하기 보다는 유기적으로 연결되어져 있다. 즉, 세 요인들이 절의 타동성을 나타내는 하나의 요인으로 해석될 수 있기 때문에 설명 변수들의 3차 상호작용을 나타낼 수 있는 새로운 변수를 만들어 재해석할 필요가 있다. 본 논문에서는 보통 2개의 가지로만 분할하는 이진분할방법을 사용하는 나무모형으로는 이러한 분석이 이루어지기가 어려우며, 참조 범주를 설정하고 상호작용항을 추가하는 기존의 회귀 분석으로도 주격 관계사절과 목적격 관계사절이 다른 언어적 맥락에서 사용된다는 사실을 밝히기에 적절치 않음을 보여주었다. 대안으로 상수항을 제거하고 모든 변수들의 수준을 결합한 하나의 변수를 만드는 방식의 로지스틱회귀모형을 제안하였다. 이 모형에서는 나무모형과 달리 준-완전분리 문제가 발생하지만 코퍼스 데이터의 규모를 키우거나 원인이 되는 경우를 제거함으로써 해결될 수 있었다. 이 새로운 대안이 나무모형 보다 통합적 상호작용 효과를 가지고 오는 요인들 간의 유기적 관계를 좀 더 정밀하고 쉽게 파악할 수 있게 해준다. 로지스틱회귀모형의 또 다른 장점은 기존의 관계사절 빈도 연구에서 고려되지 않았던 비언어학적 요소인 레지스터(register)의 임의효과를 추가한 혼합로지스틱회귀분석을 할 수 있다는 점이다.

참고문헌

Ai, C. and E. C. Norton. 2003. Interaction Terms in Logit and Probit Models. *Economics Letters* 80.1, 123-129.

Aissen, J. 1999. Markedness and Subject Choice in Optimality Theory. *Natural Language and Linguistic Theory* 17.4, 673-711.

Aissen, J. 2003. Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory* 21.3, 435-483.

Albert, A. and J. Anderson. 1984. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika* 71.1, 1-10.

Barth, D. and V. Kapatsinski. 2018. Evaluating Logistic Mixed-effects Models for Corpus-linguistic Data in Light of Lexical Diffusion. In D. Speelman, K. Heylen, and D. Geeraets (eds), *Mixed-effects Regression Models in Linguistics*. Springer. 99-116.

Comrie, B. 1989. *Language Universals and Linguistic Typology*. Chicago: University of Chicago Press.

Eddington, D. 2010. A Comparison of Two Tools for Analyzing Linguistic Data: Logistic Regression and Decision Trees. *Italian Journal of Linguistics* 22.2, 265-286.

Fox, J., W. Sanford, B. Price, J. Hong, R. Anderson, D. Firth, S. Taylor, and the R Core Team. 2020. Effect Displays for Linear, Generalized Linear, and Other Models. UTC.

Gennari, S. and M. MacDonald. 2008. Semantic Indeterminacy in Object Relative Clauses. *Journal of Memory and Language* 58.2, 161-187.

Gennari, S. and M. MacDonald. 2009. Linking Production and Comprehension Processes: The Case of Relative Clauses. *Cognition* 111.1, 1-23.

Gordon, P., R. Hendrick, and M. Johnson. 2001. Memory Interference during Language Processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27.6, 1411-1423.

Gordon, P., R. Hendrick, and M. Johnson. 2004. Effects of Noun Phrase Type on Sentence Complexity. *Journal of Memory and Language* 51.1, 97-114.

Gordon, P. and R. Hendrick, 2005. Relativization, Ergativity, and Corpus Frequency. *Linguistic Inquiry* 36, 456-463.

Gries, S. 2015. The Most Under-used Statistical Method in Corpus Linguistics: Multi-Level (and mixed-effects) Models. *Corpora* 10.1, 95-125.

Gries, S. 2020. On Classification Trees and Random Forests in Corpus Linguistics: Some Words of Caution and Suggestions for Improvement. *Corpus Linguistics and Linguistic Theory* 16.3, 617-647.

Gries, S. 2021. (Generalized linear) Mixed-effects Modeling: A Learner Corpus Example. *Language Learning* 1-42.

Hörberg, T. 2018. Functional Motivations Behind Direct Object Fronting in Written Swedish: A Corpus-Distributional Account. *Glossa* 3.1, 81.

Jenset, G., B. McGillivray, and M. Rundell. 2018. The English Dative Alternation Revisited: Fresh Insights from Contemporary British Spoken Data. In V. Brezina, R. Love, and K. Aijmer (eds.), *Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC 2014*. London: Routledge, 185-207.

Just, M. and P. Carpenter. 1992. A Capacity Theory of Comprehension: Individual Differences in Working Memory Capacity. *Psychological Review* 99.1, 122-149.

King, J. and M. Just. 1991. Individual Differences in Syntactic Processing: The Role of Working Memory. *Journal of Memory and Language* 30.5, 580-602.

Levy, R. 2012. Probabilistic Methods in Linguistics. Lecture 14: Logistic regression. Manuscript. UC San Diego.

Manning, C. 2007. Logistic regression (with R). Manuscript. Stanford University.

Norton, E. C., H. Wang, and C. Ai. 2004. Computing Interaction Effects and Standard Errors in Logit and Probit Models. *Stata Journal* 4.2, 154-167.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Reali, F. and M. Christiansen. 2007. Processing of Relative Clauses is Made Easier by Frequency of Occurrence. *Journal of Memory and Language* 57, 1-23.

Roland, D., F. Dick, and J. Elman. 2007. Frequency of Basic English Grammatical Structures: A Corpus Analysis. *Journal of Memory and Language* 57, 348-379.

Roland, D., G. Mauner C., O'Meara, and H. Yun. 2012. Discourse Expectations and Relative Clause Processing. *Journal of Memory and Language* 66, 479-508.

Shin, K. 2019. An Expectation-Based Account for the Processing Difficulty of it Object-Extracted Relative Clauses. *Korean Journal of Linguistics* 44.4, 807-829.

Shin, K. 2020a. Some Remarks on Gries's Criticism on a Tree-Based Approach of Multifactorial Data. *Language and Information* 24.1, 15-28.

Shin, K. 2020b. Non-linear Interactions of Factors Influencing Relative Clasue Distribution and Their Implications on Relative Clause Processing. *Korean Journal of Linguistics* 45.1, 919-940.

Silverstein, M. 1976. Hierarchy of Features and Ergativity, In R. Dixon (ed.), *Grammatical Categories in Australian Languages*. Canberra: Australian Institute of Aboriginal Studies, 112-171.

Traxler, M., R. Morris, and R. Seely. 2002. Processing Subject and Object Relative Clauses: Evidence from Eye Movements. *Journal of Memory and Language* 47.1, 69-90.

Traxler, M., R. Williams, S. Blozis, and R. Morris. 2005. Working Memory, Animacy, and Verb Class in the Processing of Relative Clauses. *Journal of Memory and Language* 53.2, 204-224.

신근영, 교수
광주 광역시 북구 용봉로 77
전남대학교 영어영문학과
E-mail: kyshin@chonnam.ac.kr