선행 발화의 중심 전이를 이용한 영형 생성
(Generation of Zero Pronouns using Center Transition of Preceding Utterances)

노 지 은 † 나 승 훈 † 이 종 혁 ††
(Ji Eun Roh) (Seung Hoon Na) (Jong Hyeok Lee)

요 약 자연스러운 텍스트를 생성하기 위해서는, 한번 언급된 대상을 지시하기 위한 대용화(pronominalization)과정이 필수적이며, 특히 한국어에 빈번히 발생하는 영형(zero pronoun)을 자연스럽게 생성하는 것이 중요하다. 본 논문에서는, 비용기반 중심화 이론(cost-based centering theory)을 적용하여, 선행 발화의 중심 전이(center transition)가 현 발화의 영형에 미치는 영향을 살펴본다. 이를 위해, 영형으로 실현될 수 있는 명사를 중심화 이론에 기반해 문장간 현저성, 문장내 현저성, 문장간/내 현저성을 가지는지의 여부로 4가지 유형(Npair, Ninter, Nintra, Nnon)으로 정의하고, 유형별로 영형 현상을 고찰하였다. 그 결과, 기존에 중심화 이론에서 배제되었던 명사들이 선행 발화의 중심 전이로 설명될 수 있음을 밝혔다. 또, 선행 발화의 중심 전이를 이용한 영형 생성 모델을 구축하여 다양한 자질을 적용한 영형 생성 모델의 성능과 비교하였다.

키워드 : 영형, 비용기반 중심화 이론, 중심 전이, 문장간 현저성, 문장내 현저성, 문장간/내 현저성

Abstract To generate coherent texts, it is important to produce appropriate pronouns to refer to previously-mentioned things in a discourse. Specifically, we focus on pronominalization by zero pronouns which frequently occur in Korean. This paper investigates zero pronouns in Korean based on the cost-based centering theory, especially focusing on the center transitions of adjacent utterances. In previous centering works, only one type of nominal entity has been considered as the target of pronominalization, even though other entities are frequently pronominalized as zero pronouns. To resolve this problem, and explain the reference phenomena of real texts, four types of nominal entity (*Npair, Ninter, Nintra,* and *Nnon*) from centering theory are defined with the concept of *inter-*, *intra-*, and *pairwise salience*. For each entity type, a case study of zero phenomena is performed through analyzing corpus and building a pronominalization model. This study shows that the zero phenomena of entities which have been neglected in previous centering works are explained via the center transition of the second previous utterance. We also show that in Ninter, Nintra, and Nnon, pronominalization accuracy achieved by complex combination of several types of features is completely or nearly achieved by using the second previous utterance's transition across genres.

Key words : zero pronoun, cost-based centering theory, center transition, inter-salience, intra-salience, pairwise-salience

1. 서 론

자연스러운 텍스트를 생성하기 위해서는, 한번 언급된 대상을 지시하기 위한 대용화(pronominalization)과정이 필수적이며, 특히 한국어에 빈번히 발생하는 영형(zero pronoun)을 자연스럽게 생성하는 것이 중요하다. 한국어는 문맥 의존 언어(context-sensitive language)로, 영어와 달리 조응 표현으로 영형의 사용이 두드러지며, 주어진 문맥에서 회복 가능한 임의의 논항(argument)들은 기본적으로 모두 생략 가능하다. 본 논문에서는 영형의 대상으로 '명사'(nominal entity)만을 고려하며 '구'나 '절' 등의 생략은 다루지 않는다.

어떤 명사의 대용화 현상을 설명하기 위해, 명사의 현

· 본 연구는 첨단정보기술연구센터를 통한 과학재단 및 2002년도 두뇌한국21사업에 의하여 지원되었음
† 비 회 원 : 포항공과대학교 컴퓨터공학과
jeroh@postech.ac.kr
nsh1979@postech.ac.kr
†† 종신회원 : 포항공과대학교 컴퓨터공학과 교수
jhlee@postech.ac.kr
논문접수 : 2005년 5월 4일
심사완료 : 2005년 8월 18일

저성(salience of nominal entity)의 관점에서 많은 기존 연구들이 선행되었는데[1-6], 그 중, 중심화 이론(centering theory)[6]은 언어학, 심리 언어학, 계산 언어학 등에서 가장 활발하게 적용되고 있는 이론이다. 중심화 이론은 이론 자체의 단순함에도 불구하고, 자연어 처리의 여러 분야에 응용되는 강력한 계산 모델이다. 초기에는 대용 해석을 위해 주로 활용되었지만, 최근에는 텍스트 생성의 여러 과정 - 텍스트 구조화(text structuring)[7], 문장 계획(sentence planning)[8], 지시어 생성(referring expression generation)[9-14] - 에 널리 적용되고 있다.

본 논문은, 실제 코퍼스 상에서 발생하는 영형 현상을 규명하기 위해 중심화 이론을 적용하는 기존 연구들의 2가지의 문제점에서 출발한다. 첫째, 중심화 이론을 토대로 대용 현상, 한국어에서는 특히 영형 현상을 규명하는 대부분의 기존 연구들은 중심화 이론의 한가지 논항, 즉, CONTINUE 전이의 Cb에 대해서만 대용 현상을 규명하고 있다[9-13,15-18]. 하지만, [13], [14], [19]에서 지적한 것처럼, 실제 텍스트 상에서 영형의 발생은 CONTINUE 전이의 Cb에서만 발생하는 것이 아니라 다른 유형의 명사, 즉 Cb가 아닌 다른 논항에서도 빈번히 발생한다. 기존 연구에서는 이런 명사들의 생략 현상을 중심화 이론으로 설명하지 못한다. [14]에서는, 영형 생성의 대상으로 모든 구정보를 다 포함시키되, 세 종류 *Cb*, *oldCp*, *oldR*로 나누어서 영형을 고찰하였지만, CONTINUE 전이의 Cb외에 다른 종류의 명사의 영형에 미치는 자질을 중심화 이론으로 설명하지 못하였다. 따라서, 본 연구의 첫 번째 동기는, 기존에 중심화 이론에서 간과되어 왔던 명사 부류의 영형 현상을 중심화 이론으로 어떻게 설명할 것인가 하는 것이다. 둘째, 중심화 이론을 통해 텍스트의 응집성, 대용 해결, 대용어 생성들을 설명할 때 대부분의 논문들은 인접한 두 발화(U_{i-1}, U_i)의 전이를 함께 고려하는 것이 유용하다는 것에 의견의 일치를 보인다[6,9,10,13,20-27]. 하지만, U_i에서 대명사를 생성(또는 해결)할 때, 이전 두번째 발화의 전이유형이 미치는 영향에 대한 고찰은 이루어 지지 않았다. [14]에서 유일하게 영형 생성 모델을 구축할 때 U_{i-1}의 전이유형뿐만 아니라, U_{i-2}의 전이유형을 추가적으로 고려하여, 기계 학습을 위한 자질중의 하나로 사용하였다. 하지만, 실제 U_{i-2}의 전이유형이 U_i에서 영형을 생성하는데 미치는 영향을 설명하지 못하였다. 따라서, 본 연구의 두 번째 동기는, U_i에서 영형을 생성할 때 U_{i-1}의 전이유형뿐만 아니라 범위를 더 확대하여 U_{i-2}의 전이유형이 U_i에서 영형의 생성에 어떤 영향을 미치는지를 고찰하는 것이다.

이런 두 가지 이슈를 다루기 위해, 본 논문에서는 먼저, 중심화 이론에 기반해서 영형 생성 가능한 4가지 유형 - Npair, Ninter, Nintra, Nnon - 의 명사를 정의한다. 이 네 유형은 서로 다른 현저성의 강도를 갖는데 이를 문장간 현저성(inter-salience), 문장내 현저성(intra-salience), 문장간/내 현저성(pairwise salience)으로 설명한다. 각 명사 유형별로, 선행 발화의 전이에 초점을 맞추어 코퍼스를 분석하고, 지시어 생성 모델을 구축하여 기존에 중심화 이론에서 배제되었던 명사들의 영형 현상을 설명한다. 또, 선행 발화의 중심 전이를 이용해 구축된 영형 생성 모델과, 다양한 자질을 적용한 영형 생성 모델의 성능을 비교, 그 상대적인 중요성을 파악한다.

2. 중심화 이론과 비용기반 중심화 이론

중심화 이론[6]은 텍스트를 구성하고 있는 발화의 각 명사(구) 관점에서 응집성(cohesion)과 현저성(salience)의 상호작용을 통해, 텍스트의 국소적 결속성(local coherence)을 모델링한 담화 해석의 계산 모델이다.

중심화 이론에서 분석의 최소 기본 단위는 발화(utterance)로, 한 개의 발화는 세 개의 중심구조 - Cf(forward-looking center), Cb(backward-looking center), Cp(preferred center) - 를 가진다. 이때 중심(center)은 발화 시점에서 화자의 의식이 활성화되고 집중되어 있는 대상물들을 말한다. Cf는 현 발화에서 실현된 객체 지시물들이고, Cf-list는 발화에 실현된 객체 지시물들을 화자의 의식 내에서 활성화된 정도에 따라 서열을 매긴 것으로, 다음 발화에 나타나게 될 지시물에 대한 선행사(antecedent)의 집합이다. Cf-list에 있는 지시물 중에서 가장 높은 서열에 있는 지시물은 Cp (preferred center)가 되며, Cp는 다음절에서 주제로 논의될 가능성이 가장 높은 후보자이다. Cb는 문장의 주제(topic)와 유사한 개념으로, 많은 경우에 바로 앞의 발화의 Cp가 다음 발화에서 Cb가 된다.

중심화 이론에서 가정하는 세 가지 제약과 두 가지 규칙은 다음과 같다.

- **제약(constraints)**
  1. 각 발화 내에 하나의 Cb가 있다.
  2. 각 발화의 Cf 목록의 모든 요소는 반드시 현 발화 안에서 실현되어야 한다.
  3. 각 발화의 Cb는 현 발화(U_i)에서 실현된, 바로 전 발화(U_{i-1})의 Cf에서 가장 높은 순위의 담화 요소이다.
- ***규칙(rules)***
  1. 앞 발화(U_{i-1})의 Cf의 어떤 요소가 현 발화(U_i)에서 대명사화 되었다면, 현 발화(U_i)의 Cb도 역시 대명사화 된다.
  2. 발화간의 전이유형은 다음 순서로 선호된다.

CONTINUE > RETAIN > SMOOTH-SHIFT > ROUGH-SHIFT

전이유형은 $Cb(U_i)$와 $Cb(U_{i-1})$의 일치 여부와, $Cb(U_i)$와 $Cp(U_i)$의 일치 여부에 의해 결정된다(표 1). CON은 화자가 특정 지시물에 대해 이야기하고 있으면서, 다음 발화에서도 계속 그 지시물에 대해 이야기 하겠다는 의도를 표시하며, 그 지시물은 현 발화에서 Cb인 동시에 Cp로 표현된다. RET는 화자가 다음절에서 의식의 중심을 새로운 대상으로 옮기고 싶다는 의도를 표시하는 것으로, Cb는 그대로 유지되지만, 현재의 중심을 Cf-list에서 낮은 서열에 배치함으로써 Cp가 바뀌는 경우이다. SSH는 이전 발화와 비교해 중심은 변했지만, 새로운 중심에 대해 이야기하며 다음 발화에서도 계속 새롭게 바뀐 현재의 중심에 대해 이야기 하겠다는 의도를 표시한다. RSH는 중심도 변하고 새로운 중심이 Cf-list에서 낮은 서열에 배치되는 경우이다.

[20]에서는 추론 비용(inference cost)을 고려하여 중심화 이론을 재고안 하였는데, $Cb(U_i)$와 $Cp(U_{i-1})$의 일치 여부에 따라 기존의 네 가지 전이유형을 여섯 가지로 확대하였다(표 2). [14]에서는 이렇게 여섯 가지의 전이유형으로 확장된 중심화 이론을 '비용기반 중심화 이론'이라고 명명하고 한국어의 영형을 설명하는데 적용하였다. [13], [14]에서 한국어의 영형의 대부분은 E-CON의 Cb에서보다 C-CON의 Cb에서 빈번히 발생한다고 밝힌 바 있어, 본 논문에서도 비용기반 중심화 이론의 확장된 여섯 가지 전이유형을 채택한다.

중심화 이론에서 발화의 단위와 Cf-list에서 순위 결정은 언어에 따라 조금씩 다르며, 같은 언어에 대해서도 학자들마다 그 설정 기준이 다르다. 본 연구에서는 발화의 단위를 시제절로 정의하고[13-18], [15]에서 설정한 다음과 같은 순위로 Cf-list의 순위를 결정한다.

주제[1] > 주어 > 직접 목적어 > 간접 목적어 > 관형어 > 부사어

1) 본 논문에서 '주제(topic)'는 주제 표지 (topic marker) '은/는'을 조사로 갖는 단어를 의미한다.

실제 [15]에서는 관형어와 부사에 대해서는 순위를 명시하지 않았지만, 본 논문에서는 직관적으로 관형어를 부사어보다 높은 순위에 두었다. 각각의 동일 순위에 여러 명사가 존재할 경우, 문장에서 먼저 실현된 것을 우선 순위에 두어 순서를 매겼다. 다음 예제 텍스트를 통해 비용기반 중심화 이론이 어떻게 적용되는지 살펴보자.

(텍스트 1)

U_1: 자귀는 우리나라의 전통적인 농기구이다.
⇨ Cf: 자귀 > 우리나라 > 농기구, Cp: 자귀

U_2: 자귀의 날은 절삭날이라고도 한다.
⇨ Cf: 날 > 자귀 > 절삭날, Cp: 날, Cb: 자귀, 전이: RET

U_3: 자귀는 형태가 도끼와 비슷하게 생겼는데,
⇨ Cf: 자귀 > 형태 > 도끼, Cp: 자귀, Cb: 자귀, 전이: E-CON

U_4: [자귀는(ø)] 크기에 따라 대자귀, 중자귀, 소자귀로 나누어진다.
⇨ Cf: 자귀 > 크기 > 대자귀 > 중자귀 > 소자귀, Cp: 자귀, Cb: 자귀, 전이: C-CON

U_5: 대자귀는 날의 볼이 얇은 것과 두꺼운 것이 있다.
⇨ Cf: 대자귀 > 볼 > 날, Cp: 대자귀, Cb: 대자귀, 전이: E-SSH

위의 텍스트는 한국의 전통적인 농기구 '자귀'에 대한 설명으로, 텍스트 전체의 주제는 '자귀'인데, U_2에서는 화제가 '자귀'에서 '날'로 자연스럽게 바뀐 다음, U_3에서 다시 '자귀'로 돌아오고 U_5에서는 자귀의 한 종류인 '대자귀'로 화제가 바뀌는 흐름을 가진다. U_1은 첫 발화이므로 Cb가 없다. U_2에서는, U_2에 실현된 명사 중, U_1에서 Cf의 서열상 가장 높게 실현된 명사가 '자귀'이므로 $Cb(U_2)$는 '자귀'가 되고, U_2에서 가장 높게 실현된 명사는 '날'이므로 $Cp(U_2)$은 '날'이 된다. $Cb(U_2) \neq Cp(U_2)$이고 $Cb(U_1)$이 정의되지 않았으므로 U_2의 전이는 RET이 된다. U_3에서 $Cb(U_3)$은 같은 방식으로 '자귀'가 되고 $Cp(U_3)$도 '자귀'로 $Cb(U_3) = Cp(U_3) = Cb(U_2)$이지만, $Cb(U_3) \neq Cp(U_2)$이므로 E-CON이 실현된다. U_4에서 대괄호 안의 명사는 생략된 명사, 즉 영형으로 실현된

표 1 발화간의 전이유형(transition type)

	$Cb(U_i) = Cb(U_{i-1})$ 또는 $C(U_{i-1}) = NULL$	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i) = Cp(U_i)$	CONTINUE (CON)	SMOOTH-SHIFT (SSH)
$Cb(U_i) \neq Cp(U_i)$	RETAIN (RET)	ROUGH-SHIFT (RSH)

표 2 추론 비용을 고려한 발화간의 전이유형 [20]

	$Cb(U_n) = Cb(U_{n-1})$	$Cb(U_n) \neq Cb(U_{n-1})$
$Cb(U_n) = Cp(U_n)$ and $Cb(U_n) = Cp(U_{n-1})$	Cheap-CONTINUE (C-CON)	Cheap-SMOOTH-SHIFT (C-CON)
$Cb(U_n) = Cp(U_n)$ and $Cb(U_n) \neq Cp(U_{n-1})$	Expensive-CONTINUE (E-CON)	Expensive-SMOOTH-SHIFT (E-SSH)
$Cb(U_n) \neq Cp(U_n)$	RETAIN (RET)	ROUGH-SHIFT (RSH)

명사를 표시한다. U_4에서는 $Cb(U_4) = Cp(U_4) = Cb(U_3) = Cp(U_3)$이므로 C-CON이 실현되고, U_5에서는 $Cb(U_5) = Cp(U_5) = Cp(U_4)$이지만 $Cb(U_5) \neq Cb(U_4)$이므로 E-SSH가 실현된다.

3. 문장간/내 현저성에 기반한 네 가지 명사 유형

중심화 이론에서 규칙 1은, 중심화 이론이 대명사 해석 또는 생성에 적용될 수 있는 주요한 근거를 제공한다. 규칙 1은, U_i에서 대명사가 한 개 이상 실현되었다면 그 중 하나는 반드시 Cb라는 의미이다. 이는, 대명사 생성 측면에서 U_i의 Cb는 대명사화 할 수 있다는 것을 의미한다. 특히, 중심화 이론을 대명사 생성에 적용한 기존 연구[9-13,15-18]에서 대용형이 영형이든 영형이 아닌 대명사든 간에, 좁게는 CON의 Cb만, 넓게는 CON외에 SSH의 Cb를 영형으로 생성하고자 했다. 하지만 [13], [14], [19]에서 지적한 것처럼 실제 텍스트 상에서는 Cb가 항상 생략되는 것은 아니며, 또 Cb가 아닌 것들도 빈번히 생략될 수 있다. 이런 측면에서 기존의 특정 전이(CON 또는 CON과 SSH)에서의 Cb로 영형 현상을 제약하는 것은 문제가 있다. 따라서, 현 발화에서 대명사화 가능한 명사들을 다 영형 생성 가능한 범주에 포함시키되 각 명사들이 갖는 현저성의 정도를 구별하여 영형 현상을 관찰할 필요가 있다. 이를 위해 본 논문에서는 중심화 이론에 기반해 네 가지 명사 유형을 표 3과 같이 정의하였다.

네 유형을 설명하기에 앞서, 본 논문에서는 U_i에서 대명사화 가능한 명사를 U_{i-1}에서 실현된 명사로 제한한다. 즉, U_i에 실현된 명사들 중 바로 직전 발화 U_{i-1}에서 실현된 명사들만을 U_i에서 대명사화 가능한 명사들로 정의하며, 이런 명사들을 구명사(old nominal entity)라 칭한다. 따라서 U_i의 구명사에 한해, 각각이 구명사들은 표 3의 네 가지 유형 중 하나에 해당되며 U_i에서 신명사(U_{i-1}에 나타나지 않은 명사)들은 대용화 대상이 아니므로 위의 명사 유형에 해당되지 않는다.

각 명사 유형은, 각각이 $Cb(U_i)$, $Cp(U_i)$와 일치하는지의 여부로 구별된다. 중심화 이론에서 $Cb(U_i)$는 U_i에 실현된 명사 중 U_{i-1}에서 가장 현저하게 실현된 명사이므로, 그 명사는 U_{i-1}의 관점에서 U_i에서 가장 현저한 대상을 의미한다. 반면, $Cp(U_i)$는 U_{i-1}에서 그것의 현저함과는 상관없이 U_i 내부에서 가장 현저한 대상을 의미한다. 따라서, 첫 번째 유형으로, $Cb(U_i)$이지만 $Cp(U_i)$가 아닌 명사 부류를 *문장간에 현저한 명사*(약어로, *Ninter*)라 부른다. 이는, U_i에서는 가장 현저하진 않지만 U_{i-1}에 의해서는 가장 현저한 명사를 일컫는다. 즉, Ninter는 현 발화에서, 직전 발화에(inter-sententially) 의해 현저성을 갖지만, 현 발화(intra-sententially)내에서 가장 현저하지는 않다는 뜻에서 문장간에서 현저하다고 정의하였다.

두번째 유형으로, $Cb(U_i)$는 아니지만 $Cp(U_i)$인 명사 부류를 *문장내에 현저한 명사*(약어로, *Nintra*)라 부른다. 이는, U_i에서는 가장 현저하지만 U_{i-1}의 관점에서는 가장 현저하지 않은 명사를 일컫는다. 즉, Nintra는 현 발화에서, 직전 발화에(inter-sententially) 의해 현저성을 갖지는 않지만, 현 발화(intra-sententially)내에서는 가장 현저하다는 뜻에서 문장내에서 현저하다고 정의하였다.

세 번째 유형으로, $Cb(U_i)$이기도 하고 $Cp(U_i)$이기도 한 명사 부류를 *문장간/내에서 현저한 명사*(약어로, *Npair*)라 부른다. 이는, U_i에서 가장 현저한 동시에 U_{i-1}에 의해서도 가장 현저한 명사를 일컫는다. 즉, 문장간/내에서 현저한 명사는 인접한 문장상에서 문장내적(intra-sententially), 외적(inter-sententially)으로 동시에 현저한 명사로 정의된다.

마지막으로 $Cb(U_i)$도 $Cp(U_i)$도 아닌 명사 부류를 *현저하지 않은 명사*(약어로 *Nnon*)라 부른다. 이는, U_i에서 가장 현저하지도 않으면서 U_{i-1}에 의해서도 가장 현저하지 않은 명사를 일컫는다.

Nintra(U_i), Ninter(U_i), Npair(U_i)의 개수는 U_i에서 많아야 1개인데, 이는 $Cb(U_i)$, $Cp(U_i)$ 각각이 U_i에서 1개 이상일 수 없다는 중심화 이론의 가정 때문이다. 반면에, Nnon(U_i)의 개수는 1개 이상일 수 있다. 같은 이유로 Npair(U_i)와 Ninter(U_i)가 U_i에서 함께 존재할 수 없으며 Npair(U_i)와 Nintra(U_i)도 마찬가지이다. 본 논문에서는 Npair, Ninter, Nintra, Nnon을 일컫기 위해 *PSNT(Pairwise Salience based Nominal Type)*라는 용어를 사용한다. 현저성을 가진 Npair, Ninter, Nintra를 일컫기 위해 Nsalient라는 용어를 사용한다. 다음 텍스트에서 이 PSNT의 네 가지 유형이 어떻게 정의되는 살펴보자.

표 3 문장간/내 현저성에 기반한 명사 n의 네 가지 유형

	$n = Cb(U_i)$	$n \neq Cb(U_i)$
$n = Cp(U_i)$	문장간/내에서 동시에 현저한 명사(pairwise salient nominal, 약어로 Npair)	문장내에서 현저한 명사 (intra-salient nominal, 약어로 Nintra)
$n \neq Cp(U_i)$	문장간에 현저한 명사 (inter-salient nominal, 약어로 Ninter)	현저하지 않은 명사 (non-salient nominal, 약어로 Nnon)

(텍스트 2)

U_1: 자귀는 우리나라의 전통적인 농기구이다.

⇨ Cf: 자귀 > 우리나라 > 농기구, Cp: 자귀

⇨ Nintra: 자귀, Nnon: 우리나라, 농기구

U_2: 자귀의 날은 절삭날이라고도 한다.

⇨ Cf: 날 > 자귀 > 절삭날, Cp: 날, Cb: 자귀, 전이: RET

⇨ Nintra: 날, Ninter: 자귀, Nnon: 절삭날

U_3: 자귀는 형태가 도끼와 비슷하게 생겼는데,

⇨ Cf: 자귀 > 형태 > 도끼, Cp: 자귀, Cb: 자귀, 전이: E-CON

⇨ Npair: 자귀, Nnon: 형태, 도끼

U_4: [자귀는(ø)] 크기에 따라 대자귀, 중자귀, 소자귀로 나누어진다.

⇨ Cf: 자귀 > 크기 > 대자귀 > 중자귀 > 소자귀, Cp: 자귀, Cb: 자귀, 전이: C-CON

⇨ Npair: 자귀, Nnon: 크기, 대자귀, 중자귀, 소자귀

U_5: 대자귀는 소자귀보다 볼이 얇다.

⇨ Cf: 대자귀 > 볼 > 소자귀, Cp: 대자귀, Cb: 대자귀, 전이: E-SSH

⇨ Npair: 대자귀, Nnon: 볼, 소자귀

$Cb(U_i)$와 $Cp(U_i)$는 여섯 가지 전이유형에 따라 Npair, Ninter, Nintra가 될 수 있으며, Cf-list의 나머지 명사는 Nnon에 대응된다. 밑줄이 그어져 있는 명사는 잠정적으로 영형의 생성 대상이 될 수 있는 구정보를 의미한다.

요약하면, 각 $PSNT(U_i)$는 그 현지성이 획득되는 발화의 범위에 따라 서로 다른 현저성의 강도를 가지고 있다. 이런 네 가지 유형에 대한 분류의 동기는 단순하면서도 직관적이다. 어떤 명사의 대용화 현상은 그 명사가 현재 발화 상태에서 얼마나 화자 또는 청자의 인지상에서 명확하게 자리잡고 있느냐로 설명될 수 있다. 또, 이는 인접한 발화 사이에서 그 명사의 현저성이 어떻게 변화하고 있는지의 여부를 통해서 판단하게 된다. 이 때 해당 명사가 인접한 발화 사이에서 그 현저성이 감소, 증대, 유지 될 수 있는데 이렇게 다른 현저성의 강도는 실제 텍스트의 대용화 현상을 설명할 수 있는 가장 기본적인 기준이 된다. 더불어, 기존의 중심화 이론에 기반한 영형 연구의 대부분은 CON의 Cb, 즉 Npair만을 고려했다는 점을 고려할 때 네 가지 명사 부류 각각에 대해 영형을 분석하는 것은 실제 텍스트 상에서 Cb뿐만 아니라 나머지 명사들에서 발생하는 영형을 분석하는 기회를 제공한다.

4. 코퍼스의 수집 및 간단한 분석

각 PSNT별로 영형 현상을 분석하고, 특히 발화들의 전이유형이 영형의 발생에 미치는 영향을 파악하기 위해 세 개의 장르 - 묘사문(한국 민속 박물관 웹 페이지에서 텍스트 추출), 뉴스 기사, 이야기(짧은 이솝 우화) - 로부터 총 93개의 텍스트를 수집하였다. 수작업으로 모든 문장을 발화 단위인 시제절 기준으로 분할하고, 영형 및 영형 외의 대명사에 대한 선행사(antecedent)를 찾는 대용 해결(reference resolution)을 처리하였다. 다음, 대용 해결된 각각의 일련의 시제절에 대해 구문 분석기를 통해 각 단어들의 구문 관계를 획득하였다. 이 중 명사를 대상으로 2장에서 정의한 Cf의 순위 매김에 따라 우선 순위를 정하였다. 다음, 인접한 시제절들을 대상으로 비용기반 중심화 이론을 적용하여, 각 발화에 대해 Cb, Cp, 전이유형을 자동으로 구하였다. 마지막으로, 모든 구명사들을 네 가지 PSNT로 구별하였다.

표 4는 수집된 텍스트의 기본적인 통계치들을 보여준다. 앞서 서두에서 언급한 것처럼, 한국어에서는 영형이 아닌 대명사의 발생이 두드러지지 않는다. 묘사문에서 모든 구명사의 3%(20/680), 모든 대명사의 6% 정도가 비영형 대명사로 실현되고, 전체 장르에서 구명사의 6%, 모든 대명사의 13% 정도가 비영형 대명사로 실현되었다. 이야기 장르에서 비영형 대명사의 발생이 다소 높은 편으로 구명사의 10%, 전체 대명사의 20% 정도가 비영형 대명사이다. 이는, 이야기 장르에서 동물성(animacy)을 가진 명사들이 자주 등장하고 이런 동물성을 가진 명사들은 비동물성을 가진 명사에 비해 영형이 아닌 대명사로 실현되는 경우가 잦기 때문이다. 하지만 본 논문에서는 텍스트 생성을 위한 하나의 과정으로 영형을 다루고, 텍스트 생성 시스템에서 생성될 텍스트의 성격은 이야기라기보다는 묘사문에 가깝기 때문에 비영형 대명사의 생성은 다루지 않는다. 지금부터 비영형 대명사와, 대명사화 되지 않고 본래의 명사 표현 그대로

표 4 수집된 텍스트의 정보

장르	묘사문			뉴스			이야기			전체		
대용형	N	P	Z	N	P	Z	N	P	Z	N	P	Z
구명사의 개수	344	20	316	173	21	146	291	65	245	808	106	707
발화의 개수		659			225			439			1323	
텍스트의 개수		53			20			20			93	

N: 본래 명사 표현 그대로, P: 비영형 대명사, Z: 영형 대명사 (단위: 개수)

표현된 명사들을 통틀어 '비영형 대용형'이라고 칭하겠다.

표 5는 장르별 각 PSNT에서 대용형의 분포를 보여준다. PSNT의 개수는 Npair(764), Ninter(419), Nnon(261), Nintra(186) 순으로 많다. Npair에서 61%, Ninter에서 33%, Nintra에서 29%, 그리고 Nnon에서 19%가 영형으로 발생되었다. 즉, Npair에서만 영형이, 나머지 PSNT에서는 비영형 대용형이 더 빈번히 발생한다. Npair에서 61%의 명사만이 영형으로 실현되었으므로, CON 또는 SSH에서의 Cb를 영형으로 생성하는 기존 방법들은 Npair에서 약 39%의 영형을 과다 생성(overgeneration)하는 문제점을 가짐을 알 수 있다.

수집된 텍스트에서 약 8%의 영형은 PSNT의 어느 유형으로도 설명할 수 없는 것이었다. 이 8%의 영형은 U_i의 관점에서 구명사가 아닌 것들, 즉 U_{i-1}에서 실현되지 않은 명사들에서 출현한 것으로, 이런 영형은 본 논문에서는 다루지 않는다. 이를 제외한 모든 영형에서의 PSNT의 분포가 표 6에 정리되어 있다. 표 5에서 얻어진 표 6에서, 영형의 66%(464/(464+55+136+52))는 Npair에서, 19%(136/(464+55+136+52))는 Ninter에서 실현된 것으로, 약 85%의 영형이 Npair와 Ninter에서 실현되었으며, 나머지 15%의 영형이 Nintra와 Nnon에서 실현되었다. 이러한 사실을 통해, CON 또는 SSH에서의 Cb를 영형으로 생성하는 기존 연구들이 다른 명사 부류에서 발생하는 약 34%의 영형을 처리할 수 없다는 사실을 확인하였다.

5. 선행 발화의 전이유형이 현 발화의 영형에 미치는 영향

이 장에서는, 영형 생성에 관여하는 여러 자질을 통해 학습된 영형 생성 모델의 성능과, 선행 발화의 전이유형만으로 학습된 영형 생성 모델의 성능을 비교하여, 선행 발화의 전이유형의 상대적인 중요성을 알아 본다.

5.1 다양한 자질들을 이용한 영형 생성 모델 구축

[14]는 한국어에서 영형 생성에 미칠 수 있는 12개의 자질을 가장 광범위하게 제안하였으며, 다양한 기계학습을 통해 영형 생성 모델의 성능을 최대화 하였다. 따라서, 본 논문에서는 [14]에서 제안한 12개의 자질(표 7에서 1~12)을 포함하고, 다음과 같은 4개의 자질(표 7에서 13~16)을 추가하여 총 16개의 자질을 적용하였다.

- (13) 평행성(parallelism): 선행사와 현재 대용형 처리를 위해 고려중인 명사의 구문 관계의 변화 [19,29]
- (14) 선행사와의 거리 [19,28,29]
- (15) 선행사의 구정보 여부 [19]
- (16) 연결어(connective) [30]

표 5 PSNT에서 대용형 분포 (단위: 개수)

	묘사문		뉴스		이야기		전체	
	N	Z	N	Z	N	Z	N	Z
Npair	114	189	66	95	120	180	300	464
Nintra	51	20	20	15	60	20	131	55
Ninter	116	85	47	21	111	30	274	136
Nnon	83	22	61	15	65	15	209	52

표 6 대용형에서 PSNT 분포 (단위: %)

	묘사문		뉴스		이야기		전체	
	N	Z	N	Z	N	Z	N	Z
Npair	31	60	34	65	34	73	33	66
Nintra	14	6	10	10	17	8	14	8
Ninter	32	27	24	14	31	12	30	19
Nnon	23	7	31	10	18	6	23	7

N: 비영형 대용형 (명사표현 그대로 + 비영형 대명사), Z: 영형 대명사

표 7 U_i의 구명사 n의 지시어 생성을 위해 고려하는 자질들

자질들	설명
1. p2_gr	U_{i-2}에서 n의 구문 관계
2. p1_gr	U_{i-1}에서 n의 구문 관계
3. gr	U_i에서 n의 구문 관계
4. modifee	n의 피수식어 여부
5. animacy	n의 동물성(animacy) 여부
6. p2_proform	U_{i-2}에서 n의 대용형
7. p1_proform	U_{i-1}에서 n의 대용형
8. p2_trans	n을 Nsalient(U_{i-2})로 가지는 U_{i-2}의 전이유형
9. p1_trams	n을 Nsalient(U_{i-2})로 가지는U_{i-1}의 전이유형
10. trans	U_i의 전이유형
11. equality(Cb(U_i), Cp(U_{i-1}))	Cb(U_i)와 Cp(U_{i-1})의 일치 여부
12. cost(U_i, U_{i-1})	U_i와 U_{i-1}사이의 추론 비용
13. parallelism	gr(n, U_{i-1})과 gr(n, U_{i-2})의 변화 관계
14. distance	n와 n의 선행사 사이의 거리
15. ant_state	n의 선행사의 정보가 구정보인지 아닌지
16. connective	U_{i-1}과 U_i사이의 연결어

표 8 각 기계 학습의 적용에 따른 PSNT에서의 영형 생성 모델의 성능

	Npair				Nintra				Ninter				Nnon			
	D	N	S	T	D	N	S	T	D	N	S	T	D	N	S	T
베이스라인	62.5	51.1	56.3	58.2	72.7	60.0	76.9	71.1	57.2	73.8	79.7	64.1	85.8	87.2	83.1	85.3
결정 트리 (Decision Tree)	82.5	93.1	64.8	79.5	72.7	70.0	76.9	77.8	66.2	76.2	78.4	72.9	92.8	83.3	83.8	89.9
핵밀도추정 (Kernel Density)	79.2	88.5	63.8	74.9	81.8	70.0	69.2	73.3	65.2	76.2	78.4	68.1	95.9	92.6	89.2	89.4
개체 중심 학습(K*)	77.9	87.4	69.3	74.4	81.8	70.0	69.2	71.1	65.2	78.6	79.7	69.1	95.9	92.6	89.2	91.5
로지스틱 회귀(Logistic Regression)	80.2	87.4	70.5	80.2	68.2	70.0	84.6	73.3	66.7	71.4	70.3	71.0	88.7	85.2	83.8	88.8
베이지안 분류기 (Naïve Bayesian)	79.5	95.1	72.7	77.0	77.3	80.0	84.6	80.0	69.7	76.2	77.0	72.2	93.8	85.2	89.2	88.3
SVM	*83.2*	*93.1*	*74.4*	*81.6*	*86.4*	*80.0*	*84.6*	*84.4*	*69.7*	*78.6*	*81.1*	*73.2*	*95.9*	*94.4*	*94.6*	*92.0*
교대성 결정 트리 (Alternating Decision Tree)	81.8	90.8	68.8	78.8	86.4	70.0	84.6	75.6	69.2	79.2	75.7	71.0	95.9	92.6	89.2	88.3
개체 중심 학습 (K-NN)	76.9	87.4	63.6	72.6	81.8	70.0	76.9	73.3	63.7	73.8	81.1	67.5	94.8	92.6	89.2	91.0
부스팅 (LogitBoosting)	83.2	92.0	69.9	79.0	81.8	70.0	84.6	80.0	68.2	73.8	79.7	72.6	95.9	90.7	89.2	91.5
스택킹 (Stacking)	82.5	93.1	67.0	79.9	72.7	70.0	76.9	75.6	65.7	76.2	79.7	72.6	92.8	88.9	83.8	89.9

D: 묘사문, N: 뉴스, S: 이야기 T: 전체 장르 (단위: %)

기존 연구에서 고려되었던 대부분의 자질들은, 표 7의 16가지의 범주에서 크게 벗어나지 않는다. '연결어'와 '동사 원인성'을 제외한 나머지 자질들은 실제 대용형 생성 모델 구축에 활용되었거나 코퍼스 분석을 통해 대용형의 생성에 영향을 미친다고 판단된 것들이다. '연결어'와 '동사 원인성'과 관련해, 어떤 대명사에 대한 선행사를 찾을 때 연결어와 동사 원인성이 중요하게 작용한다는 사실은 심리언어학(psycholinguistics) 분야에서 먼저 연구되었다[30-33]. [30]에서는 이런 사실을 토대로 U_{i-1}이 주어졌을 때 U_i의 주어를 결정하기 위한 규칙과 이때 U_i의 주어의 대용화에 관한 규칙을 제안하였다. 예를 들어, U_{i-1}에 이동 동사(transfer verb)가 있고, U_i와 U_{i-1}이 'so'로 연결되어 있을 때, U_{i-1}에 목적('goal')의 의미 관계를 가진 명사는 U_i에서 주어로 선호되며 이때 U_{i-1}과 U_i에서 주어가 일치하면, U_i에서 주어를 대명사화한다. 하지만 [30]의 이런 규칙들은 실제 코퍼스를 통해 평가되지 않았다.

학습 모델을 구축하기 위해 [14]에서 적용한 기계 학습 방법 외에 WEKA 3.4[2]에서 제공하는 총 11개의 기계 학습을 적용하여 최고 성능을 내는 알고리즘을 찾고자 하였다. 기계 학습을 위한 각 알고리즘들은, 학습 데이타의 수, 자질의 수, 자질의 중복성 등에 의해 서로 다르게 동작될 수 있어 그 효과가 달라질 수 있다. 본 논문에서는 기계 학습의 본질적인 비교 없이, 단순한 성능 비교를 통해 영형 생성에 적합한 학습 모델을 찾는 것에 그 목적을 둔다.

각 기계 학습에 대한 성능 평가는, 10-분할 교차 검증(10-fold Cross-Validation)을 총 10회 시행한 평균값으로 이루어 졌다. 표 8은 각각의 기계 학습 방법의 적용에 따른 각 PSNT에서 영형 생성 모델의 평균 성능을 보여준다. 베이스라인은, 각 PSNT에서 가장 많이 발생하는 대용형을 선택할 때의 정확률을 의미한다. 이는, Npair에서는 무조건 영형을 나머지 PSNT에서는 비영형 대용형을 선택할 때의 정확률이다. Nnon에서 베이스라인의 성능이 가장 높다. 이유는, 다른 PSNT에 비해 가장 많이 발생하는 대용형(여기서는 비영형 대용형)의 비율이 다른 대용형에 비해 가장 높기 때문이다. 표 8에서 보여지는 것처럼, 모든 학습 방법 중, SVM이 모든 장르, 모든 PSNT에서 평균적으로 좋은 성능을 낸다. 표 8의 SVM 학습에 의해 얻어진 정확률을 간단히 'SVM-전체-성능' 이라 부르고, 이를 나중에 선행 발화의 전이만을 고려한 영형 생성 모델(절 5.2)의 성능과 비교할 것이다.

5.2 발화의 전이유형을 이용한 영형 생성 모델 구축

표 9는 선행 발화 각각의 전이유형과 현행 발화의 전이유형만을 이용해 영형 생성 모델을 구축했을 때의 성능을 보여 준다. P2_trans는 U_{i-2}의 전이유형만을, p1_trans는 U_{i-1}의 전이유형만을, trans는 U_i의 전이유형만을 이용했을 때의 성능을 나타낸다. 영형 생성 모델의 규칙을 쉽게 설명하기 위해 결정 트리(DT)를 SVM과

2) WEKA는 뉴질랜드의 Waikato 대학에서 java로 개발된 기계 학습 소프트웨어로, 자연언어처리의 다양한 분야에 편리하게 적용할 수 있어, 기계 학습을 이용한 많은 논문에서 활용되고 있다. (http://www.cs.waikato.ac.nz/~ml/weka/index.html)

표 9 각각의 선행/현행 발화의 전이유형만을 고려했을 때 영형 생성 모델의 성능

	Npair				Nintra				Ninter				Nnon			
	D	N	S	T	D	N	S	T	D	N	S	T	D	N	S	T
svm-전체-성능	83	93	74	81	86	80	84	84	69	78	81	73	95	94	94	92
p2_trans (DT)	64	58	55	58	86	71	80	83	60	78	79	68	93	87	91	90
p1_trans (DT)	70	66	65	66	72	65	76	75	56	73	77	64	91	87	91	88
trans (DT)	72	62	66	68	72	69	76	71	57	73	77	64	-	-	-	-
p2_trans (SVM)	63	57	52	58	86	71	77	83	58	79	80	68	94	87	92	90
p1_trans (SVM)	69	67	65	66	73	66	69	73	51	71	78	65	92	87	92	88
trans (SVM)	72	61	66	67	73	68	69	71	51	74	78	65	-	-	-	-

D: 묘사문, N: 뉴스, S: 이야기 T: 전체 장르 (단위: %)

함께 적용해서 성능을 도출하였다. Trans에서 Nnon의 성능이 없는 이유는, U_i의 전이유형은 Cb와 Cp에 의해 결정될 뿐, Cb도 Cp도 아닌 Nnon(U_i)와는 아무 상관이 없기 때문이다.

Npair에서 장르를 통틀어 p2_trans의 성능이 p1_trans와 trans의 성능보다 낮은 반면, p1_trans와 trans는 p2_trans에 비해 장르별로 골고루 높은 성능을 보인다. 묘사문과 이야기 장르에서 trans의 성능이 가장 높은데, 이는 Npair(U_i)의 영형을 결정하기 위해 U_i의 전이유형이 C-CON, E-CON, C-SSH, E-SSH중에 어떤 것인지를 구별하는 것이 중요하다는 의미이다. Npair에서 결정 트리에 의해 도출된 trans의 모든 성능은 다음의 규칙에 의해 학습된 것으로 확인되었다. Trans(U_i)는 U_i의 전이유형을 의미한다.

(Trans-규칙) Trans(U_i)가 C-CON이거나 C-SSH일 때, Npair(U_i)를 영형으로 생성하라.

Nintra에서는 장르를 통틀어 p2_trans의 성능이 가장 높다. 또한, p2_trans에 의한 모든 결정 트리의 성능은 다음의 'p2-trans-규칙'에 의해 도출되었음을 확인하였다. 특히 묘사문에서는 이 규칙만을 이용하여 SVM-전체-성능에 도달하였으며, 또한 뉴스 장르에서 Ninter도 다음 규칙을 통해 SVM-전체-성능에 도달하였다.

(P2-trans-규칙, a) U_i의 임의의 명사가 Npair(U_{i-2})로 실현되고, 그 때 U_{i-2}의 전이유형이 C-CON이면, 그 명사를 영형으로 생성하라.

Nnon에서 p2_trans와 p1_trans의 정확률은 거의 비슷하고, 특히 묘사문과 전체 장르에서 다음 규칙에 의한 정확률은 SVM-전체-성능에 거의 근접한다.

(P2-trans-규칙, b) U_{i-2}의 전이에 상관없이 Nnon(U_i)가 Nsalient(U_{i-2})로 실현되었으면, Nnon(U_i)를 영형으로 생성하라.

요약하면, 각 PSNT에서 영형을 결정할 때, PSNT별로 서로 다른 이전 발화의 전이에 영향을 받는데, Npair는 U_{i-2}보다는 U_{i-1}과 U_i의 전이에 의해 더 영향을 받는 반면, Ninter, Nintra는 U_{i-1}, U_i보다는 U_{i-2}의 전이에 의해 더 영향을 받는다. 또한, 16가지의 자질로 학습된 영형 생성 모델의 성능, 즉 SVM-전체-성능과 비교했을 때, 특정 장르에서는, 선행 발화의 전이유형만을 이용했을 때의 Ninter, Nintra, Nnon의 성능이 SVM-전체-성능과 같거나, 거의 근접한 성능을 냄을 확인하였다. 지금부터는, Npair의 영형이 왜 trans와 밀접한 관련이 있는지, 또 Ninter, Nintra의 영형이 왜 p2-trans의 전이유형과 밀접한 관계에 있는지를 수집된 코퍼스 분석을 통해 살펴보도록 하자.

5.3 현 발화의 전이가 현 발화의 PSNT의 영형에 미치는 영향

수집된 텍스트에서 발생한 전이는 C-CON(36%), RET(21%), E-CON(9%), RSH(8%), C-SSH(5%), E-SSH(4%) 순으로 많고, Cb가 없는 발화는 17%였다. 표 10은, 각 전이의 PSNT에서 영형의 분포를 보여 준

표 10 각 전이의 PSNT에서 대용형의 분포

	DESC		NEWS		STORY		TOTAL	
	N	Z	N	Z	N	Z	N	Z
Npair in C-CON	25	75	47	53	34	66	31	69
Npair in E-CON	63	37	60	40	92	8	62	38
Ninter in RET	53	47	72	28	80	20	60	40
Nintra in RET	87	13	33	67	87	13	78	22
Npair in C-SSH	37	63	30	70	64	36	42	58
Npair in E-SSH	90	10	80	20	80	20	85	15
Ninter in RSH	76	24	76	24	81	19	78	22
Nintra in RSH	62	38	71	29	67	33	65	35

N: 비영형 대용형 (명사표현 그대로 + 비영형 대명사), Z: 영형 대명사, D: 묘사문, N: 뉴스, S: 이야기 T: 전체 장르 (단위: %)

다. 영형이 비영형 대용형보다 많이 발생한 전이는 C-CON과 C-SSH 두 전이이다. 또한 추론 비용이 비싼 전이보다 추론 비용이 싼 전이에서 영형이 발생하는 비율이 더 높은데, C-CON과 C-SSH에서는 각각 영형 비율이 69%, 58%인 반면, E-CON과 E-SSH에서는 이에 훨씬 못 미치는 38%와 15%이다. 특히 묘사문에서는 이런 특징이 두드러진다. C-CON과 C-SSH에서는 각각 영형 비율이 75%, 63%인 반면, E-CON과 E-SSH에서는 이보다 훨씬 작은 37%와 10%이다. 각 장르별로 영형 비율이 가장 높은 전이도, 묘사문은 C-CON에서 75%, 뉴스는 C-SSH에서 70%, 이야기는 C-CON에서 66%, 그리고 전체적으로 C-CON에서 69%로 전부 추론 비용이 싼 전이유형임을 알 수 있다. 이는 5.2절의 Trans-규칙과 일치하는 결과이며, 기존의 4가지 전이유형보다 추론 비용에 기반하여 나누어진 6개의 전이유형이 한국어의 영형을 설명하는데 효과적임을 알 수 있다.

5.4 선행 발화의 전이가 현 발화의 PSNT의 영형에 미치는 영향

표 11은 전체 장르에서 PSNT(U_i)의 대용형에 대해 U_{i-1}과 U_{i-2}의 전이유형 분포를 보여 주고 있다. PSNT(U_i)가 이전 발화 각각에서 Nsalient가 아니라 Nnon으로 실현될 때는 이전 발화의 전이유형과는 아무런 의미가 없으므로 표 11의 이전 발화에서 전이 분포는, PSNT(U_i)가 이전 발화 각각에서 Nsalient로 실현된 경우를 토대로 구해진 것이다. 예를 들어, <Npair, U_{i-2}, Z, C-CON>의 조건에서 51%는, Npair(U_i)가 영형이고 그것이 U_{i-2}에서 Npair일 때 U_{i-2}의 전이 중 51%가 C-CON임을 의미한다.[3)]

표 11의 데이타를 분석하기 전에 먼저 몇 가지 용어를 정의한다.

- SEPn (Salient Entity Probability): PSNT(U_i)가 U_{i-n}에서 Nsalient(U_{i-n})로 실현될 확률
- N-PSNT: 비영형 대용형으로 실현된 PSNT
- Z-PSNT: 영형 대용형으로 실현된 PSNT
- Z-RSEP: N-PSNT의 SEP에 대한 Z-PSNT의 SEP 비율 (이전 발화에서 Nsalient로 실현된 PSNT(U_i)가 영형으로 실현될 확률)

그럼, 지금부터 표 11로부터 도출되는 다음의 결과들을 살펴보자.

1) Z-PSNT의 SEP은 항상 N-PSNT의 SEP보다 모든 경우에서 다 크다. 이는 영형으로 실현된 PSNT가 그렇지 않은 것보다 더 자주 이전 발화에서 Nsalient로 실현됨을 의미한다.. 이런 결과는 당연한데, 이전 발화에서 Nsalient로 실현된 것들은 그렇지 않은 것에 비해 화자(또는 청자)의 인지 속에 명확하게 자리잡혀 있어 생략하더라도 담화 또는 문맥상에서 회복 가능하기 때문이다.
2) Nintra와 Ninter의 경우, U_{i-2}에서 Z-RSEP이 U_{i-1}에서 Z-RSEP보다 높은 반면, Npair에서는 역으로 U_{i-1}에서 Z-RSEP이 U_{i-2}에서 Z-RSEP보다 높다. 다시 말하면, Ninter(U_i)와 Nintra(U_i)는 Nsalient(U_{i-1})로 실현되었을 때보다, Nsalient(U_{i-2})로 실현되었을 때 영형이 되는 확률이 높다. 반면, Npair(U_i)은 그것이 Nsalient(U_{i-2})로 실현되었을 때보다 Nsalient (U_{i-1})로 실현되었을 때 영형이 되는 확률이 더 높다. 특히, 이런 특징은 Nintra에서 더 두드러지게 나타나는데, Nintra(U_i) 가 Nsalient(U_{i-2})로 실현되었을 때 영형이 되는 확률은 73%로 다른 것에 비해서 높다. Nnon 관해서는, U_{i-2}에서 Z-RSEP이 U_{i-1}에서 Z-RSEP보다 약간 높으며, Nnon(U_i)가 Nsalient(U_{i-2})로 실현되었을 때 영형이 될 확률은 89%로 역시 높다.
3) U_{i-2}에서 전이유형은 Nintra와 Ninter의 영형과 밀접한 관련이 있다. 영형으로 실현된 Npair(Z-Npair)는 이전 발화에서 특정한 전이유형에 대한 선호도나 나타나지 않는다. Z-Npair 이 경우, 이전 발화에서 C-CON의 비중이 다른 전이에 비해 높긴 하지만,

3) 전이가 C-CON인 임의의 발화 U_n에서 Esalient(U_n)중 존재할 수 있는 것은 Npair(U_n)뿐이므로, trans(U_{i-2})가 C-CON일 때 Npair(U_i)가 Esalint(U_{i-2})와 같다는 것은 Npair(U_{i-2})와 같다는 의미이다.

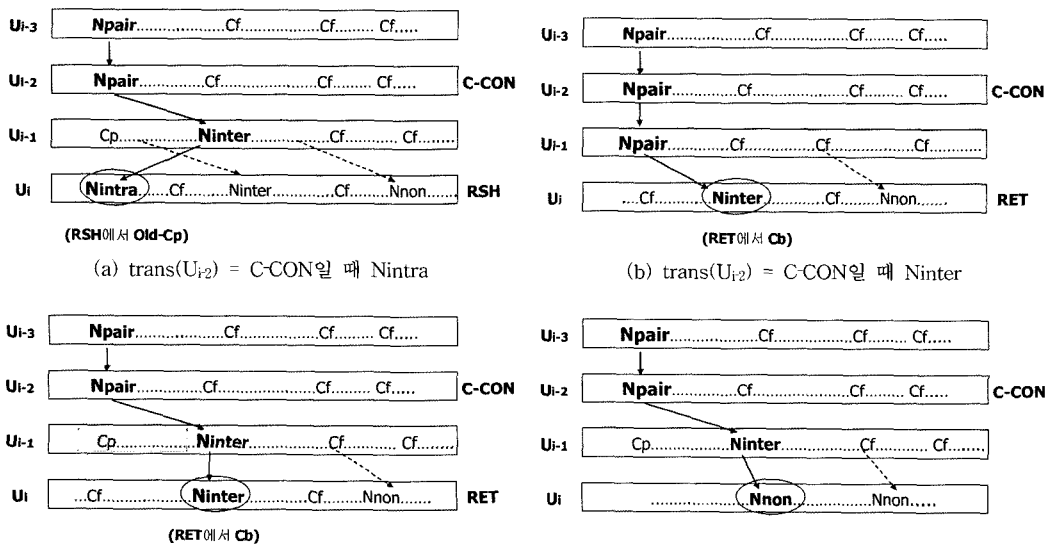
표 11 전체 장르에서 PSNT(U_i)의 대용형에 대한 이전 발화의 전이 분포

	Npair				Ninter				Nintra				Nnon			
	U_{i-2}		U_{i-1}		U_{i-2}		U_{i-1}		U_{i-2}		U_{i-1}		U_{i-2}		U_{i-1}	
U_i에서 PSNT의 대용형	N	Z	N	Z	N	Z	N	Z	N	Z	N	Z	N	Z	N	Z
C-CON	44	51	42	49	23	53	28	29	0	56	0	0	20	21	0	0
E-CON	14	9	11	13	15	9	13	10	13	0	0	0	10	11	0	0
RET	25	24	27	20	39	25	27	34	25	13	64	56	20	37	60	79
C-SSH	4	3	4	5	4	5	6	3	25	13	0	0	10	16	0	0
E-SSH	2	5	3	5	0	2	5	5	0	0	0	0	10	5	0	0
RSH	11	8	13	8	19	6	21	19	37	18	36	44	30	11	40	21
SEP	51	54	69	86	39	58	65	83	25	69	44	69	6	49	8	49
RSEP (SEP의 비율)	49	51	45	55	40	60	44	56	27	73	39	61	11	89	14	86

N: 비영형 대용형 (명사표현 그대로 + 비영형 대명사), Z: 영형 대명사 (단위: %)

C-CON의 비중이 N-Npair의 경우에 비해 큰 차이가 나지 않는다. 반면에, Z-Nintra와 Z-Ninter은 U_{i-2}에서 C-CON이 다른 전이에 비해 선호도가 두드러진다. Nintra가 U_i에서 영형이면서 Nsalient(U_{i-2})로 실현되었을 때, U_{i-2}에서 발생한 전이의 56%가 C-CON이었던 반면, 같은 조건에서 그것이 비영형 대용형으로 실현되었을 때는 U_{i-2}에서 C-CON이 한 번도 발생하지 않았다. Ninter도 마찬가지로, Ninter가 U_i에서 영형이면서 Nsalient(U_{i-2})로 실현되었을 때, U_{i-2}에서 발생한 전이의 53%가 C-CON이었던 반면에, 같은 조건에서 그것이 비영형 대용형으로 실현되었을 때는 U_{i-2}에서 C-CON이 23%정도로 낮게 발생하였다. 표 11은 전체 장르에 대해서만 표기하였지만, 장르별로 조사한 결과, 이야기의 Z-Ninter인 경우 U_{i-2}에서 C-CON이 75%인 반면, N-Ninter경우 약 27%, 뉴스의 Z-Ninter인 경우 C-CON이 약 50%인 반면, N-Ninter에서는 한번도 C-CON이 발생하지 않았다. 요약하면, Z-Nintra와 Z-Ninter는 전이가 C-CON인 U_{i-2}에서 Nsalient(U_{i-2})로 실현된 경우가 많다. 이는, *영형 생성의 관점에서*, Nintra(U_i) 또는 Ninter(U_i)가 Nsalient(U_{i-2})로 실현되었으며 그 때 U_{i-2}의 전이가 C-CON이면, Nintra(U_i) 또는 Ninter(U_i)는 영형으로 생성될 가능성이 높다는 것을 의미한다. 이것이 바로 앞서, 결정트리를 통해 구축되었던 영형 생성 모델의 규칙, p2-trans-규칙으로 드러난 것이다.

Trans(U_n) = C-CON인 임의의 U_n에 대해, Nsalient(U_n)은 Npair(U_n)을 의미하고, 이 때 Npair(U_n)은 C-CON의 정의에 의해 Npair(U_{n-1})과 같다. Npair(U_n)이 U_{n+1}에 실현될 경우, Npair(U_{n+1})또는 Ninter(U_{n+1})중의 하나가 된다. 즉 C-CON의 Npair는, 크게는 현 발화와 그것의 전후 발화, 즉 인접한 3개의 발화에서, 작게는 현 발화와 직전의 발화, 즉 인접한 2개의 발화에서 문장내에서 가장 현저하게 실현됨을 보장한다. 이것을 염두에 두고, Nintra와 Nintra의 영형에 U_{i-2}의 C-CON이 왜 결정적인 역할을 하게 되는지 살펴보자. 그림 1(a)는 U_i의 전이유형이 RSH인 경우, Nintra(U_i)가 trans(U_{i-2}) = C-CON인 U_{i-2}에서 Nsalient(U_{i-2})로 실현되었을 때, 인접한 4개의 발화에서 Nintra(U_i)의 변화 경로를 보여준다. Nintra(U_i)가 U_{i-1}에서는 Npair가 아닐지라도 그 이전 두 발화 U_{i-2}와 U_{i-3}에서 연속적으로 Npair인 경우, Nintra(U_i)는 이 두 발화를 통해 이미 화자 또는 청자의 담화 인식 상에서 화제로 자리잡을 수 있다. 때문에, 현재 발화상에서 영형으로 생성될 가능성이 높다. 문장간/내 *현저성의 관점에서 다시 해석해 보면, 현재 발화상에서 가장 현저하지만 직전 발화에 의해 현저하지 않은 명사가, 그 이전의 두 발화에서 문장내 현저성과 문장간 현저성을 둘 다 가지는 경우 그것은 영형으로 생성될 가능성이 높다.* 다음 텍스트에서 그림 1(a)에 대응되는 RSH의 Nintra(U_i)의 생략을 살펴보자.



(a) trans(U_{i-2}) = C-CON일 때 Nintra

(b) trans(U_{i-2}) = C-CON일 때 Ninter

(c) trans(U_{i-2}) = C-CON일 때 Ninter

(d) trans(U_{i-2}) = C-CON일 때 Nnon

각 발화에서 가장 왼쪽에 있는 명사가 가장 현저하고, 오른쪽으로 갈수록 현저성이 떨어지며, 가장 왼쪽에 위치한 명사는 Cp가 된다고 가정한다. 음영이 들어간 부분은 다음 발화에서 실현될 수 없는 부분을 의미한다.

그림 1 U_{i-2}가 C-CON일 때 Ninter, Nintra, Nnon의 변화

(텍스트 3)

U_1: 향갑은 삼국시대부터 사용되었는데,

⇨ Cf: 향갑 > 삼국시대 > 사용, Cp: 향갑

⇨ Nintra: 향갑, Nnon: 삼국시대, 사용

U_2: [향갑은(ø)] 주로 금사로 만들어졌다.

⇨ Cf: 향갑 > 금사, Cp: 향갑, Cb: 향갑, 전이: C-CON

⇨ Npair: 향갑, Nnon: 금사

U_3: 향갑의 주 용도는 향을 보관하는 것이다.

⇨ Cf: 용도 > 향 > 향갑 > 보관, Cp: 용도, Cb: 향갑, 전이: RET

⇨ Ninter: 향갑, Nintra: 용도, Nnon: 보관

U_{4-1}: 또, [향갑은(ø)] 노리개의 용도로도 썼다.

⇨ Cf: 향갑 > 노리개 > 용도, Cp: 향갑, Cb: 용도, 전이: RSH

⇨ Ninter: 용도, Nintra: 향갑, Nnon: 노리개

U_{4-2}: 향은 위급 시 비상약의 용도로도 썼다.

⇨ Cf: 향 > 비상약 > 용도, Cp: 향, Cb: 용도, 전이: RSH

⇨ Ninter: 용도, Nintra: 향, Nnon: 비상약

위 텍스트의 전체 주제는 '향갑'으로, U_1에서 U_2까지 화제가 '향갑'으로 유지 되다가 U_3에서 '용도'로 바뀌고 U_{4-1}에서 다시 '향갑'으로 돌아 온다. 이 때, U_{4-1}에서 '향갑'의 생략은 자연스럽게 느껴지는데, 바로 직전 발화에서 '향갑'이 문장내에서 가장 현저하지 않지만 그것이 C-CON인 U_2에서 Npair로 실현되어 U_1과 U_2에서 화자의 인식 상태에 문장간/문장내에서 동시에 가장 현저한 명사로 자리잡았기 때문이다. U_{4-1} 대신에 U_{4-2}를 고려해 보자. U_{4-2}에서 '향'은 U_{4-1}의 '향갑'과 마찬가지로 RSH의 Nintra이긴 하지만, 이 때 '향'의 생략은 자연스럽지 않은데, 이는 '향'이 이전 발화에서 한번도 현저하지 못했기 때문이다.

Ninter에서도 이와 같은 방식으로 U_{i-2}에서 C-CON의 두드러짐 현상을 설명할 수 있다. 그림 1(b)와 그림 1(c)는 RET의 Ninter(U_i)가 trans(U_{i-2}) = C-CON인 U_{i-2}에서 Npair(U_{i-2})로 실현되었을 때, 인접한 4개의 발화에서 Ninter(U_i)의 변화 경로를 보여 준다. 그림 1(b)에서 Ninter(U_i)는 이전 세 발화에서 연속적으로 문장간/내 현저성을 다 가지고 있고 그 결과 화자 또는 청자의 머리 속에 이전 발화를 통해 이미 화제로 자리 잡았다. 때문에, 현 발화에서 가장 현저하지 않더라도 영형으로 생성될 가능성이 높다. 다음 텍스트에서 그림 1(b)에 대응되는 RET에서 Ninter(U_i)의 생략을 살펴보자.

(텍스트 4)

U_1: 쳇다리는 간장을 걸러낼 때 체와 그릇 사이에 받쳐놓는 기구로,

⇨ Cf: 쳇다리 > 간장 > 체 > 그릇 > 기구, Cp: 쳇다리

⇨ Nintra: 쳇다리, Nnon: 간장, 체, 그릇, 기구

U_2: [쳇다리는(ø)] 쳇발이라고도 한다.

⇨ Cf: 쳇다리 > 쳇발, Cp: 쳇다리, Cb: 쳇다리, 전이: C-CON

⇨ Npair: 쳇다리, Nnon: 쳇발

U_3: [쳇다리는(ø)] 주로 나무로 만든다.

⇨ Cf: 쳇다리 > 나무, Cp: 쳇다리, Cb: 쳇다리, 전이: C-CON

⇨ Npair: 쳇다리, Nnon: 나무

U_4: [쳇다리의(ø)] 중심부분에 홈이 있어,

⇨ Cf: 쳇다리 > 홈 > 중심부분, Cp: 홈, Cb: 쳇다리, 전이: RET

⇨ Ninter: 쳇다리, Nintra: 홈, Nnon: 중심부분

위 텍스트의 전체 주제는 '쳇다리'로 U_1에서 U_3까지 문장내 가장 두드러진 명사가 '쳇다리'로 유지 되다가 U_4에서 '홈'으로 바뀌면서 U_4에서 '쳇다리'는 문장내에서 가장 현저한 명사가 아니다. 하지만 U_1에서 U_3까지 '쳇다리'가 화제로 유지되어서 화자의 인식 상태에 가장 현저한 명사로 자리잡았기 때문에 U_4에서 '쳇다리'의 생략이 자연스럽다. 3장의 텍스트 2에서 U_2의 Ninter와 비교해 볼 때, 같은 RET의 Ninter라 하더라도 텍스트 2의 U_2에서 생략은 어색하다. 이유는, 그 Ninter가 앞선 단 하나의 발화에서만 문장내에서 가장 현저한 명사로 나타났기 때문에, 화자의 인식 상태에 충분히 현저하게 자리잡지 못했기 때문이다.

그림 1(c)에서, Ninter의 변화 경로는 그림 1(a)와 동일한데, 1(b)와 비교해 봤을 때 이전 발화에서의 두드러짐은 약하지만, 역시 1(a)와 같은 이유로 영형의 가능성이 높다. 그림 1(d)는 Nnon의 변화 경로를 보여 준다. 역시 1(a), 1(c)와 동일하지만, Nnon(U_i)가 Ninter(U_i), Nintra(U_i)에 비해 현저성이 떨어지기 때문에 1(a), 1(c)보다는 영형 생성 가능성이 낮을 수 있다. 그렇지만, 1(d)와 다른 경로를 갖는 Nnon(U_i)에 비해 영형이 될 가능성은 높다.

요약하면, U_i에서 문장간 현저성을 갖지 않거나 문장내 현저성을 갖지 않는 명사들의 영형은, 선행하는 두번째 발화 U_{i-2}의 전이에 절대적인 영향을 받는다. 그런 명사들이 U_{i-2}에서 Nsalient로 실현되고 그 때 U_{i-2}의 전이유형이 C-CON이면, 그 명사들은 영형으로 생성될 가능성이 높다. 그 이유는, 현 발화에서 또는 직전 발화에서 문장간 현저성을 갖지 않거나 문장내 현저성을 갖지 않는 명사라 하더라도, 그것이 C-CON인 U_{i-2}에서 Nsalient로 실현되어, 이전 발화를 통해 화자의 인식 속에 현저한 명사로 자리잡았기 때문이다. 기존 연구가

RET과 RSH에서 Cb(Ninter)와 구명사인 Cp(Nintra), 또 전이에 관계없이 Cb, Cp가 아닌 구명사들의 대용형을 처리하지 못했다는 점에서 위의 코퍼스 분석 결과와 함께, 이런 명사들의 영형 생성을 위한 p2-trans-규칙은 의미 있는 결과라 할 수 있겠다.

6. 결 론

본 논문에서는, 자연스러운 영형 대명사를 생성하기 위해 비용기반 중심화 이론의 6가지 전이유형을 이용해 선행 발화의 전이에 초점을 맞추어 영형을 분석하였다. 중심화 이론에 기반해 영형으로 실현될 수 있는 명사를 4가지 유형(Npair, Ninter, Nintra, Nnon)으로 정의한 다음, 각 유형별로 코퍼스를 분석하고 영형 생성 모델을 구축하였다. 그 결과, 기존에 중심화 이론으로 설명할 수 없었던 영형 명사들을 이전 두번째 발화의 전이유형으로 설명할 수 있음을 밝히고, 이런 명사들의 효과적인 영형 생성을 위한 p2-trans-규칙을 발견하였다. 또한, 선행 발화의 중심 전이를 이용한 영형 생성 모델의 성능과 다양한 자질을 적용한 모델의 성능을 비교하여, 전자의 성능이 후자의 성능과 일치하거나 거의 근접함을 확인하였다.

본 논문에서는 이전 두발화의 전이유형만을 고찰하였으나 그 범위를 더 확대하여 연구해 볼 필요가 있다. 또한 본 논문의 결과를 일본어와 같은 한국어와 비슷한 유형의 다른 언어에 적용해서 그 효과를 입증하는 것도 향후 과제로 남아 있다.

참 고 문 헌

[1] M. Ariel, "Accessing noun phrase antecedents," Routledge, London (Croom Helm Linguistics series), 1990.

[2] W. Chafe, Discourse, consciousness, and time, University of Chicago Press, Chicago, IL; London, 1994.

[3] E.F. Prince, Towards a taxonomy of given-new information, pp.223-225, in P.Cole(ed.), Radical Pragmatics, Academic Press, New York, N.Y., 1981.

[4] J.K. Gundel, N. Hedberg, and R. Zacharski, "Cognitive status and the form of referring expressions in discourse," Proc. Language, vol.69, no.2, pp.279-307, 1993.

[5] M.A.K. Haliday, "Notes on transitivity and theme in English," Proc. Linguistics, vol.3, no.2, pp.199-244, 1967.

[6] B.J. Grosz, A.K. Joshi, and S. Weinstein, "Centering: a framework for modeling the local coherence of discourse," Proc. Computational Linguistics vol.21, no.2, pp.203-225, 1995.

[7] H. Cheng, "Experimenting with the interaction between aggregation and text planning," Proc. ANLP-NAACL Student Research Workshop, USA. 2000.

[8] V. Mittal, J. Moore, G. Carenini, and S. Roth, "Describing complex charts in natural language: a caption generation system," Proc. Computational Linguistics, Special issue on Natural Language Generation, vol.24, no.3, pp.431-467, 1998.

[9] R. Kibble and R. Power, "Using centering theory to plan coherent texts," Proc. 12th Amsterdam colloquium, pp.187-192, 1999.

[10] R. Kibble and R. Power, "An integrated framework for text planning and pronominalization," Proc. 1st International Natural Language Generation, Mitzpe Ramon, Israel, pp.77-84, 2000.

[11] Y. T. Mitsuko, M. Fujiwara, and T. Aizawa, "Centering as an anaphora generation algorithm: a language learning aid perspective," Proc. 6th Natural Language Processing Pacific Rim, Tokyo, Japan, pp.557-562, 2001.

[12] R. Prasad, "Constraints on the generation of referring expressions, with special reference to Hindi," U of Pennsylvania, PhD Thesis, 2003.

[13] J.E. Roh and J.H. Lee, "An empirical study for generating zero pronoun in Korean based on Cost-based Centering Model," Proc. Australasian Language Technology Association, Melbourne, Australia, pp.90-97, 2003.

[14] J.E. Roh and J.H. Lee, "Generation of natural referring expressions by syntactic information and Cost-based Centering Model," Journal of KISS: Software and Applications, vol.21, no.12, pp.1649-1659, 2004.

[15] M.Y. Kim, "The centering of Korean discourse," Seoul National University, M.S. Thesis, 1994.

[16] M.K. Kim, "Conditions on deletion in Korean based on information packaging," Proc. Discourse and Cognition, vol.1, no.2, pp.61-88, 1999.

[17] B.R. Ryu, "Centering and zero anaphora in the Korean discourse," Seoul National University, M.S. Thesis, 2001.

[18] M.K. Kim, "Zero vs. overt NPs in Korean discourse: a centering analysis," Korean Journal of Linguistics, vol.28, no.1, pp.29-49, 2003.

[19] R. Henschel, H. Cheng, and M. Poieso, "Pronominalization revisited," Proc. 18th International Conf. on Computational Linguistics, Saarbruecken, pp.306-312, 2000.

[20] M. Strube and U. Hahn, "Functional centering: grounding referential coherence in information structure," Proc. Computational Linguistics, vol.25, no.3, pp.309-344, 1999.

[21] M. Walker, M. Iida, and S. Cote, "Japanese discourse and the process of centering," Proc. Computational Linguistics, vol.20, no.2, pp.193-232,

1994.
[22] R.J. Passonneau, "Getting and keeping the center of attention," In Bates, M. and Weischedel, R.R., editors, Challenges in Natural Language Processing, Cambridge University Press, pp.179-227, 1993.
[23] D. Byron and A. Stent, "A preliminary model of centering in dialog," Proc. 36th Annual Meeting of the Association for Computational Linguistics, Montreal, Canada, pp.1475-1477, August. 1998.
[24] B. Di Eugenio, "Centering in Italian," In Walker, M.A., Joshi, A.K., and Prince, E.F., editors, Centering Theory in Discourse, chapter 7, pp.115-138, Oxford, 1998.
[25] M. Kameyama, "Intra-sentential centering: a case study," In Walker, M.A., Joshi, A.K., and Prince, E.F., editors, Centering Theory in Discourse, chapter 6, pp.89-112, Oxford, 1998.
[26] J.R. Tetreault, "A corpus-based evaluation of centering and pronoun resolution," Proc. Computational Linguistics, vol.2, no.4, pp.507-520, 2001.
[27] M. Poesio, R. Stevenson, H. Cheng, B.D. Eugenio, and J. Hitzeman, "Centering: a parametric theory and its instantiations," Proc. Computational Linguistics, vol.30, no.3, pp.309-363, 2004.
[28] K.F. McCoy and M. Strube, "Generating anaphoric expressions: pronoun or definite description?," Proc. Workshop on the Relation of Discourse/Dialogue Structure and Reference, held in conjunction with Annual Meeting of the Association for Computational Linguistics, pp.63-71, 1999.
[29] M. Strube, and M. Wolters, "A probabilistic genre-independent model of pronominalization," Proc. 1st Meeting of the North American Chapter of the Association for Computational Linguistics, Seattle, WA, USA, pp.18-25, April. 2000.
[30] R. Stevenson, "The role of salience in the production of referring expressions," In Kees van Deemter and Rodger Kibble(eds), Information Sharing, CSLI Publications, 2002.
[31] B.K. Lee, "The effect of verb causality upon pronoun disambiguation in sentences with causal, adversative, and conjunctive relation," Department of Psychology Graduate School of Seoul National University, M.S Thesis, 1989.
[32] H.E. Yun, "The sentence reading time and the comprehension of anaphoric pronouns as a function of the causality implicit in verbs," Department of Psychology Graduate School of Seoul National University, M.S Thesis, 1984.
[33] H.R. Kwon, "The effects of semantic factors upon comprehension of relative-clause sentence: the semantic factors of co-reference and cause," Department of psychology graduate school of Seoul National University, M.S. Thesis, 1988.

노 지 은
2000년 2월 부산대학교 컴퓨터공학과 학사. 2000년 3월~현재 포항공과대학교 컴퓨터공학과 석박사통합과정. 관심분야는 텍스트 생성, 기계 번역, 자연언어처리, 한국어처리

나 승 훈
2001년 2월 아주대학교 컴퓨터공학과 학사. 2003년 2월 포항공과대학교 컴퓨터공학과 석사. 2003년 3월~현재 포항공과대학교 컴퓨터공학과 박사과정. 관심분야는 정보검색, 자연언어처리, 한국어처리, 기계 번역

이 종 혁
정보과학회논문지 : 소프트웨어 및 응용 제 32 권 제 9 호 참조