구문 관계 지식 추출을 위한 코퍼스 정규화에 대한 연구*

조 정미, 조 영환, 김 길창
한국과학기술원, 전산학과

A Corpus Formalization for Extracting the Syntactic Relations

Jeong Mi Cho, Young Hwan Cho, Gil Chang Kim
Dept. of Computer Science, KAIST

요 약

대량의 코퍼스를 이용해 여러 가지 일반적인 언어 현상을 관찰하고, 언어 지식을 자동으로 획득하여 자연 언어 처리의 여러 분야에 이용하는 등의 연구가 활발히 진행되고 있으며, 이에 따라 코퍼스에 대한 필요성이 날로 증가하고 있다. 코퍼스에서 추출할 수 있는 유용한 지식 중의 하나가 구문 관계 지식이다. 그러나 한국어에 자주 나타나는 격이동이나 생략 현상, 복합어 의 이형태 등은 정확한 지식 획득을 어렵게 할 뿐 아니라 자료 희귀 문제를 더욱 심화시킨다. 본 논문에서는 한국어의 문법적인 특징을 반영한 코퍼스 정규화에 의해 이러한 문제를 해결하고자 한다.

1. 서론

자연언어 처리 분야에서 실험적인 방법이 확산되면서 코퍼스로부터 필요한 언어 지식을 자동적으로 획득하는 일이 많아지고 있다[Marcus 1993]. 코퍼스를 이용한 자연언어 처리에 관한 많은 연구들은 단어 자체를 언어 지식으로 사용하기보다는 단어가 가지고 있는 품사나 그 밖의 여러 형태의 자질 정보를 이용한다. 이러한 방법론에서는 학습 대상이 되는 코퍼스에 여러 형태의 자질 정보를 추가하는 코퍼스 가공 작업이 필요하다. 코퍼스는 이 가공의 정도에 따라 텍스트 코퍼스, 형태소 분석이 된 품사 부착 코퍼스, 문장 단위로 구문 구조가 분석된 구문 구조 부착 코퍼스 혹은 의미 표지 부착 코퍼스 등으로 분류 할 수 있다. 자동으로 할 수 있는 가공이 제한되어 있으므로, 많은 가공이 필요한 코퍼스일 수록 그 코퍼스를 얻기까지 많은 노력이 요구된다.

구문 관계 지식이란 문장의 성분간의 문법적 관계에 대한 지식으로, 여러 자연언어 처리 시스템에서 이용된다. 코퍼스로부터 구문 관계 지식을 자동으로 획득하기 위해서는 구문 구조에 대한 정보가 포함된 구문 구조 부착 코퍼스가 필요하다. 그러나 구문 구조 부착 코퍼스 구축은 매우 어려운 일이다.

한국어는 의미를 갖는 실질 형태소의 문법적 기능이 형식 형태소인 조사나 어미 등에 의해 결정된다. 따라서 문장의 구성 성분간의 구문 관계는 조사에 의해 표현되고, 그렇기 때문에 구문 분석을 거치지 않은 품사 부착 코퍼스로부터도 조사와 단어로 표현되는 구문 관계 지식을 획득할 수 있다. 예를 들면, "철수가 밥을 먹는다."라는 문장에서는 '철수'는 주격 조사 '가'에 의해 동사 '먹다'와 주어 관계를 이루며 '밥'은 목적격 조사 '을'에 의해 목적어 관계를 이룬다.

본 논문에서는 한국어 품사 부착 코퍼스로부터 구문 관계 지식을 추출하고자 할 때 필요한 코퍼스 정규화에 대하여 설명한다.

2. 코퍼스 정규화란?

자동으로 구문 관계 지식을 획득하고자 할 경우, 다음과 같은 고려 사항들이 있다.
첫째, 격이동에 의해 조사 변경 현상이 발생한다. 격이동이란 본용언에 보조 용언이나 접사, 또는 보조 용언 상당어구 등이 덧붙여지면서 조사의 역할이 바뀌거나 추가적인 성분을 요구하게 되는 현상을 의미한다[최기선,1989]. 다음의 예를 보자.

*본 연구는 과학재단의 목적 기초 과제 "한국어 이해에 나타나는 중의성 문제 처리 모델에 관한 연구"의 부분 지원을 받은 것입니다.

1. 얼음이 녹는다.

2. 얼음을 녹게 하다.

문장 1)에서 동사 '녹다'는 녹는 대상을 요구하는 자동사이다. 문장 2)는 동사 '녹다'에 사동 보조 용언 '게 하다'가 덧붙여져 문장의 양상이 사동화 된 것이며, 녹는 대상에 대한 조사가 '이'에서 '을'로 변하였다. 즉 문장이 사동화되면서 대상을 나타내는 조사가 바뀐 것이다. 이런 현상을 격이동이라고 한다. 문장 2)에서 사동화를 고려하지 않고 선택 제한 지식을 추출하면, [녹다-(목적어)-얼음]과 같이 자동사인 '녹다'에 대해 목적어가 추출되는 잘못된 결과를 초래한다. 따라서 조사로 표현되는 선택 제한 지식을 추출하고자 할 경우에는 격이동 현상을 반드시 고려하여야 한다.

둘째, 복합어의 이형태 표현 현상이 발생한다. 단어 형성이란 새로운 단어를 만들어 내는 것을 뜻하며 이 때 만들어진 단어를 복합어라고 한다. 즉, 이것은 어근과 어근의 결합, 어근과 접사의 결합, 어근 창조 등의 방법으로 새로운 단어를 만들어 내는 것이다. 복합어들은 그 쓰임에 있어서 여러 가지 이형태를 갖는다. 예를 들면, 동작성 명사와 동사 파생 접미사 '하다'가 결합하여 동사가 되는 경우, 이것은 한 단어 동사로 나타날 수도 있고 목적격 관계로 나타날 수도 있으며 혹은 목적격 관계를 나타내는 조사가 생략되어 나타날 수도 있다. 즉, '준비'라는 동작성 명사에 '하다'라는 동사 파생 접미사가 결합된 경우, '준비하다', '준비를 하다', '준비 하다', 이 세 가지 형태가 모두 코퍼스에 나타날 수 있다. 복합어의 이형태는 지식 획득시 자료 부족 현상을 심화시킨다. 이형태를 한 가지 형태로 통일한다면 자료 부족 현상을 어느 정도 감소시킬 수 있다.

세째, 둘 이상의 문장이 결합하여 더 큰 문장을 만드는 경우, 선행문과 후행문의 공통된 성분의 생략 현상이 발생한다. 성분의 생략 현상 역시 지식 획득시 자료 부족 현상을 심화시키다. 따라서 생략된 성분을 명확하게 알 수 있는 경우는 생략된 성분을 복구하여 코퍼스로부터 풍부한 지식을 얻을 수 있도록 한다.

위와 같은 현상에 대한 처리는 다분히 어휘에 의존적이기 때문에 코퍼스에 대한 구문 분석으로도 해결하기가 어렵다. 본 논문에서는 한국어의 문법적 특징에 맞게 이런 현상들에 적절한 처리를 하는 과정을 코퍼스의 정규화라고 정의한다. 코퍼스 정규화란 코퍼스의 내용에 대한 수정이 아니라 코퍼스로부터 지식을 획득할 때 적용되는 규칙을 의미한다. 코퍼스의 정규화 범위는 다음과 같다.

- 보조 용언에 의한 격이동의 복원
- 복합어의 이형태 통일
- 접속문에 의해 생략된 주어의 복구

3. 코퍼스 정규화 규칙

3.1 보조용언에 의한 격이동의 복원

격이동 복원 규칙은 격이동 현상이 나타난 문장을 격이동이 일어나기 전의 형태로 복원한다.

3.1.1 단일 보조 용언인 경우

1. 피동을 나타내는 보조 용언
   피동이란 어떤 주체가 동작 또는 상태 변화를 입음을 표현하는 태의 일종이다. 즉, 이것은 주체의 동작이나 상태의 변화가 다른 행위자에 의하여 이루어짐을 나타내는 문법 범주이다. 이와 달리 스스로의 힘으로 행하는 행위나 동작을 능동이라 한다[이주행 1993]. 피동문의 유형[1]은 다음과 같다.

   - 능동사 어간 + 피동 접미사 '-아/어지다'
   - 동작성명사 + '되다', '당하다', '받다'

   다음은 피동문과 그에 해당하는 능동문의 예이다.

[1] 피동문의 유형에서 능동사 어간 + 피동 접미사 '이,히,리,기'는 코퍼스 정규화에서 제외된다. 이 유형은 하나의 품사로 간주되기 때문에 잘못된 지식 추출과 관계가 없다.

규칙	피동문		능동문
	능동사 어간 + 피동 접미사 '-아/어지다'	→	능동사 어간
	동작성 명사 + '되다', '당하다', '받다'	→	동작성 명사 + '하다'
P1	에게/한테/에/에 의해	→	가
	가	→	를
P2	에게/에 의해	→	가
	가	→	로
	에서	→	를

표 1: 피동 복원 규칙

피동문	능동문
1) 얼굴이 창백해지다.	얼굴이 창백하다.
2) 그들에 의해 협상이 깨어지다.	그들이 협상을 깨다.
3) A가 영희에 의해 B로 변경되다.	영희가 A를 B로 변경하다.
4) B가 영희에 의해 A에서 변경되다	영희가 A를 B로 변경하다.
5) 그가 검찰에 구속당하다.	검찰이 그를 구속하다.
6) 환자가 의사에게 치료받다.	의사가 환자를 치료하다.

예문에서 보듯이, 피동화는 격이동을 유발시킨다. 즉, 능동문의 목적어가 피동문의 주어가 되며 능동문의 주어는 부사격 조사에 의해 부사어가 된다. 그러나 동사의 유형에 따라 격이동이 발생하지 않는 것들도 있다. 문장 1)에서 보듯이 형용사나 자동사에 피동 접미사 '-아/어지다'가 결합될 경우는 격이동이 발생하지 않는다. 이를 정리한 피동 복원 규칙은 표 1과 같다.

2. 사동을 나타내는 보조 용언
남으로 하여금 어떤 동작을 하게 하는 동작을 사동이라고 한다. 반면에 어떤 동작이나 행위를 자기 스스로 행하는 것을 주동이라고 한다[이주행 1993]. 사동문의 유형[2]은 다음과 같다.

- 능동사 어간 + '게 하다'
- 동작성 명사 + '시키다'

다음은 피동문과 그에 해당하는 주동문의 예이다.

사동문	주동문
1) 철수가 얼음을(이) 녹게 하다.	얼음이 철수에 의해 녹다.
2) 어머니가 철수에게(가) 책을 읽게 하다.	철수가 어머니에 의해 책을 읽다.
3) 산성비가 농작물을 고사시키다.	산성비에 농작물이 고사하다.
4) 그가 아이에게 우유를 배달시키다.	그에 의해 아이가 우유를 배달한다.

예에서 보듯이 복원 후의 용언의 형태가 자동사, 형용사인 경우와 타동사인 경우의 격이동 양상이 다르다. 복원 후의 용언이 형용사나 자동사가 되는 경우(예문 1,3) 사동문의 목적어가 주어가 되며, 복원 후의 용언이 그대로 타동사인 경우(예문 2,4) 사동문의 부사어가 주동문의 주어가 된다. 이를 규칙으로 정리하면 표 2와 같다.

3. 희망을 나타내는 보조 용언 '고 싶다'
보조 용언 '고 싶다'는 희망을 나타내는 보조 용언이다. 이 보조 용언이 문장의 용언부에 결합될 경우도 격이동이 발생한다. 다음의 예문을 보자.

[2]사동문의 유형에서 능동사 어간 + 사동 접미사 '이,히,리,기,우,구,추'는 피동문에서와 같은 이유로 제외된다.

규칙	사동문		주동문
C1	자동사/형용사 + '게 하다'	→	자동사/형용사
	동작성 명사 + '시키다'	→	동작성 명사 + '하다'
	가(선행)	→	에 의해/에
	가(후행)	→	가 ('게 하다'의 경우만 적용)
	를/에게	→	가
C2	타동사 + '게 하다'	→	타동사
	동작성 명사 + '시키다'	→	동작성 명사 + '하다'
	가	→	에 의해/에
	가/에게	→	가

표 2: 사동 복원 규칙

규칙	결합된 형태		결합되지 않은 형태
G1	본용언 + 보조 용언 '고 싶다'	→	본용언
	가/는(선행)	→	가
	가(후행)	→	를

표 3: '고 싶다'에 의한 격이동 복원 규칙

'고 싶다'가 결합된 형태 '고 싶다'가 결합되지 않은 형태
1) 철수는(가) 사과가(를) 먹고 싶다. 철수가 사과를 먹다.
2) 철수는(가) 그녀가(를) 보고 싶다. 철수가 그녀를 보다.

예문에서 보듯이 보조 용언 '고 싶다'가 결합할 경우에는 문장 1)의 "철수는 사과가 먹고 싶다."와 같이 격이동에 의해 조사의 변화가 일어나기도 하고 "철수는 사과를 먹고 싶다."와 같이 그렇지 않기도 한다. 격이동이 발생할 경우는 표 3과 같은 규칙을 따른다.

4. 형용사의 동사화 보조 용언 '어 하다'
보조 용언 '어 하다'는 형용사에 결합되어 그 형용사를 동사화한다. 이 과정에서도 격이동이 발생한다. 다음의 예문을 보자.

'어 하다'가 결합된 형태 '어 하다'가 결합되지 않은 형태
1) 나는 철수를 부러워 한다. 나는 철수가 부럽다.
2) 그는 이 책을 재미있어 한다. 그는 이 책이 재미있다.

문장 1)을 보면, 부러워하는 대상이 보조 용언 '어 하다'에 의해 주격 조사 '가'에서 목적격 조사 '를'로 변화함을 알 수 있다. 이와 같은 격이동을 복원하기 위한 규칙은 표 4와 같다.

3.1.2 복수 보조 용언인 경우

일반적으로 문장의 용언부에 하나 이상의 보조 용언들이 결합되어 나타난다. 용언부에 결합된 보조 용언 상당어구 중에 격이동을 유발하는 보조 용언이 하나만 포함되어 있는 경우는 단일 보조 용언 격이동 복원 규칙을 따른다. 그리고

규칙	결합된 형태		결합되지 않은 형태
G2	본용언 + 보조 용언 '어 하다'	→	본용언
	가/는	→	는
	를	→	가

표 4: '어 하다'에 의한 격이동 복원 규칙

규칙	이형태	통일된 형태
B1	동작성 명사 + X (X ∈ {하다, 되다, 시키다, 당하다})	
	동작성 명사 + X 동작성 명사 + 목적격 조사 [space] X 동작성 명사 [space] X	→ 동작성 명사 + X
B2	명사와 용언의 결합	
	명사 + 용언 명사 + 목적격 조사/주격 조사 [space] 용언 명사 [space] 용언	→ 명사 + 용언

표 5: 복합어의 이형태 통일을 위한 규칙

격이동을 유발하는 보조 용언이 두 개 이상 용언부에 나타날 경우는 각 보조 용언의 복원 규칙을 보조 용언이 적용된 순서의 역순서로 적용한다.

용언부에 피동 보조 용언이 두 번 이상 나타날 경우는 임홍빈의 단회피동제약을 따른다[이주행 1993]. 이것은 일단 피동문이 되면 다시 피동화하지 않는다는 제약이다. 다음의 예문을 보자.

1. 학생들이 이것을 쓰다.
2. 이것이 학생들에게 쓰인다.
3. 이것이 학생들에게 쓰이어진다.

문장 1-3)은 용언부에 접미사 '이, 히, 리, 기'에 의한 피동과 접미사 '어지다'에 의한 피동이 모두 발생한 경우이다. 그러나 접미사 '이, 히, 리, 기'에 의해 한번 피동화된 문장(1-2)과 비교하여 보면 조사의 변화가 없다. 이것은 접미사 '이, 히, 리, 기'에 의해 피동화가 되고, 접미사 '어지다'가 피동화된 상태를 강조한다고 볼 수 있다. 이중 피동의 경우를 제외하고는 서술부에 적용된 보조 용언들에 대해 역순으로 단일 보조 용언 격이동 복원 규칙을 적용한다. 다음의 예문을 보자.

1-1) 그 사건이 학생들을 검찰에 구속당하게 한다. ↓ 사동 복원 규칙 C1 적용
1-2) 그 사건에 의해 학생들이 검찰에 구속당하다. ↓ 피동 복원 규칙 P1 적용
1-3) 그 사건에 의해 검찰이 학생들을 구속하다.

문장 1-1)은 피동 접미사 '당하다'와 사동 보조 용언 '게 하다'가 순서적으로 용언부에 나타난다. 따라서 사동 보조 용언 '게 하다'에 의한 격이동 복원 규칙 C1과 피동 접미사 '당하다'에 의한 격이동 복원 규칙 P1을 순서적으로 적용하여 최종적으로 문장 1-3)과 같이 복원한다. 다음도 용언부에 격이동을 유발하는 보조 용언이 하나 이상 나타날 경우의 처리 예이다.

2-1) 그는 어머니를 보고 싶어 한다. ↓ '어 하다' 복원 규칙 G2 적용
2-2) 그는 어머니가 보고 싶다. ↓ '고 싶다' 복원 규칙 G1 적용
2-3) 그는 어머니를 보다.

3.2 복합어의 이형태 통일

본 논문에서 대상으로 하는 복합어로는 '맛-있다, 끌-내다, 성-나다' 등 명사에 용언이 결합된 경우와 '일-하다, 일-시키다'와 같이 동작성 명사에 동사 파생 접미사, '하다, 되다, 시키다, 당하다'가 결합된 경우가 있다. 이러한 복합어는 '맛-있다, 끌-내다, 일-하다, 일-시키다' 등과 같이 하나의 단일어로 나타나기도 하고, 조사가 추가되어 '맛이 있다, 끌을 내다, 일을 하다, 일을 시키다' 등의 형태로 나타나기도 하고, 또는 조사가 생략되어 '맛 있다, 끌 내다, 일 하다,일 시키다' 등의 형태로 쓰이기도 한다. 이형태를 갖는 복합어 모두 단일어 형태로 통일한다. 복합어의 이형태를 통일하기 위한 규칙은 표 5와 같다.

규칙	주어 복구 전		주어 복구 후
S1	$H_{subject}\ H_{predicate}\ E_{same}\ T_{predicate}$	$\rightarrow$	$H_{subject}\ H_{predicate}\ E_{same}\ H_{subject}\ T_{predicate}$
S2	$H_{predicate}\ E_{same}\ T_{subject}\ T_{predicate}$	$\rightarrow$	$T_{subject}\ H_{predicate}\ E_{same}\ T_{subject}\ T_{predicate}$

표 6: 주어 복구 규칙

3.3 접속문에 의해 생략된 주어의 복구

접속(conjunction)은 둘 또는 그 이상의 문장이 연결 어미에 의해 대등적으로 혹은 종속적으로 결합되어 더 큰 문장이 되는 것이다. 둘 이상의 문장이 연결 어미에 의해 결합될 때, 특정 연결 어미는 동일 주어 제약을 따른다. 다음의 문장을 보자.

1. 책을 읽으려고 철수는 도서관에 갔다.
2. 철수는 책을 읽는다.
3. 철수는 도서관에 갔다.

문장 1)에서 의도를 나타내는 연결 어미 '려고'는 동일 주어 제약을 따른다. 따라서 문장 1)은 문장 2), 3)과 같은 두 문장으로 나뉠 수 있다. 코퍼스 정규화에서는 동일 주어 제약을 적용할 수 있는 접속문의 경우 생략된 주어를 복구해 낸다. 동일 주어 제약을 따르는 연결 어미는 다음과 같다.

- 인과 관계를 나타내는 '-느라고'에 의한 접속
- 의도 관계를 나타내는 '-려고, -고자, -러'에 의한 접속
- 순차 관계를 나타내는 '-고(서)'에 의한 접속

선행문의 주어와 서술어를 $H_{subject}$, $H_{predicate}$, 후행문의 주어와 서술어를 $T_{subject}$, $T_{predicate}$, 동일 주어 제약을 따르는 어미를 E_{same}라 할 때, 주어 복구 규칙은 표 6과 같다. 즉, 후행문의 주어가 생략되면, 선행문의 주어가 후행문의 주어를 대신하고 선행문의 주어가 생략된 경우에는 후행문의 주어가 선행문의 주어를 대신한다.

4. 코퍼스 정규화 규칙 적용 순서

코퍼스 정규화가 필요한 현상이 한 문장에 하나 이상 나타날 경우 각 규칙간 적용 순서가 존재한다. 그 적용 순서는 그림 1과 같다.

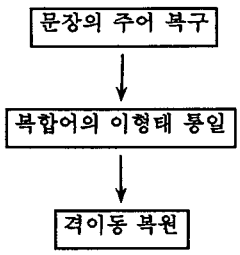

그림 1: 코퍼스 정규화 규칙 적용 순서

격이동을 복원하기 전에 문장의 주어를 복구해야 한다. 복구된 주어가 격이동 복원에 참여할 수도 있기 때문이다. 마찬가지로 격이동 복원 전에 복합어의 이형태를 통일해야 한다. 즉, '공부를 시키다'를 '공부시키다'로 통일시켜야 '공부하다'로의 격이동 복원이 가능하게 된다. 그리고 복합어의 이형태는 '맛이 있다-맛있다'와 같이 주어-서술어 관계에서도 발생하므로 복합어 이형태 통일 전에 문장의 주어를 먼저 복구한다. 하나 이상의 정규화 규칙을 적용해야 하는 예를 들어 보자.

1-1) 아들에게 공부를 시키느라고 그는 힘들었다. ↓ 문장의 주어 복구 (규칙 S2)
1-2) 그는 아들에게 공부를 시키느라고 그는 힘들었다. ↓ 복합어의 이형태 통일 (규칙 B1)
1-3) 그는 아들에게 공부시키느라고 그는 힘들었다. ↓ 격이동 복원 (규칙 C1)
1-4a) 그에 의해 아들이 공부하다.
1-4b) 그는 힘들었다.

문장 1-1)의 어미 '느라고'는 동일 주어 제약을 따르므로 주어 복구 규칙(S2)을 적용하여 선행문의 주어를 '그는'으로 복구한다. 문장 1-2)에서는 '공부-시키다'의 이형태인 '공부를 시키다'가 나타나므로 복합어 이형태 통일 규칙(B1)을 적용하여 이형태를 통일한다. 문장 1-3)에서는 동작성 명사에 사동 접미사 '시키다'가 결합한 격이동 현상이 발생하므로 격이동 복원 규칙(C1)을 적용하여 문장 1-4a)와 같이 격이동이 발생하기 전의 문장으로 복원한다.

5. 코퍼스 정규화에 대한 실험 및 평가

코퍼스 정규화 실험은 자료 부족 문제에 초점을 맞추어 진행되었고, 정규화를 하기 전과 하고 난 후의 자료 부족 문제의 감소 정도를 평가하였다.

5.1 학습 대상 코퍼스

본 논문에서 이용하고 있는 코퍼스는 한국과학기술원에서 작성한 KAIST 코퍼스이다[김재훈 1995]. 이 코퍼스는 약 20만 어절로 구성되어 있으며 신문 사설, 국민학교 교과서, 소설, 수필 등의 장르를 포함하고 있는, 품사가 부착된 것이다. 코퍼스의 품사 부착에 이용된 한국어 품사 체계는 언어학적인 측면을 고려하여 분류된 [김재훈 1994]를 이용했다. 이 품사 체계는 자동적인 방법에 의한 것이 아니며, 이미 언어학자들이 제시한 여러 분류 체계와 실제의 예문을 참조하여 분류된 것이다.

5.2 격이동 복원 실험

사동과 피동을 나타내는 보조 용언에 의한 격이동 복원 실험을 하였다. 동사 '파괴하다'의 사동, 피동형인 '파괴당하다, 파괴되다, 파괴시키다'는 이 실험에 의해 원형인 '파괴하다'로 복원된다. 보조 용언에 의해 사동, 피동화된 용언들을 원형으로 복원한 결과, 코퍼스 내에서의 용언의 전체 갯수가 5041개에서 4422개로 12% 줄어 들었다. 코퍼스 내에서의 용언의 전체 빈도수는 66655개이며 따라서 용언의 평균 빈도수는 복원 전 13.22개에서 복원 후 15.07개로 14% 증가하였다. 또한 자료 부족 문제는 빈도수가 적은 사건에 의해 발생하므로 정규화에 의해 빈도수가 적은 용언들의 분포가 어떻게 달라지는지를 살펴보았다. 표 7은 빈도수가 10 이하인 용언에 대해 살펴본 것이다. 표에서 보듯이 빈도수가 적은 용언의 갯수가 줄어듬을 알 수 있다.

5.3 복합어의 이형태 통일에 대한 실험

동작성 명사에 동사화 접미사 '하다'가 결합되는 형태(실험 I)와 '끌내다'와 같이 명사에 용언이 결합한 형태(실험 II)로의 통일 실험을 수행하였다. 실험 I의 경우, 661개의 이형태 유형으로부터 307개의 통일된 유형을 추출하였다. 유형의 수는 53.56 % 감소하였으며 유형의 평균 빈도수는 7.57개에서 16.30개로 증가하였다. 실험 II의 경우, 166개의 이형태 유형으로부터 69개의 통일된 유형을 추출하였다. 유형의 수는 58.53 % 감소하였으며, 평균 빈도수는 8.37개에서 20.1개로 증가하였다. 이들에 대해서도 자료 부족 문제의 감소 정도를 살펴 보기 위해 적은 빈도수를 나타내는 복합어의 이형태 통일하기 전과 후의 빈도수 차이를 실험해 보았다. 그래프 2, 3은 빈도수 50까지의 복합어를 대상으로 한 결과이다. 그래프에서 알 수 있듯이 이형태 통일에 의해 빈도수가 적은 복합어의 갯수가 많이 줄어듬을 알 수 있다.

빈도수	격이동 복원 전 용언의 갯수	격이동 복원 후 용언의 갯수	줄어든 정도
1	1930	1504	22.07 %
2	791	698	11.76 %
3	430	370	13.95 %
4	306	278	9.15 %
5	218	209	4.13 %
6	144	136	5.56 %
7	118	114	3.39 %
8	90	79	12.22 %
9	76	67	11.84 %
10	67	66	1.49 %

표 7: 격이동 복원 실험 결과

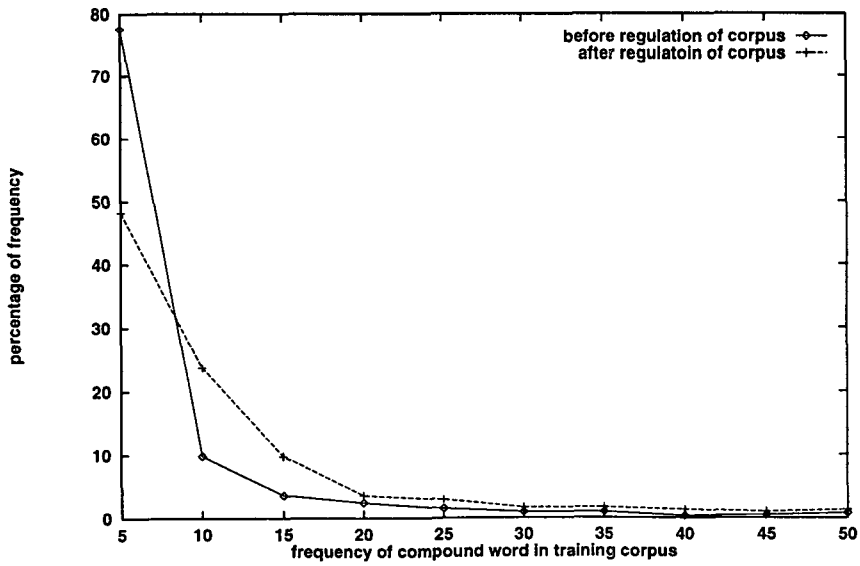

그림 2: 동작성 명사+'하다'에 대한 이형태 통일 실험 결과(실험 I)

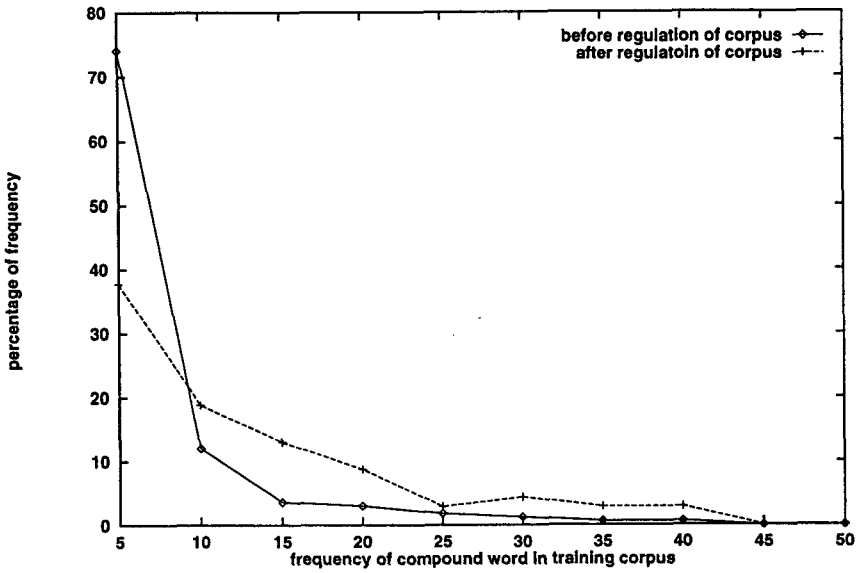

그림 3: 명사+용언에 대한 이형태 통일 실험 결과(실험 II)

6. 결론

본 논문에서는 한국어 품사 부착 코퍼스로부터 구문 관계 지식을 추출하고자 할 경우 필요한 코퍼스 정규화에 대해 살펴보았다. 한국어에서 문장의 구성 성분들 간의 관계는 조사에 의해 표현되므로 정규화의 범위를 격이동의 복원, 복합어의 이형태 통일, 접속문의 생략된 주어 복구로 결정하였다. 그리고 이러한 범위에 대해 한국어의 문법적 특징을 반영한 정규화 규칙을 정의하였다. KAIST 코퍼스에 대해 정규화 실험을 수행하여 약 41.3 %의 용언의 수를 감소할 수 있었으며, 용언의 평균 빈도수를 증가시켰다. 또한 정규화 실험에 의해 빈도수가 적은 용언의 갯수가 월등히 줄어드는 결과를 얻었다.

모든 정규화 규칙에 대한 실험을 한 것이 아니므로 적용되지 않은 정규화 규칙에 대한 실험이 필요하며, 정규화된 결과의 정확성에 대한 평가도 함께 진행되어야 할 것이다.

참고 문헌

[Marcus 1993] M. P. Marcus, B. Santorini and M. A. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank," *Computational Linguistics*, 1993.

[김재훈 1994] 김재훈, 서정연, 자연언어 처리를 위한 한국어 품사 태그, 한국과학기술원, 인공지능센터, 기술 보고서, CAIR-TR-94-55, 1994.

[김재훈 1995] 김재훈, 김길창, 한국어에서의 품사 부착 말뭉치의 작성 요령 : KAIST 말뭉치, 한국과학기술원, 기술 보고서, CS/TR-95-99, 1995.

[이주행 1993] 이주행, 현대국어문법론, 1993.

[최기선,1989] 최기선, 한국어 해석을 위한 격이동 패턴의 고찰, 인지과학, 민음사, 1989.