표현 패턴에 의한 한국어-영어 기계 번역을 위한 개념 구성

이 호 석
뉴미디어학과 공과대학 호서대학교
hslee@office.hoseo.ac.kr

A Conceptual Framework for Korean-English Machine Translation using Expression Patterns

Ho Suk Lee
New Media Dept. College of Engineering Hoseo University

요 약

본 논문은 표현 패턴(expression patterns)을 이용한 한국어-영어 기계 번역 방법에 대하여 논의한다. 표현 패턴은 한국어-영어 기계 번역을 위하여 의미적이고 표현적인 관점에서 한국어 표현에 적합한 영어 표현을 대응하여 정의하였다. 그리고 새로운 한국어 파싱 방법을 개발할 것을 제안하였다. 한국어 파싱 방법은 한국어의 교착어로서의 특징, 표현 패턴 개념, 문장 분할 개념, 그리고 파싱 과정에 의미 구조를 포함하는 개념 등을 포함하여 개발할 것을 주장하였다. 논문의 마지막 부분에는 간단한 한국어 문법을 정의하여 새로운 한국어 구문 분석 방법의 가능성을 제시하였다.

Abstract

This paper discusses a Korean-English machine translation method using expression patterns. The expression patterns are defined for the purpose of aligning Korean expressions with appropriate English expressions in semantic and expressive senses. This paper also argues to develop a new Korean syntax analysis method using agglutinative characteristics of Korean language, expression pattern concept, sentence partition concept, and incorporation of semantic structures as well in the parsing process. We defined a simple Korean grammar to show the possibility of new Korean syntax analysis method.

1. 서 론

한국어-영어 기계 번역을 위해서는 번역의 품질을 결정하는 주요 문장 요소를 파악하는 것이 중요하다. 본 논문에서는 우선 한국어의 문법에 대하여 조사하였다. 참고 문헌 [1]은 한국어의 통사 구조, 통사 규칙, 그리고 통사적 사실들의 지배원리를 규명하고 기술하였다. 참고 문헌 [2]는 한국어의 형용사에 대하여 형태론, 의미/화용론, 통사론적인 관점에서 고찰하여 논의하였다. 참고 문헌 [3]은 한국어의 부사절에 대하여 상세하게 논의하였다. 참고 문헌 [4]는 한국어의 종결 어미와 보조동사와 한국어의 복합문에 대하여 논의하였다. 참고 문헌 [5]는 외국인(일본인)의 관점에서 한국어 단어의 형태와 문법에 대하여 용례를 들어서 간결하면서도 쉽게 설명하고 있다. 참고 문헌 [6]은 한국어 문법 체계에 대하여 폭넓게 설명하고 있다. 참고문헌 [7]은 한국어와 만주 퉁구스어 특히 그 중에서 어웡키 어와의 문법 구조 대조 연구가 상세히 설명되어 있다. 참고문헌 [8]은 우리말의 형태에 대하여 형태소론, 단어의 구조와 파생법, 단어 합성법, 굴곡법(어미활용), 준굴곡법(조사활용), 품사론으로 나누어서 설명하고 있다. 한국어 단어와 어미의 형태를 이해하기 위해서는 반드시 참고하여야 될 책으로 생각된다. 참고문헌 [9]는 10세기 말엽까지의 연대를 고대국어의 시기로 정의하고 이 시기 이전의 국어에 대하여 논의하고 있다. 1.1.3 삼국시대의 언어에서는 여러 단어와 지명을 예로 들어서 삼국의 언어들 사이에는 방언적인 차이이외에 그다지 차이가 없었다고 설명하고 있다. 참고문헌 [10]은 한국어의 문법에 대하여 음운론과 자소론, 형태론, 통사론과 의미론, 텍스트과학으로 나누어서 설명하고 있다. 참고 문헌 [11]은 한국어에 나타나는 개념의 분류 구조에 대하여 논의한다. 이 책의 내용은 의미 표지(semantic feature)를 작성하는데 좋은 참고가 될 것으로 보인다. 참고 문헌 [12]는 한국어 문장을 종류별로 분류하여, 문장의 유형에 따른 의미 분석, 문장 간의 의미 관계, 문장의 화용 의미, 문장의 논리성, 문장의 유형에 따른 기능 등을 논의하고 있다. 참고 문헌 [13]은 한국어 문법에 대하여 간결하고 체계적으로 설명하여 쉽게 살펴 볼 수 있도록 하였다. 참고문헌 [14]는 대조언어학의 관점에서 한국어와 스페인어의 문장들을 대조하여 문법, 문장 형태, 그리고 의미를 논의하고 있다. 참고 문헌 [15]는 한국어에 대하여 단어 형태의 구성법, 굴곡법, 영굴곡법(단어 자체로서 조사나 어미의 활용이 없이 뜻을 나타내는 방법)으로 나누어서 논의하고 있다. 참고문헌 [16]은 국어에 대하여 형태론과 통사론으로 구분하여 논의하고 있다. 참고 문헌 [17]은 독특한 용어와 방법으로 국어의 문법을 설명하고 있다. 예를 들어 문장을 "월" 이라고 하고 주어 부분을 "임자조각" 그리고 술어 부분을 "풀이조각" 이라고 한다. 주어는 "임자말"이라고 하고 술어는 "풀이말"이라고 한다. 국어 문법은 형태소와 낱말, 낱말 만들기, 씨의 범주, 준굴곡법, 굴곡법, 영굴곡법, 월의 성분, 월의 갈래, 말본 요소의 통어적 기능의 순서로 논의하였다. 참고 문헌 [18]은 국어 명사의 의미 체계, 의미 구조, 의미 구조의 구성, 명사 연결 구성과 의미 구조의 관계에 대하여 논의하고 있다. 우선 의미 영역에 대하여 분류하였는데, 의미 영역을 실체와 양식으로 분류하였고, 양식은 다시 사태와 관계로 분류하였으며, 사태는 다시 사건과 상태로 구분하였다. 의미 구조의 구성에서는 의미 네트워크(semantic network)에 대하여 논하였으며, 명사의 의미 구조인 속성 구조(qualia structure) 그리고 의미 자질(semantic feature)에 대하여서도 논의하였다. 명사의 의미 구조 참고에 유용한 도서이다. 참고 문헌 [19]는 국어의 자동사를 비행위성 자동사와 행위성 자동사로 분류하여 논의하였다(80쪽 <표 1>). 비행위성 자동사는 대상자동사, 소재자동사, 비교자동사, 변성자동사, 대칭자동사, 피동자동사, 심리자동사로 구분하였다. 대상자동사의 예로는 "곪다", "늙다" 등을 제시하였다. 소재자동사의 예로는 "붙다", "속하다" 등을 제시하였다. 행위성 자동사는 행위자동사, 위치자동사, 이동자동사, 대칭자동사로 분류하였다. 행위자동사의 예로는 "날다", "뛰다" 등을 제시하였으며, 위치자동사의 예로는 "이르다", "닿다" 등을 제시하였다. 그 밖에 자동과 타동, 행위성과 비행위성이 미분화된 동사들도 있는데 이러한 동사들은 중립동사로 분류하였다. 중립동사의 예로는 "가시다", "글썽거리다" 등을 들었다. 마지막으로 결론 부분에는 논의한 자동사들의 자세한 분류표가 제시되어 있다(367쪽). 참고 문헌 [20]은 비교적 전통적인 관점에서 국어의 문법에 대하여 논의하였다. 국어 문법을 형태론, 품사론, 의존명사론, 불규칙활용론, 의존용언론, 서법과 양태론 등으로 설명하였다. 참고 문헌 [21]은 국어의 화용론(pragmatics)에 대하여 서술하였다. 화용론은 언어의 사용에 대한 것으로 음운론적, 형태론적, 통사론적, 의미론적, 담화/텍스트론적, 사회언어학적, 인지심리학적으로 화용론을 분류하여 논의하였다. 3.5절에는 대화의 유형이 분류되어 있으며, 7장에는 화용론에 대한 종합적 분석의 예가 제시되어 있는데, MBC 100분 토론(2006년 3월 6일) "과자 유해 논란" 토론 중에서 처음 30분 동안의 대화가 제시되어 있다. 이 대화 텍스트에 대하여 음운론적, 형태론적, 통사론적, 담화론적 분석이 제시되어 있다. 담화론적 분석에서는 95개의 담화 표지가 사용되었다고 제시되어 있다. 전산 언어학 연구의 관점에서는 중요한 자료가 될 것이다. 참고 문헌 [22]는 국어 조사 중에서 "-에" 와 "-로"의 용법에 대하여 매우 상세하게 설명되어 있다. 한국어의 번역을 위해서는 참고할 만한 도서이다. 참고 문헌 [23]은 한국어 문장 분석의 새로운 접근(27~46쪽) 방법을 논의하고 있다. 한국어 문장의 분석을 위하여 거시구조의 틀과 미시구조의 틀을 제시하고 있다. 거시구조의 틀에는 어미 구조, 어미와 어미의 상관성, 한정조사와 어미의 상관관계, 부사기능어(문장 부사, 양태 부사)와 어미의 상관관계 등이 포함된다. 미시구조의 틀에는 연어구성, 관용구 구성, 구동사(phrasal verb), 술어와 논항 구조(argument structure), 문법관계표지(격조사, 격어미) 등이 포함된다. 한국어의 기계 처리를 위해서는 참고할 만한 도서이

다. 참고 문헌 [24]는 국어의 첨가어(교착어)적인 특징을 잘 설명하고 있다. 국어의 중요한 특징으로는 (1) 국어는 근간 성분의 생략이 용이한 언어이다. (2) 국어는 첨가어이다. (3) 주어가 잇달아 나타나는 문장 구성 방법이 있다. (4) 또한 주어가 없는 문장도 많이 쓰인다. (5) 조사와 어미가 매우 발달하여 다른 유형의 언어에서는 부사나 기타 독립된 어휘로 나타나야 할 의미가 조사나 어미의 형태로 나타난다. (6) 연결어미가 매우 많아서 여러 가지 뜻으로 두 문장을 접속시킨다. (7) 문장의 통사적 서법이나 절의 통사적 자격이 문말 형태소에 의하여 외현적으로 구현된다. (8) 논항이 문맥적 상황에 따라 공범주(empty category)로 실현되는 경우가 많다. (9) 형용사가 동사와 거의 비슷한 문법적 역할을 한다. (10) 구성 성분의 순서가 기저구조에서는 엄격히 정해져서 "주어+목적어+동사"의 순서를 가진다. 기본 구성 성분의 순서는 핵말 언어(head-final language)로서 모든 구에서 핵은 보충어의 뒤에 나타난다. 등이 제시되어 있다. 국어의 이러한 특징은 분명히 인정되어야 하며 서구 언어의 이론에 입각하여 국어를 보는 것은 바람직하지 않다고 논의하고 있다. 다음은 문헌에서 제시하는 국어 문장의 재미있는 예이다.

(ㄱ). 그가 와 기쁘다. (ㄴ). 그가 와서 기쁘다.
(ㄷ). 그가 오니 기쁘다. (ㄹ). 그가 오니까 기쁘다.
(ㅁ). 그가 오매 기쁘다. (ㅂ). 그가 오므로 기쁘다.
(ㅅ). 그가 온지라 기쁘다. (ㅇ). 그가 오는데 기쁘다.

"오다" 동사 어미의 활용에 따라서 문장의 의미가 조금씩 다르다는 것을 나타내고 있다. 국어가 다른 언어에서는 가능하지 않은 문장의 섬세한 의미의 차이를 어미의 활용으로 나타내고 있다는 것을 보여준다. 참고할 만한 자료라고 본다. 참고 문헌 [25]는 국어의 어미의 활용에 대하여 서술하였다. 서술한 어미에는 "-고 하-"의 생략 현상, "-ㄴ단다, -ㄴ다네, -ㄴ다오, -ㄴ답니다", "-다고, -으라고, -자고", "-다면서, -으라면서, -자면서" 등등이 있다. 참고 문헌 [26]은 한국어의 능동과 피동의 상호 관계를 서술하고 있다. 한국어의 피동서술어와 피동문의 목적어에 대하서도 서술하고 있다. 한국어는 영어에 비하여 다양한 형태의 피동문이 있으므로 번역의 관점에서는 참고하여야 할 문헌이다. 다음의 예문을 보자.

(ㄱ). 영선이는 {옷이(O), 옷을(X)} 질질 끌렸다.
(ㄴ). 영선이는 한 어린아이에게 {손목이(O), 손목을(O)}끌려 가게로 향하고 있었다.

본 문헌에서는 두 문장의 경우 모두 동사 "끌리다"에 의한 피동문이지만, (ㄱ)의 경우에는 목적격 조사 "을"을 허용하지 않지만 (ㄴ)의 경우에는 목적격 조사 "을"을 허용한다고 설명하고 있다. 그런데 요사이는 (ㄱ)의 경우에도 목적격 조사를 허용하는 것이 어색하지 않은 것으로 느껴진다. 참고 문헌 [27]은 화용표지어에 대하여 서술하고 있다. 화용표지의 예로는 "말이야", "이제", "-로", "-요" 등을 설명하고 있다. 화용표지의 예로는 다음과 같은 문장이 있다.

(ㄱ) 그렇게 어려운 문제는 말이야, 자네가 풀어야 된다.
(ㄴ) 이제는 말이야, 우리는 아프리카를 다시 봐야 한다.

참고 문헌 [28]은 현대 국어의 의존 명사에 대하여 논의하고 있다. 연구는 말뭉치를 분석하여 의존 명사의 용례를 찾는 방법으로 진행하였다. 의존 명사의 정의로는 "명사의 통사론적 기능을 가지나 어휘 의미가 분명하지 않고 문장에서 쓰일 때 다른 환경요소의 도움을 받아 그 의미가 뚜렷해지는, 자립성이 완전하지 못한 형태를 지칭한다." 로 하였다. 국어의 의존 명사에는 "깐, 것, 게, 길, 김, 나름, 나위, 녀석, 년, 녘, 노릇, 듯, 데, 바, 줄, 만큼" 등등이 있다. 의존 명사에 대한 연구 결과는 자연언어처리에 유용하게 쓰일 수 있을 것이다. 참고 문헌 [29]는 자연언어처리 입장도 고려하여 한국어 문장 분석 방법에 대하여 자세하게 서술하였다. 먼저 기존의 한국어 구문 분석에 대한 연구들을 인용하였으며, 이들을 구절구조문법(PSG, phrase structure grammar), 중심어 주도 구구조 문법(HPSG, head-driven phrase structure grammar), 어휘 기능 문법(LFG, lexical-functional grammar), 범주 문법(CG, categorial grammar), 의존 문법(DG, dependency grammar), 그리고 몬태규 문법(Montague grammar) 등의 범주로 분류하였다. 대부분의 전산학적인 연구는 구절구조문법, 중심어 주도 구구조 문법, 의존 문법 범주로 분류하였다. 다음에는 구문 분석 표지 작성 원칙, 통사 범주와 구문 분석 표지 정의, 구문 분석의 원리와 방법, 특이한 구성에 대한 분석, 문장과 문장 성분 등으로 구분하여 설명하였다. 한국어 구문 분석에 대하여서는 참고할 가치가 있는 문헌이라고 생각한다. 그러나 조사와 어미에 있어서 분류 체계와 문법 표지만 제시되었을 뿐 실제 예가 제시되어 있지 않는 점은 매우 아쉽다. 참고 문헌 [24][25][26]27][28][29]를 보면 근래에 들어와서 다수의 국어 학자들에 의하여 서구 언어 문법의 관점이 아니고 한국어의 교착어로서의 특징을 반영하는 문법 체계가 연구되고 있는 현상을 볼 수 있다. 국립국어원에서는 1998년부터 2007년까지 10년 동안 "21세기 세종계획"을 수행하여 한민족 언어 정보화 사업을 완수하였다. 이 계획에서는 한국어 말뭉치와 활용도구 등을 개발하였다[30].

다음에 본 논문에서는 효과적인 한국어-영어 기계 번역을 위하여 한국어 문장들에 대하여 의미가 동일한 영어 문장을 대조하여 동사와 형용사의 하위범주화 패턴(subcategorization pattern)을 고려하였으며 또한 표현 패턴(expression pattern)도 고려하였다. 우선 5절까지는 하위범주화 패턴을 이용한 기계 번역 방식을 논의하였고, 6절에서는 한국어의 기본 문장 형태에 대하여 제시하였다. 7절에서는 새로 도입한 개념인 표현 패턴 방식에 의한 기계 번역 방법을 논의하였다. 표현 패턴은 한국어와 영어의 표면 표현(surface expression)을 대조하고 비교하여 의미가 거의 동일한 표현의 쌍을 파악하여 구성하였다. 표현 패턴은 거의 모든 관용어를 포함할 수 있다. 특히, 한국어 동사 "논의하다"와 "조사하다"의 예를 들어 표현 패턴이 사전에 등재되는 모양에 대하여 제시하였다. 사전 내용의 비교를 위하여 하위범주화 패턴과 표현 패턴을 함께 제시하였다. 8절에서는 근래에 진행된 기계 번역 관련 연구에 대하여 논의하였으며, 표현 패턴의 이론적인 의미와 앞으로의 기계 번역 연구의 방향에 대하여서도 간단하게 논의하였다. 9절에서는 문장 분할 개념과 이를 적용한 기계 번역 방법에 대하여 논의하였다. 10절에서는 교착어로서의 특징, 표현 패턴 개념, 문장 분할 개념, 그리고 파싱 과정에 의미 구조를 포함하는 개념 등을 포함하는 새로운 한국어 파싱 방법의 개발에 대하여 논의하였다. 마지막으로 11장에서는 결론을 제시하였다.

이제 2절에서 조사를 논의하고 5절까지 동사, 보문, 부사와 함께 하위범주화 패턴과 표현 패턴에 대하여 논의하기로 한다.

2. 조 사

한국어의 품사에는 명사, 대명사, 수사, 동사, 형용사, 부사, 조사(후치사), 접속사, 감탄사, 관사 등이 존재한다. 이중에서 한국어-영어 기계 번역의 관점에서 한국어 문장의 문장 구조와 의미 구조에 중요한 영향을 끼치는 품사는 조사, 동사, 형용사, 그리고 부사 등이라고 할 수 있다. 동사나 형용사 어미의 활용, 한국어 보어 문장, 그리고 불완전 명사도 중요한 요소이다.

조사는 한국어 문장에서 기능적으로나 의미적으로 매우 중요한 역할을 수행하는 품사로서 격조사와 보조조사로 나눌 수 있다. 격조사에는 주격조사, 서술격조사, 목적격조사, 보어격조사, 관형격조사, 부사격조사, 접속격조사 등이 있다. 한국어 조사는 영어의 전치사보다 종류가 많고 의미도 다양하고 전치사로 번역할 수 없는 것들도 많이 있기 때문에, 한국어-영어 기계 번역을 수행할 때에 조심스럽게 다루어야 한다.

3. 동 사

동사에 대하여 살펴보자. 우선, 다음 두 개의 문장을 예를 들어 보자.

(문장 1) 철수는 정말로 회사에 갔나 싶다.
(문장 2) 철수가 이미 죽었는가 싶다.
(문장 3) 철수가 이미 죽었는지 싶다.

(문장 1)을 영어로 번역하면 "I wonder whether 철수 really has gone to company."가 될 것이다. (문장 2)와 (문장 3)을 번역하면 "I suppose that 철수 has already died."가 될 수 있다. 따라서 동사 "싶다"의 사전에는 다음과 같이 등재되어 있어야 할 것이다.

●싶다 : 동사
. 하위범주화 패턴1 : ($N_1 M_{subj}$ $N_2 M_{obj}$ V)
. if (V = 선행동사(보다)),
영어 번역 : 싶다 -> want to(infinitive),
V -> see
(sent (np(subj(N_1))
vp(infinitive(want to see) obj(N_2)))
. if (V = 선행동사(하다)),
영어 번역 : 싶다 -> want to(infinitive),
V -> do
(sent (np(subj(N_1))
vp(infinitive(want to do) obj(N_2)))
. if (V= 선행동사(존경하다)),
영어 번역 : 싶다 -> want to(infinitive),
V -> respect
. 하위범주화 패턴2 : ($N_1 M_{subj}$ CLAUSE)
// CLAUSE = ($N_2 M_{subj}$ $N_3 M_{adv}$ V)
. if (V = 선행동사(가다)), 영어 번역 :
N_1 suppose that N_2 go to N_3
(sent (np(subj(N_1)) vp(vt(suppose)
clause(that(that) np(subj(N_2))
vp(vi(go)) prep(to(to) np(N_3)))))
. if (V = 선행동사(그밖의동사)), 영어 번역 :
N_1 suppose that N_2 V N_3
(sent (np(subj(N_1)) vp(vt(suppose)
clause(that(that) np(subj(N_2))
vp(vt(V)) obj(N_3))))
. if (V = (선행동사(그밖의동사)) &
부사(정말로)),
영어 번역 : N_1 wonder
whether N_2 really V N_3
(sent (np(subj(N_1) vp(vt(wonder)
clause(whether(whether)
np(subj(N_2)) vp(adv(really)

v(V) obj(N_3))
. if (V = 선행동사(그밖의동사)) &
부사(이미)),
영어 번역 : N_1 wonder
whether N_2 already V N_3
(sent (np(subj(N_1)) vp(vt(wonder)
clause(whether(whether)
np(subj(N_2)) vp(adv(already)
v(V) obj(N_3))

따라서 위의 예들을 보면 적절한 영어 번역을 위해서는 한국어 동사의 하위범주화 패턴이 매우 중요하며, 이러한 하위범주화 패턴 정보가 반드시 사전에 등재되어야 한다.

결국 하위범주화 패턴은 동사나 형용사 활용의 모든 패턴을 나타내는 것이다. 따라서 하위범주화 패턴은 절(clause)을 포함할 수 있다. 이런 경우에는 절을 미리 영어로 번역한 다음에 최종적으로 만들어진 영어 번역 문장에 삽입하여 전체 영어 문장을 구성하는 경우도 있다.

4. 보 문

한국어-영어 기계 번역을 위해서는 보문에 대한 이해가 필요하다. 예를 들어, 다음의 문장들을 보자. 이 문장들은 보문에 대한 좋은 예이다.

(문장 4) 그는 학자가 되었다.
(문장 5) 코끼리는 코가 길다.
(문장 6) 이 나무에 새 잎이 돋는다.

(문장 4)를 영어로 번역하면 "He became a scholar." 가 되고, (문장 5)를 영어로 번역하면 "The nose of elephant is long." 이 되고, (문장 6)을 영어로 번역하면 "A new leaf is coming up on this tree." 가 될 것이다. 따라서 한국어에서는 동일한 구문 구조를 가지고 있지만, 영어로 번역하면 완전히 다른 구문 구조로 바뀌어 나타난다. 따라서 한국어 보어 문장의 적절한 번역에는 동사와 그리고 동사 앞에 위치하는 두 개의 명사 사이의 관계가 중요하다고 하겠다. 따라서 관련된 언어적 정보는 동사와 명사 두 부분에 모두 등재되어야 한다. 이 경우에도 파서는 문장의 구문 구조를 파악하는 역할을 수행하고, 실질적인 문장의 의미 구조는 사전에 등재되어 있는 동사나 형용사의 하위범주화 패턴과 문장의 주격 명사와 보격 명사의 관계에 의하여 결정되어야 할 것이다. 즉, 위의 (문장 4)에서는 명사 "그" 와 명사 "학자" 의 관계를 알아야 하며, (문장 5)에서는 명사 "코끼리" 와 명사 "코"의 관계를 알아야 한다. 이 두 예에서는 명사의 조사는 주격조사라는 것을 주의하여야 한다. 따라서 명사는 영어로 번역된 문장에서도 주어가 된다. (문장 6)에서도 명사 "나무" 와 명사 "잎" 의 관계를 알아야 한다. 그리고 이 문장에서는 명사의 조사는 처소격임도 알아야 한다. 이 경우에는 영어로 번역하였을 경우에 전치사구가 될 수 있다. 이 예는 문장의 구문과 의미 정보이외에도 보어 문장에서는 조사의 역할이 중요함을 나타낸다.

(문장 4)의 예에서 알 수 있듯이, 동사 "되다" 에는 사전에 다음 정보가 등재되어 있어야 한다.

●되다 : 동사
. 하위범주 패턴1 : (N_1M_{subj}, N_2M_{subj})
. 영어 번역 : N_1 become N_2

그리고 (문장 5)의 형용사 "길다" 에는 다음 정보가 등재되어 있어야 한다.

●길다 : 형용사
. 하위범주화 패턴1 : (N_1M_{subj})
. 영어 번역 : N_1 be long.
. 하위범주화 패턴2 : (N_1M_{subj}, N_2M_{subj})
. if(yes=Noun(N_1,N_2)), 영어 번역 :
N_2 of N_1 be long.

(문장 6)의 동사 "돋다" 에는 다음 정보가 등재되어 있어야 한다.

●돋다 : 동사
. 하위범주화 패턴1 : (N_1M_{subj})
. 영어 번역 : N_1 come up.
. 하위범주화 패턴2 : ($N_1M_{adv(에)}$, N_2M_{subj})
. 영어 번역 : N_2 come up ($Prep_{adv(에)}$) N_1

보문에 대한 하위범주화 정보도 사전에 등재되어 있어야 문장에 대한 구문 분석이 끝나고 필요한 구문 정보를 사전에서 추출하여 영어로의 번역에 사용할 수가 있다.

5. 부 사

다음에 한국어 부사절에 대하여 논의해 보자. 우선 다음 문장을 보자.

(문장 7) 물이 너무 맑으면 고기가 적게 든다.
(문장 8) 그는 글도 잘하며, 말도 잘한다.
(문장 9) 그 아이가 형과는 달리 사교에 능하다.

(문장 7)을 영어로 번역하면, "If water is too clear, fishes little gather." 가 될 수 있다. 이런 번역을 가능하게 하는 문장 요소는 부사적 어미 "-면" 이다. "-면"에 의하여 "if"가 도입된다. 동사 "듣다"에는 다음 언어 정보가 등재되어 있어야 한다. 그리고 명사 "고기"에는 다음 언어 정보가 등재되어 있어야 한다.

●듣다 : 동사
. 하위범주화 패턴1 : (N_1M_{subj})
. 영어 번역 : N_1 gather.

●고기 : 명사
. 관련명사 : 물
. 영어 번역 : fish
. 관련명사 : 동물
. 영어 번역 : meat

(문장 8)의 영어 번역은 "He writes well and talks well." 이 될 수 있다. 이 경우에는 "잘하다" 동사에 다음 언어 정보가 등재되어 있어야 한다.

●잘하다 : 동사
. 하위범주화 패턴1 : (N_1M_{obj} , N_2M_{obj})
. if (N_2 = 글), 영어 번역 : N_1 write well
. if (N_2 = 말), 영어 번역 : N_1 talk well
. if (N_2 = 피아노), 영어 번역 :
N_1 play piano well
. if (N_2 = 노래), 영어 번역 : N_1 sing well

(문장 9)는 "The boy is sociable unlike his old brother." 번역이 될 수 있다. 이 문장의 경우에는 동사 "능하다" 와 "다르다"는 다음과 같이 되어야 한다.

●능하다 : 동사
. 하위범주화 패턴1 : (N_1M_{subj}, N_2M_{place})
. if (N_2 = 사교), 영어 번역 : N_1 be sociable
. if (N_2 = 사격), 영어 번역 : N_1 shoot well
. if (N_2 = 일), 영어 번역 : N_1 be good at N_2

●다르다 : 동사
. 하위범주화 패턴1 : (N_1M_{subj}, N_2M_{adv})
. if (문장 = 단문), 영어 번역 :
N_1 be different from N_2
. if (문장 = 부사절), 영어 번역 :
N_1 unlike N_2

이제까지 한국어-영어 기계 번역에 있어서 한국어의 조사, 동사, 부사, 그리고 (문장 7)에서는 어미가 중요한 역할을 한다는 것을 제시하였다.

6. 한국어 기본 문장 문형

다음은 한국어의 기본 문장 문형들이다[6].

1. S -> NP VP
2. NP -> (Det) N_1M_{subj}
(Det) N_2M_{obj}
(Det) N_2M_{com}
(Det) $N_2M_{obj}N_3M_{com}$
3. VP -> (Adb) ± V AUX
(Det) N_1M_o, AUX
(Det) N_2M_{obj} (Adb) + V AUX
(Det) N_2M_{com} (Adb) ± V AUX
(Det) N_2M_{obj} (Det) N_3M_{com} Adb + V AUX
4. S -> (Det) N_1M_{subj} ± V AUX
(Det) N_1M_{subj} (Det) N_2M_o AUX
(Det) N_1M_{subj} (Det) N_2M_{obj} (Adb) + AUX
(Det) N_1M_{subj} (Det) N_2M_{com} (Adb) ± V AUX
(Det) N_1M_{subj} (Det) N_2M_{obj} (Det) N_3M_{com}
(Adb) + V AUX

여기서, S는 문장, NP는 명사구, VP는 동사구, Det는 관형어, Adb는 부사어, M은 지표, M_{subj}은 주격지표, M_{obj} 은 목적격지표, M_{com}은 보격지표, M_o은 서술격지표, N_1 은 첫 번째 명사(주어), N_2 는 두 번째 명사(목적어나 보어), N_3는 세 번째 명사(보어), +V는 동사(+행위), -V는 동사(-행위), AUX는 어미(Auxiliary), ( )는 임의적인 꾸밈말 등을 의미한다.

한국어의 기본 문형은 구절구조 문법에 입각하여 제시되었다. 문형을 보면 한국어의 문장은 명사구와 동사구와 구성되어 있다. 그리고 명사구는 (1) 관형어 + 주어 역할을 하는 명사 + 주격지표(조사), (2) 관형어+ 목적어나 보어 역할을 하는 명사 + 목적격이나 보격 지표(조사), 혹은 (3) 관형어 + 목적어나 보어 역할을 하는 명사 + 보어 역할을 하는 명사 + 보격지표(조사)로 구성되는 것을 알 수 있다. 동사구는 (1) 부사어 + 동사 + 어미, 혹은 (2) 관형어 + 목적어나 보어 역할을 하는 명사 + 부사어 + 동사 + 어미 가 반복되어 구성되는 것을 알 수 있다. 한국어 파서를 구성할 때에는 이러한 관계를 파악하여 구성하여야 할 것이다.

7. 표현 패턴에 의한 한국어-영어 기계 번역

지금까지는 하위범주화 패턴 방식의 한국어-영어 기계 번역에 대하여 논의하였다. 즉, 한국어 표면 문장을 동일한 의미를 나타내는 영어 표면 문장으로 변환시키기 위하여 트랜스퍼 사전에 수록된 한국어 동사와 형용사의 하위범주화 패턴을 활용하였다.

그러나 한국어 표현 패턴을 의미가 동일한 영어 표현 패턴과 1:1로 대응시켜서 표현 패턴 사전을 구성하고 이를 기계 번역에 활용하는 방법을 생각할 수 있다. 다음의 한국어 문장의 예를 보자.

(문장 10) 그는 아침을 먹고 출근하였다.
(문장 11) 그는 오늘 아침을 거르고 출근하였다.
(문장 12) 우리는 역사에 대하여 논의하였다.
(문장 13) 우리는 개념에 대하여 조사하였다.
(문장 14) 우리는 개념을 조사하였다.
(문장 15) 그는 더위를 먹었다.
(문장 16) 그는 더위를 심하게 먹었다.

(문장 10)을 영어로 번역하면 "He went to office after having breakfast."가 될 것이다. 그러나 한국어 문장에는 실제로 "사무실(office)"이라는 단어도 없고 "아침밥(breakfast)"이라는 단어도 없다. 한국어에서 "출근하였다"라는 단어는 "사무실에 일하러 간다"는 의미이며, "아침을 먹다"라는 동사구 속에는 "아침 식사를 하였다"는 의미가 포함되어 있다. 따라서 동일한 의미를 나타내는 한국어 표현과 영어 표현은 표현 패턴이 다른 것이다. 따라서 이 경우에는 "아침을 먹고 출근하다." 라는 한국어 표현 패턴 전체를 "go to office after having breakfast"라는 영어 표현 패턴 전체로 번역하면 될 것이다. (문장 11)의 경우에는 "Today he went to office without having breakfast."로 번역이 될 것이다. 한국어 타동사 "거르다"를 영어 전치사 "without"으로 번역한 셈이다. (문장 12)의 경우에는 "역사에 대하여"라는 구문이 있다. 이 구문에는 한국어에 있어서 조사와 동사로 구성된 "에 대하여"라는 표현 패턴이 있다. 이 표현 패턴은 영어로는 "about"이라는 하나의 전치사로 번역이 될 수 있다. 그러나 (문장 13)의 경우에 "We checked the cause of accident."로 번역할 수 있다. 이 경우에는 "에 대하여"를 "about"으로 번역하면 안 된다. (문장 14)의 경우도 영어로 동일하게 번역할 수 있다. 따라서 한국어의 "에 대하여" 표현 패턴이 항상 영어의 "about"으로 번역되는 것은 아니라는 것을 알 수 있다. 영어로 번역되는 동사가 타동사인가 자동사인가가 중요한 요소로 작용하는 것을 알 수 있다. 이 문제는 표현 패턴이 지배하는 문장 성분의 문법적 성분과 기능을 표현 패턴과 함께 표시함으로서 해결할 수 있다. 사전 내용의 비교를 위하여 하위범주화 패턴 방식과 표현 패턴 방식을 함께 표시하였다. 영어에 한국어의 "조사하다"와 유사한 의미의 단어에는 check, investigate, examine, inquire into, probe 등이 있다고 할 수 있다. 아래의 예에서는 이들 단어들을 모두 나열하였다. 이러한 경우에 어느 표현 패턴을 선택하여 영어 문장을 생성하는가가 중요한 문제가 될 수 있다. 이것은문체와 관련이 있다고 할 수 있다. 예를 들어, 문체를 평범한 문체와 격식을 갖춘 문체 등으로 구분하여 생각할 수 있다. 영어 표현 "inquire into"는 다른 영어 표현에 비하여 비교적 격식을 갖춘 표현이라고 한다. 아래 사전의 내용에는 표현 패턴에 (N_2(noun obj)) 라고 표시하여 표현 패턴의 지배를 받는 N_2가 명사(noun)이고 목적어(obj)라는 것을 함께 나타내었다. 이것은 단수와 복수 일치와 시제 때문에 필요하다.

●논의하다 : 동사
. 하위범주화 패턴1 : ($N_1 M_{subj}$ $N_2 M_{adv(에)}$ $V_{(대하여)}$)
. 영어 번역 : N_1 discuss about N_2
. 표현 패턴1 : (($N_1 M_{subj}$ N_2Pattern$_{(에\ 대하여)}$) (N_2(noun obj)))
. 영어 번역 : N_1 discuss about N_2

●조사하다 : 동사
. 하위범주화 패턴1 : ($N_1 M_{subj}$ $N_2 M_{adv(에)}$ $V_{(대하여)}$)
. 영어 번역 : N_1 check N_2
. 하위범주화 패턴2 : ($N_1 M_{subj}$ $N_2 M_{obj(를)}$)
. 영어 번역 : N_1 check N_2
. 표현 패턴1 : (($N_1 M_{subj}$ N_2Pattern$_{(에\ 대하여)}$) (N_2(noun obj)))
. 영어 번역 : N_1 check N_2
. 표현 패턴2 : (($N_1 M_{subj}$ N_2Pattern$_{(에\ 대하여)}$) (N_2(noun obj)))
. 영어 번역 : N_1 investigate N_2
. 표현 패턴3 : (($N_1 M_{subj}$ N_2Pattern$_{(에\ 대하여)}$) (N_2(noun obj)))
. 영어 번역 : N_1 examine N_2
. 표현 패턴4 : (($N_1 M_{subj}$ N_2Pattern$_{(에\ 대하여)}$) (N_2(noun obj)))
. 영어 번역 : N_1 inquire into N_2
. 표현 패턴5 : (($N_1 M_{subj}$ N_2Pattern$_{(에\ 대하여)}$) (N_2(noun obj)))
. 영어 번역 : N_1 probe N_2

(문장 15)는 앞에서 트랜스퍼 사전 방식으로도 다루었다. 그러나 "더위를 먹다" 전체를 표현 패턴으로 간주하여 "suffer from a hot weather" 표현으로 표현 패턴 방식에 의하여 번역할 수 있다. (문장 16)의 경우는 "더위를 심하게 먹다" 전체를 표현 패턴으로 간주하여 "suffer from a hot weather severely"로 번역할 수 있다.

표현 패턴은 명사구, 동사구, 혹은 그 밖의 문장 표현에 존재할 수 있으며 관형어도 넓은 의미에서는 표현 패턴으로 간주할 수 있다. 따라서 기존의 하위범주화 패턴에 의한 기계 번역 방식을 확장하여 표현 패턴에 의한 기계 번역 방법을 구성할 수 있다. 그러나 표현 패턴에 의한 기계 번역 방법의 효용성에 대하여서는 추후 연구와 조사가 필요하다. 표현 패턴 사전은 기존의 트랜스퍼 사전에 구성할 수가 있으며 또한 별도의 사전으로 구성할 수도 있다. 트랜스퍼 사전의 하위범주화 패턴이 구문적인 관점에 좀 더 치중하여 두 언어 간의 차이점을 다룬 것이라면, 표현 패턴은 표현과 의미적인 관점에 좀 더 치중하여 두 언어 간의 차이점을 다룬 것이라고 할 수 있다. 표현 패턴의 장점은 일단 이해가 용이하고 사전에 등재시키기도 비교적 용이하며 번역에 있어서의 효과도 매우 크다는 것이다. 한국어에 있어서는 조사와 어미 활용이 상당히 복잡하고 때로는 매우 미묘한 의미의 차이를 나타내기도 하는데, 표현 패턴은 이러한 한국어의 독특한 언어적 현상을 하나의 단위로 인식하여 다루므로, 영어를 대상으로 개발된 하위범주화 개념보다 기계 번역에 있어서는 한국어에 더욱 적합한 개념이라고 할 수 있다. 예를 들어, 앞의 (문장 29)에서 사용한 "생각보다도"와 같은 한국어 표현은 명사와 조사와 이루어진 간단한 표현인데도 영어로 번역하면 "although thinking hard" 혹은 주어를 포함하면 "although I think hard"가 된다. 이 표현은 앞에서는 아직 표현 패턴을 도입하지 않은 상황에서 편의상 관용적 표현으로 다루었지만 사실은 한국어에서 관용적 표현은 아니다. 따라서 이러한 한국어 표현은 하나의 표현 패턴으로 간주하여 다루면 좋을 것이다. 그러나 표현 패턴에 의한 방법에도 최종적으로 한국어 문장과 영어 문장을 구성하는 어순이 다르기 때문에 어순 문제, 그리고 시제 문제와 단수/복수 일치 문제 혹은 그 밖에 문체의 문제 등을 해결해 줄 문법 검사 및 변환 규칙은 필요하다고 본다. 다음은 표현 패턴의 정의와 기능에 대한 설명이다.

[표현 패턴의 정의] : 표현 패턴은 한국어-영어 기계 번역에 있어서 표현과 의미의 차이를 나타내는 한국어 단어 문자열의 최소 단위로 정의한다. 즉, 한국어에서 중요한 문법적 기능을 수행하는 조사와 어미 등을 다른 단어들과 함께 문자열로 표기하여 동사의 하위범주화 패턴형태로 나타낸다. 표현 패턴은 표현의 차이를 나타내는 단위로서 기계 번역에서 구문 변환과 의미 전달의 기능을 수행한다.

표현 사전을 사용하여 기계 번역을 수행한다는 관점에서는 기계 번역 수행 과정을 다음과 같이 생각해 볼 수 있다.

(1) 한국어 문장을 구문 분석하여 문법을 검사하고 구문 트리를 구성한다. 문장의 표면 문자열을 표현 패턴 사전에서 찾아서 한국어 표현에 대응하는 영어 표현을 찾는다. 표현 패턴 사전은 동사, 명사, 형용사, 부사 순서로 찾는다. 표현 패턴 사전에서 찾을 수 없는 표현은 일반 한영사전을 찾아서 한국어 단어에 대응하는 영어 단어를 찾는다. 최종적인 영어 문장 생성을 위하여 생성된 영어 표현 패턴의 문법적인 기능 정보는 유지하여야 한다.
(2) 영어 생성기는 어순 문제, 시제 문제, 그리고 단수와 복수 일치 문제 등을 해결한다.
(3) 영어 생성기는 최종적인 영어 문장을 생성한다.

또한 표현 패턴에 의한 기계 번역 방식을 기존의 트랜스퍼 방식을 확장하여 구성할 수 있다. 처음에는 트랜스퍼 사전에 입각하여 한국어 구문의 분석을 시도하고 기계 번역을 수행한다. 만약 트랜스퍼 사전에서 적절한 영어 번역이 이루어지지 않으면, 다음에는 표현 패턴 사전에 의하여 다시 표면 문장의 문자열 비교를 통하여 기계 번역을 시도한다. 반면에 트랜스퍼 사전에 의한 기계 번역 방식을 지양하고 전적으로 표현 패턴 사전에 의한 기계 번역을 시도할 수 있다. 그러나 표현 패턴 방식에 의한 기계 번역이 성공을 거두기 위해서는 상당한 규모의 한국어 표현 패턴과 영어 표현 패턴 사이에 대한 대응 연구가 선행되어야 할 것이다.

8. 최근의 기계 번역 연구

근래에 중국어-영어 기계 번역 연구가 중국어권에서 비교적 활발히 진행되고 있다[33][34][35][36]. 국내에서는 2002년도에 전자통신연구소(ETRI)에서 한국어-중국어 기계 번역 시스템이 개발되기도 하였다[37]. 또한 영어-일본어 기계 번역 연구도 진행되고 있다[38]. 미국에서는 2005년도에 통계적인 방법을 사용한 한국어-영어 기계 번역 시스템이 개발되기도 하였다[39].

상품화된 기계 번역 시스템은 IBM[40], Microsoft[41], Systran[42] 등의 기관에서 제공하고 있다. 그리고 그 밖의 기관에서 기계 번역 관련 자료와 연구용 소프트웨어를 제공하고 있다[43][44][45]. 영어의 경우에는 [46][47]에서 통계적 방법을 사용하여 구현된 구문 분석기(statistical syntax analyzer)

에 대한 자료와 소프트웨어를 제공한다.

본 논문에서 논의한 표현 패턴에 의한 기계 번역은 새로운 시도로서 단어들 간의 문맥의존성(context-sensitivity)을 현저히 감소시키고 문맥자유성(context-freeness)을 증가시키려는 시도이다. 문맥자유성이 증가되면 자연언어가 문맥자유(context-free) 언어가 되어 기계에 의한 처리가 용이해지게 된다. 그러나 텍스트 전반에 걸친 문맥 구조, 의미 구조, 담론 구조 그리고 텍스트의 배경이 되는 의식 구조 등을 기반으로 한 기계 번역에 대한 연구도 논의될 필요가 있다.

9. 문장 분할 개념

문장 분할(sentence partition 혹은 segmentation) 개념은 한국어 문장 구성의 특징상 "주어+동사" 구문이 여러 번 반복되어 나타나는 경우가 많기 때문에 필요하다고 생각한다. 그리고 이것은 앞 절에서 설명한 조사와 어미 그리고 한국어 문장의 비교적 평면적인 구문 구조에 기인한다고 할 수 있다. 예를 들어 "나는 아침에 일어나서 세수를 하고 밥을 먹고 차를 타고 학교에 가서 공부를 하고 운동을 하고 공부가 끝난 후에 집에 와서 깨끗이 씻고 숙제를 하고 저녁을 먹고 잠을 잤다."와 같은 문장이 한국어에서는 가능하다. 영어에서는 이러한 긴 문장 구성이 가능하지 않다. 기계 번역의 관점에서 긴 문장은 구문적인 관점에서건 의미적인 관점에서건 다루기 어렵다. 따라서 이러한 긴 문장은 문장 분할 기법을 적용하여 문장을 다룰 수 있는 적절한 길이로 구분한 다음에 기계 번역을 수행하는 것이 적절한 표현의 영어 문장을 생성하는데 용이할 것이다. 부분 파싱과 문장 분할 개념의 필요성은 바로 한국어의 평면적인 문장 구조와 조사의 용법과 어미 활용의 발달로 인하여 문장 구성 요소의 의미와 역할이 직접적으로 표시 가능하기 때문인 것으로 이해가 된다. 한국어 파서를 개발할 때에는 이러한 문법적 기능을 나타내는 굴곡법, 준굴곡법, 영굴곡법 등을 모두 이해하고 이를 제대로 반영하는 파서를 개발하는 것이 올바른 방향이라고 생각한다. 즉, 굴절어를 위한 서구어(영어)의 문법 이론에 의하여 파서를 개발하는 것이 아니라 교착어(첨가어)인 한국어의 문법적 특징에 가장 알맞은 방법으로 파서를 개발하는 것이다. 예를 들어, 위의 문장은 다음과 같이 4개의 문장으로 분할하여 나타낼 수가 있다. "나는 아침에 일찍 일어나서 세수를 하고 밥을 먹었다." "그리고 차를 타고 학교에 가서 공부를 하고 운동을 하였다." "공부가 끝난 후에 집에 와서 깨끗이 씻었다." "그 다음에 숙제를 하고 저녁을 먹고 잠을 잤다." 사실 이렇게 문장을 분할하면 이해하기도 쉽고 다루기도 쉽다. 다만 분할하는 과정에서 생략되는 요소들이 있고, 반면에 접속사와 같이 추가되는 요소들도 있다. 이렇게 생략되거나 추가되는 요소들은 의미 구조에 포함하여 나타내고 다루어야 할 것이다. 위 4개의 문장을 영어로 번역을 하면 "I got up early in the morning, washed my face, and had a breakfast.", "Then, I went to school by car and studied and took part in sports activity.", "After finishing school, I went back home and took a shower.", "I did my homework, had supper, and went to sleep.". 이 예에서도, (1) "차를 타고 학교에 가다-go to school by car", (2) "운동을 하다-take part in sports activity", (3) "공부가 끝난 후에-after finishing school", (4) "깨끗이 씻다-take a shower", (5) "잠을 자다-go to sleep" 과 같은 한국어와 영어 사이에 대응하는 표현 패턴을 발견할 수 있다. 문장 분할 개념과 관련된 것으로 부분 파싱(partial parsing) 개념이 있다. 이 개념도 문장의 길이가 길어서 구조가 복잡한 경우에는 문장 전체를 파싱할 수가 없고 문장의 부분만을 파싱할 수 있다는 것이다. 따라서 부분 파싱 개념은 부분 파싱만이 가능한 긴 문장을 짧은 여러 개의 문장으로 분할하여 완전하게 파싱하는 기법과 관련이 있다. 문장 분할 개념은 파싱의 복잡도를 낮추는데도 크게 도움이 될 것으로 생각한다.

10. 새로운 한국어 구문 분석 시스템

참고 문헌 [2][3][4][6][10][13][15][17][18][19]와 [21]~[28]에는 현대 한국어의 문법이 상세하게 논의되어 있다. 현대 한국어 문법에서는 한국어의 교착어적인 특징을 분명히 이해하고 이를 말뭉치를 통하여 밝히는 방향으로 연구가 진행되는 경향이 있다. 참고 문헌 [29]는 유용한 자료로서 연구의 목표로는 "한국어 문장 분석에 설정되어야 할 기본적인 통사 단위와 그것이 결합되어 형성되는 통사 구성을 범주화하고 체계화하는 것이다."라고 하였다. 5장에서는 조사의 격과 어미의 활용 등에 대하여 거의 완벽한 자료 정리 결과를 제시하고 있다. 참고문헌 [48]~[53]에는 90년대의 한국어 구문 분석 시스템에 대한 연구가 제시되어 있다. 참고 문헌 [52][53][54][60]는 의존 문법에 기반을 둔 한국어 구문 분석기 개발에 대한 연구가 논의되었다. 참고 문헌 [54]~[65]는 2000년대의 연구 결과를 제시하고 있다. 특히 최근에 들어서 한국어 구문 분석에 대한 연구가 다시 진행되고 있다는 것은 주목할 만하다[58]~[65]. 참고 문헌 [62]는 한국어의 병렬 명사구 구조를 나타내는 확률적 모델을 제시하였다. 모델은 병렬 구조의 대칭성(symmetry)과 상호교환성(interchangeability)에 근거하여 구성하였다. 참고 자료 [63]의 한국어 구문 분석 시스템은 주목할 만하다. 이 시스템은 의존 문법에 기반을 두고 개발되었다. 시스템은 문장 단위로 구문 분석을 실행하여 결과를 보여준다. 참고 문헌 [64][65]는 시스템 [63]에 관련된 논문들로서 확률적 의존 문법을 사용하여 한국어 구문 분석기를 구현한 내용을 기술하고 있다. 최근에는 어휘 정보(lexical information)가 구문 분석의 정확성 향상에 일반적으로 믿는 것만큼은 기여하는 바가 없다는 연구 결과도 있다[64][66][67]. 심지어 참고 문헌 [68]에서는 깊고 복잡한 문법 분석이 간단한 문법 분석에 비하여 생각보다는 효과가 적다는 논의도 있다. 아마 이것은 어차피 자연언어는 의미에 의하여 좌우되는 부분이 많기 때문에 깊은 문법 분석 자체가 완벽할 수 없고 또한 큰 역할을 하지 못한다는 의미일 것이다.

참고 문헌 [29]의 22쪽에는 한국어 구문 분석에 있어서 HPSG나 LFG 관련 연구가 줄어든 원인으로는 "HPSG나 LFG는 국어에서 조사나 어미가 가진 핵의 성격을 부각시키기 어렵다."고 설명하고 있다. 이것은 한국어는 교착어이기 때문에 서구 언어(영어)를 설명하기 위하여 개발된 문법 이론은 한국어의 언어 현상을 설명하는 데는 적합하지 않다는 것을 의미한다고 볼 수 있다.

따라서, 새로 개발하는 한국어 구문 분석 시스템의 요구 사항은 다음과 같이 정리할 수 있다. 즉, 새로운 시스템은 언어적인 측면에서는 한국어의 교착어적인 문법 특징(참고 문헌 [24]에서 제시한 특징 포함)을 충실히 고려하여 개발되어야 할 것이다. 교착어적인 특징은 보통 순서가 자유로운(free word order) 그리고 핵이 뒤에 있는(head-final) 등의 용어로 표현된다. 이것은 당연한 것이다. 왜냐하면, 문법 기능을 나타내는 조사와 어미가 뒤에 있음으로서 순서가 자유로운 언어가 될 수 있는 것이기 때문이다. 참고 문헌 [29]의 연구 결과 자료는 중요한 자료로서 고려되어야 할 것이다. 여기에 참고 문헌 [21][27]의 화용론 개념도 고려되어야 할 것이다. 즉, 저자는 한국어의 문장 구성 체계는 의미 표현 단위(MEU, meaning expression unit)의 구성과 이들 간의 의존(연관) 관계로 되어 있다고 생각한다. 그리고 MEU들의 문법적인 기능은 조사나 어미의 활용이 나타낸다고 생각한다. MEU는 간단하게는 하나의 단어에서부터 구절 구조(phrase structure), 그리고 복잡하게는 긴 문장까지 될 수 있다고 생각한다. 이러한 MEU를 참고 문헌 [54]에서는 단순히 단위(chunk)라고 표현하였다. 그러나 논문 [54]에서 설명한 단위(chunk)보다는 훨씬 복잡하고 다양한 구조를 나타낼 수 있다. 따라서 MEU의 구성과 MEU들 간의 의존(연관) 관계를 밝히면 한국어에 대한 문법 체계를 구성할 수 있을 것으로 가정해 본다. 단어의 품사 범주를 중요하게 고려하지 않는다는 것이다. 그리고 의존(불완전) 명사에 대한 고려도 매우 중요하다. 예를 들어서 다음의 문법을 보면 일부분의 한국어 문장이기는 하지만 한국어 문장을 얼마나 간단하게 나타낼 수 있는지를 알 수 있다. DEP는 의존 관계를 나타내며 전체 문장을 나타낸다. 또한 이런 방식으로 문법을 구성하면 표현 패턴과 문장 분할을 구현하기가 용이할 수 있다.

```
MEU -> Det              // 관사
MEU -> NJosa            // 명사 조사
MEU -> VOmi             // 동사 어미
MEU -> ADJOmi           // 형용사 어미
MEU -> ADV + VOmi       // 부사 + 동사 어미
MEU -> MEU + MEU        // MEU 의 결합
DEP -> MEU + MEU        // 전체 문장
```

다음은 위의 문법을 적용하여 문장의 구조를 나타낸 것이다.

(문장 10) 그는 아침을 먹고 출근하였다.

MEU_1 -> MEU(아침을) + MEU(먹고)
MEU_2 -> MEU(그는) + MEU_1
DEP -> MEU_2 + MEU(출근하였다)

((그는) ((아침을 먹고)) (출근하였다))

(문장 16) 그는 더위를 심하게 먹었다.

MEU_1 -> MEU(심하게) + MEU(먹었다)
MEU_2 -> MEU(더위를) + MEU_1
DEP -> MEU(그는) + MEU_2

((그는) ((더위를) (심하게 먹었다)))

(문장 5) 코끼리는 코가 길다.

MEU_1 -> MEU(코가) + MEU(길다)
DEP -> MEU(코끼리는) + MEU_1

((코끼리는) (코가 길다))

(문장 17) 나는 그것을 알 수가 있다.

MEU_1 -> MEU(그것을) + MEU(알)
MEU_2 -> MEU_1 + MEU(수가)
MEU_3 -> MEU_2 + MEU(있다)
DEP -> MEU(나는) + MEU_3

((((나는) (그것을 알)) (수가)) (있다))

(문장 18) 나는 책과 연필을 샀다.

MEU_1 -> MEU(책과)
MEU_2 -> MEU_1 + MEU(연필을)
MEU_3 -> MEU_2 + MEU(샀다)
DEP -> MEU(나는) + MEU_3

((나는) ((책과 연필을) (샀다)))

(문장 19) 몇 일 후에 그가 돌아오지?

MEU_1 -> MEU(몇) + MEU(일)
MEU_2 -> MEU_1 + MEU(후에)
MEU_3 -> MEU(그가) + MEU(돌아오지)
DEP -> MEU_2 + MEU_3

((((몇) (일)) (후에)) ((그가) (돌아오지)))

(문장 20) 사고가 나자 에스컬레이터에 타고 있던 승객 1백 여명이 옆으로 급히 뛰어내리는 소동을 빚었다.

MEU_1 -> MEU(사고가) + MEU(나자)
MEU_2 -> MEU(에스컬레이터에)
+ MEU(타고) + MEU(있던)
MEU_3 -> MEU(승객) + MEU(1백 여명이)
MEU_4 -> MEU_2 + MEU_3
MEU_5 -> MEU(옆으로) + MEU(급히)
+ MEU(뛰어내리는)
DEP -> MEU_1 + MEU_4 + MEU_5 + MEU(소동을 빚었다)

((사고가 나자) ((에스컬레이터에 타고 있던)
(승객 1백여명이)) (((옆으로 급히 뛰어내리는)
(소동을)) (빚었다)))

(문장 21) 오늘따라 하늘이 맑다.

MEU_1-> MEU(오늘따라) // "따라"의 의미?
MEU_2-> MEU(하늘이) + MEU(맑다)
DEP -> MEU_1 + MEU_2

((오늘따라) (하늘이 맑다))

그리고 새로운 구문 분석 시스템은 또 다른 측면으로는 7절의 표현 패턴 개념, 9절의 문장 분할 개념, 그리고 구문 분석 과정에 의미 구조를 포함하는 개념 등을 모두 포함하여 개발되어야 할 것이다. 의미 구조는 표현 패턴 개념이나 특히 문장 분할 개념을 적용하는 경우, 생략된 문장 성분의 복원이나 새로 추가할 문장 성분이 발생하는 경우에 이에 대한 의미 정보를 표시하기 위하여 필요하다. 또한 조사나 어미의 의미를 표시할 때에도 필요하다.

11. 결 론

한국어는 문장의 의미를 결정하는 중요 요소가 단어뿐만이 아니고 조사, 어미, 그리고 문장의 뒤 부분에 위치하는 조동사와 동사에도 나타난다. 그러므로 한국어 구문 분석에 있어서는 구문 자체의 분석과 함께 조사, 어미, 그리고 동사의 의미 분석도 매우 중요하다. 근래에 많은 국어 학자들에 의하여 한국어의 교착어로서의 특징이 분명히 인식되고 이를 반영하는 국어 문법 체계가 발전되고 있는 것은 바람직하다고 하겠다. 이러한 연구 결과와 분위기를 바탕으로 새로운 한국어 파서를 개발하는 것은 큰 의의를 가진다고 하겠다.

또한 본 논문은 한국어 문장을 고찰하여 표현 패턴에 의한 새로운 기계 번역 방식을 제안하였다. 표현 패턴은 한국어와 영어의 표현을 대조하여 더욱 정확한 한국어-영어 기계 번역을 위하여 정의하였다. 또한 표현 패턴과 문장 분할을 수행하는 과정에서 파싱 과정에 의미 구조를 포함시키는 새로운 한국어 구문 분석 방법도 제안하였다. 이 방법은 기계 번역의 관점에서는 시도해 볼 만한 방법이라고 생각한다.

앞으로의 연구로는 언어적인 관점에서 표현 패턴의 구체적인 연구가 필요하다. 그리고 한국어의 교착어 특징, 표현 패턴 개념, 문장 분할 개념 그리고 의미 구조의 포함 등을 고려한 새로운 한국어 파싱 시스템의 개발이 필요하다.

참고 문헌

[1] 김영희, 한국어 통사 현상의 의의, 도서출판 역락, 2005.
[2] 김하수, 국어 형용사의 연구, 도서출판 역락, 2005.
[3] 이정복, 국어 부사절의 성립, 태학사, 2004.
[4] 엄정호, 국어의 보문과 보문자, 태학사, 2005.
[5] 노마 히데키, 한국어 어휘와 문법의 상관 구조, 태학사, 2003.
[6] 이관규, 국어문법연구 I, 경남 : 도서출판 역락, 2003.
[7] 임지룡 외 6인, 한반도와 만주의 역사 문화, 서울대학교 출판부, 2003.
[8] 한길, 현대 우리말의 형태론, 도서출판 역락, 2006.
[9] 최남희, 고대국어 형태론, 도서출판 박이정, 1996.
[10] 고영근 외 35인, 문법과 텍스트, 서울대학교출판부, 2002.
[11] 배행주, 한국어 문장구조의 이해, 푸른사상, 2005.
[12] 박영순, 한국어의 문장, 도서출판 박이정, 2001.
[13] 이익섭, 한국어 문법, 서울대학교출판부, 2007.
[14] 강현화, 진자영, 이재성, 임효상, 대조분석론(한국어.스페인어 문형 대조를 바탕으로), 도서출판 역락, 2003.
[15] 김광호, 현대표준문법 본, 한국문화사, 2003.
[16] 이광호, 현대국어문법 본, 태학사, 2004.
[17] 하치보, 현대우리말본, 도서출판 박이정, 2002.
[18] 최경봉, 국어 명사의 의미 연구, 태학사, 1998.
[19] 민현식, 현대 국어 자동사 연구, 한국문화사, 2000.
[20] 민현식, 국어 문법 연구, 도서출판 역락, 1999.
[21] 박영순, 한국어 화용론, 도서출판 박이정, 2007.
[22] 남기심, 국어 조사의 용법(-에 와 -로 를 중심으로), 도서출판 박이정, 1996.
[23] 목정수, 한국어 문법론, 도서출판 월인, 2003.
[24] 김상대, 국어 문법의 대안적 접근, 국학자료원, 2001.
[25] 김주태, 마침법 씨끝의 융합과 그 한계, 도서출판 박이정, 2006.
[26] 이정택, 현대국어 피동 연구, 도서출판 박이정, 2004.
[27] 이정애, 국어 회용표지의 연구, 도서출판 월인, 2002.
[28] 안효경, 현대국어의 의존명사 연구, 도서출판 역락, 2001.
[29] 임홍빈, 이홍식, 한국어 구문 분석 방법론, 한국문화사, 2003.
[30] 국립국어원, 21세기 세종계획, 2007.
[31] George A. Miller, "WordNet : A Lexical Database for English," CACM, Vol. 38, Issue 11, November 1995.
[32] Christiane Fellbaum (editor), WordNet : An Electronic Lexical Database, The MIT Press, 1998.
[33]Yonggang Deng, William Byrne, "HMM Word and Phrase Alignment for Statistical Machine Translation," IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, No. 3, pp.494-507, March 2008.
[34] Xiao-Shen Zheng, et al., " On difficulties of Chinese-English machine translation," Proc. of 2002 Int'l conference on machine learning and cybernetics.
[35] W Byrne, et al., "The Johns Hopkins University 2003 Chinese-English Machine Translation System, http://www.google.com.
[36] Dan Su, Target-Domain Chinese-English Machine Translation, Thesis of Master of Science, Brigham Young University, 2004.
[37]Changhyun Kim, Munpyo Hong, Yinxia Huang, Young Kil Kim, Sung Il Yang, Young Ae Seo, and Sung-Kwon Choi, "Korean-Chinese Machine Translation System based on Verb Patterns," LNCS, Vol. 2499, Springer-Verlag, 2002.
[38] Pi-Chuan Chang, Kristina Toutanova, "A Discriminative Syntactic Word Order Model for Machine Translation," Proc. of the 45th Annual Meeting of the ACL, June 2007.
[39] Yookyung kim, et al., "Senda S-MT : Incorporation of Syntax into Statistical Translation System," http://www.fluentialinc.com(Proc. of IWSLT-05 Workshop), 2005.
[40] http://www-306.ibm.com/software/pervasive/tech/demos/translation.shtml.
[41] http://research.microsoft.com/nlp/Projects/MTproj.aspx.
[42] http://www.systran-software.co.kr.
[43] http://www.statmt.org.
[44] http://www.mt-archive.info.
[45] http://www.word2word.com.
[46] http://www.codeproject.com.
[47] http://homepages.inf.ed.ac.uk/jhender6/parser/ssn_parser.html
[48] 나동열, "한국어 파싱에 대한 고찰", 한국정보과학회지, 제 12권, 제 8호, 1994년 9월.
[49] 임희경 외 8인, "기계번역 효율 향상을 위한 한국어 구문분석의 테이블 파싱 알고리즘", 한국정보과학회 봄 학술발표논문집, Vol. 21, No. 1, 한국정보과학회, 1994년.
[50] 신대규, 엄미현, 나동열, "확장 문맥자유문법을 위한 양방향 파싱 알고리즘", 한국정보과학회 가을 학술발표논문집, Vol. 23, No. 2, 한국정보과학회, 1996년.
[51] 윤준태, 김선호, 송만석, "전역적 연관표를 이용한 한국어 구문분석," 한국정보과학회 논문지(B), 제24권, 제 11호, 1997년 11월.
[52] 홍영국, 이종혁, 이근배, "의존문법에 기반을 둔 한국어 구문 분석기," 한국정보과학회 봄 학술발표대회논문집, Vol. 20, No. 1, 1993년 4월.
[53] 홍영국, 이종혁, "한국어 의존 해석을 위한 형태-통계적 품사 분류 체계," 한국정보과학회논문지(B), 제 22권, 제 9호, 1995년 9월.
[54] 김미영, 강신재, 이종혁, "단위(Chunks) 분석과 의존 문법에 기반한 한국어 구문 분석," 한국정보과학회 봄 학술발표대회논문집, Vol. 27, No. 1, 2000년 4월.
[55] 김영일, 양성일, 서영애, 김창현, 홍문표, 최승권, "한영 자동 번역을 위한 한국어 구문 분석 전처리," 한국정보과학회 가을 학술발표대회논문집, Vol. 28, No. 2, 2001년 10월.
[56] 박소영, 옥철영, 정후중, 홍영숙, 임해창, "한국어 구문 분석의 효율성을 개선하기 위한 구문 제약 규칙의 학습," 한국정보과학회논문지:소프트웨어및 응용, 제 29권 제 10호, 2002년 10월.
[57] 박정열, "2단계 방법을 사용한 빠른 한국어 TAG 구문분석기 구현," 한국정보과학회 가을 학술발표논문집, Vol. 31, No. 2, 한국정보과학회, 2004년.
[58] 김성용, 이공주, 최기선, "복합 레이블을 적용한 한국어 구문 규칙," 한국정보과학회논문지:소프트웨어및 응용, 제 31권 제 2호, 2004년 2월.
[59] 이성욱, 서정연, "한국어 문법 관계에 대한 부분 구문 분석," 한국정보과학회논문지:컴퓨터시스템및 응용, 제 32권 제 10호, 2005년 10월.
[60] 오연분, 송영인, 박소영, 옥철영, 임해창, "지배가능 경로 문맥을 이용한 의존 구문 분석의 수식 거리 모델," 한국정보과학회논문지:소프트웨어및 응용, 제 34권 제 2호, 2007년 2월.
[61] 이형영, 이용석, "내포문의 단문 분할을 이용한 한국어 구문 분석," 한국정보과학회논문지:소프트웨어및응용, 제 35권 제 1호, 2008년1월.
[62] 최용석, 신지애, 최기선, "확률모형과 수식정보를 이용한 와/과 병렬 영자구 범위결정," 한국정보과학회논문지 : 소프트웨어및 응용, 제35권, 제 2호, 2008년 2월.
[63] 고려대학교 컴퓨터학과 자연언어처리연구실, http://nlp2.korea.ac.kr/~hjchung/parsedemo.
[64] Hoojung Chung, Hae-Chang Rim, "Unlexicalized Dependency Parser for Variable Word Order Languages based on Local Contextual Pattern," LNCS 2945, pp.112-123, 2004.
[65] Hoojung Chung, Hae-Chang Rim, "A New Probabilistic Dependency Parsing Model for Head-Final, Free Word Order Languages," IEICE Transaction on Information & System, Vol. E86-D, No. 11, pp.2490-2493, 2003.
[66] Slav Petrov, Dan Klein, "Improved Inference for Unlexicalized Parsing," http://nlp.cs.berkeley.edu/Main.html#Parsing.
[67] Dan Klein, Christopher D. Manning, "Accurate Unlexicalized Parsing," Proc. of the 14th Annual Meeting on Association for Computational Linguistics, pp.423-430, Japan, 2003.
[68] Peter Dienes, Amit Dubey, "Deep Syntactic Processing by Combining Shallow Methods," Proc. of the 14th Annual Meeting on Association for Computational Linguistics, pp.431-438, Japan, 2003.