

HPSG 기반 한국어 문생성과 STYLE 자질*

송상헌

(Univ. of Washington)

Song, Sanghoun. 2009. HPSG-based Generation for Korean Sentences and STYLE Features. *Linguistic Research* 26(3), 67-90. HPSG (Head-driven Phrase Structure Grammar), a typical unification-based grammar, has proved itself to be a solid foundation for computational implementation, in terms of both parsing and generation. In particular, generation holds a significant position in recent grammar engineering using HPSG. This paper aims to try to generate Korean sentences on the basis of the Korean Resource Grammar (KRG), which is a computational grammar constructed for last several years within the HPSG framework using LKB (Linguistic Knowledge Builder). This paper, to begin with, gives an overall explanation of natural language generation, and then takes two major stages of generation into account; surface realizer and discourse planner. Building upon the fundamentals, this paper proposes how to generate Korean sentences from logical forms (Minimal Recursion Semantics, MRS), and how to choose the results in accordance with stylish properties (STYLE). The whole process has been implemented into computational systems as well, in order to check out its feasibility. The parsing module that this research employs is the PET parser, and the generation module is the LKB system. (Univ. of Washington)

Key Words generation, HPSG, KRG, PET, LKB, STYLE.

1. 서론

생성문법의 정의를 수학적으로 풀이하자면, 유한수의 규칙을 가지고 무한수의 문을 생성할 수 있는 체계라 할 수 있다. 촘스키의 여러 가설을 발전적으로 수용한 전산언어학의 입장에서 이 또한 마찬가지이기 때문에, 규칙 기반 또는 논리 기반의 자연어처리 제 방법론 역시 유한수의 규칙에 기반하여 무한수의 문장을 처리할 수 있는 시스템을 그 목표로 한다. 문법의

* 본고는 저자가 BK21 사업의 일환으로 실시된 글로벌 인턴쉽 수혜 대상자로 선정되어 일본 NiCT Language Infrastructure Group에서 진행한 연구에 기반하고 있다. 안정적인 연수를 지원해 준 Francis Bond, Kiyotaka Uchimoto, Kentaro Torisawa 및 NiCT 관계자 여러분들께 감사드린다. 또한 한국어 HPSG 연구를 적극적으로 수행할 수 있도록 물심양면의 지원을 해주신, 김종복, 양재형, 최재웅, 장석진, 이기용 교수님들께도 깊은 감사의 말씀을 올린다. 자료기반 HPSG 연구에 대해 귀중한 조언을 해 주신, 고베쇼인여자대학의 Takao Gunji 교수님과 교토대학의 Yukinori Takubo 교수님께도 감사의 인사를 올린다. 본 연구를 진행하는 데 있어, CSLI의 Dan Flickinger, NIST의 Eric Nichols, 조지아텍의 Darren Scott Appling, UW의 Michael Wayne Goodman, 고려대 임인빈, 그리고 명지대 은광희 선생님께서도 적지 않은 도움을 받았음을 밝힌다. 끝으로 익명의 심사자 여러분들께도 감사드린다. 단, 본고에서 발견되는 모든 오류와 문제점은 전적으로 저자의 몫임을 밝힌다.

역할이 ‘분석’¹⁾과 ‘생성’의 두 가지 측면으로 구분된다고 가정하면, 제안된 문법을 가지고 해당 언어의 가능한 모든 문장을 ‘분석’하여 ‘논리 형태’를 만들고, 역으로 이 논리 형태에 따라 실제 자연어 문장을 ‘생성’하는 쌍방향성이 가능할 때 (Allen, 1995:286), 비로소 그 문법은 본 연의 목표를 달성하였다고 볼 수 있다. 이는 이론언어학 및 자연어처리 양자가 공히 지니는 종국적 지향점이다.

지난 수십 년 동안 자연어처리 분야의 관심은 문의 ‘분석’에 주된 초점을 두고 연구되어 왔다. 이는 실제 자연어처리 시스템을 구현하는 데 있어 분석에 근간한 이해가 선결되지 않고서는 여타의 추가 연구가 불가하다는 점²⁾에서 일정정도 당연한 역사일 것이다. 또한, 중의성 해소와 같은 흥미로운 연구거리를 포괄하고 있는 ‘분석’에 비해, 비교적 단순한 결과만을 제공하여도 된다는 점³⁾에서 연구자들이 그동안 문생성 관련 분야에 큰 매력을 느끼지 못한바 역시 원인이 될 것이다.

그러나 생성은 실제 사용자와 가장 밀접하게 맞닿아 있는 부분이다. 기계번역을 예로 들어 보자면, 대부분의 경우 사용자는 주어진 문장의 형태·구문 분석이나 의미 표상이 어떠한 알고리즘을 통해 이루어지는지 알 수도 없고 또 굳이 알려고 하지 않는다. 다만, 그 결과로서 제시되는 번역문을 놓고 그 시스템의 활용도를 평가할 뿐이다 (김영택 외, 2001:307). 다시 말해, 제반 자연어처리 시스템을 실용화하는 단계에 있어서는 생성이 보다 주도적인 역할을 담당한다. 이러하기에 지난 수십 년의 전산언어학 연구에서 생성 분야는 비록 그 규모는 작지만 지속적인 검토 대상이 되어 왔다. 특히 최근에는 자질 구조와 통합이라는 기제를 중심 고리로 활용하는 전산 문법 분야에서 문생성은 새로운 각광을 받고 있다 (Bond *et al.* (2005), Oepen *et al.* (2007), Bond *et al.* (2008), Banik (2009) 등).

본고는 대표적인 통합기반 전산 문법인 HPSG(Head-driven Phrase Structure)를 대상으로 하여, 전산적 차원에서 한국어 문장이 어떻게 생성될 수 있는가에 대한 조망을 할 것이다. 우선, 분석과 대별하여 생성이 가지는 고려 사항들에 대한 점검을 할 것이며, 한국어 각 문장들이 HPSG에 기반하여 논리 형태로 변환되고 난 후 각 논리 형태가 복수의 한국어 문장들로 생성되는 과정을 설명할 것이다. 이러한 표층부 실현과는 별도로 실제 문생성에서는 발화 외적 요소도 중요하게 작용하는 바, STYLE 이라는 자질 구조를 상정하여 이 자질값이 생성 과정에 어떻게 관여하는지 제시할 것이다. 또한 전반적인 논의를 시스템상의 구현을 통해

1) 일반적으로 이는 ‘이해’라는 단어로도 사용된다. 그러나 본고에서는 ‘생성’과 분명한 구별점을 제시하기 위하여 ‘분석’이라는 표현을 사용하기로 한다.

2) 사람의 경우에도 말을 하는 것보다 부모의 말을 이해하는 것을 먼저 배우게 된다. 또한 말을 할 수 없는 동물의 경우에도 ‘앉아’, ‘먹어’ 등과 같은 제한된 표현의 이해는 가능하다. 이는 한편으로 분석이 생성보다 인지구조 형성에서 선행됨을 말하며, 다른 한편으로 생성이 분석보다 인지과정에 있어 보다 진화된 결과임에 대한 증거이다. 유추적으로 적용하면, 자연어처리 시스템을 구현할 때에도 ‘분석 모듈’이 ‘생성 모듈’보다 선행되어 개발되어야 함은 분명하나, ‘생성 모듈’을 갖추지 못하면 인간의 인지구조에 근접한 시스템이라 할 수 없는 것이다.

3) 생성의 결과로 제시되는 문장이 중의성을 지니든 그렇지 않든, 시스템의 평가차원에서는 중요하지 않다. 생성은 결국 자연어 또는 자연어에 준하는 단어열을 생산하는 과정이기 때문에 당연히 그 자체로 중의성을 지닌다. 이는 생성 관련 연구가 때로 연구자들에게 다분히 덜 정교한 방식인 것처럼 여겨지는 주요 원인이 되었다.

계산적으로 입증할 것이다. 본고에서 사용하고자 하는 문법은 Kim and Yang(2003) 이후 공개된 KRG(Korean Resource Grammar)이며, 논리 형태 표상은 MRS(Minimal Recursion Semantics, Copestake *et al.*(2005))를 사용한다. ‘분석’의 용도로 활용되는 시스템은 PET 분석기 (Callmeier, 2000)이며, ‘생성’의 용도로 활용되는 시스템은 LKB(Linguistic Knowledge Builder, Copestake(2002))이다.

2. 연구의 배경

2.1 생성 연구의 필요성

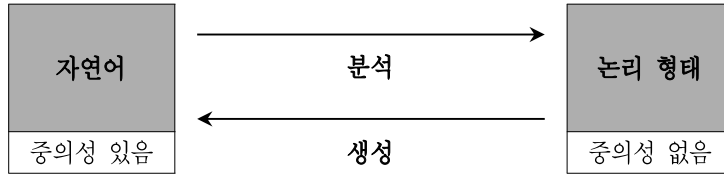
‘생성문법’이라는 단어에서 직접적으로 드러나는 바와 같이 문법을 기술하는 중요한 목적 가운데 하나가 주어진 규칙을 가지고 무수의 문장을 생성할 수 있는 것이다. 즉, 분석과 생성은 동전의 양면과도 같으며 한쪽만을 지원하는 문법은 이론적 차원 및 구현의 차원에서 그 한계를 드러낸다. 우선 이론적인 차원에서 보자면, 그 문법이 자신의 생성성을 완전하게 입증하지 못하였다는 한계를 지닌다. 어느 문법이 얼마나 많은 문장을 각 해당언어의 직관에 비추어 올바르게 만들어 낼 수 있는가는 문법의 생성 적합성을 판단하는 데 있어서 분석의 차원과 동일시되어야 한다. 다른 한편으로 전산적 구현의 측면에서 보자면, 생성을 지원하지 못하는 문법은 구문 분석기 또는 의미 해석기 등과 같은 자연어이해 단계에만 머무르게 되므로, 기계번역이나 대화형 시스템, 문서 자동 요약기 등과 같은 보다 실용적인 차원의 언어처리 시스템을 개발하는 데 활용될 수 없다. 다시 말하자면, 자연어처리 시스템의 문법은 분석은 물론이거니와 생성까지 올바르게 진행할 수 있어야 비로소 완결성을 갖추었다 할 수 있다.

2.2 분석과 생성의 차이점

분석과 생성은 전술한 바와 같이 문법의 주요한 두 축이지만, 각각에 대한 접근법 그리고 주안점은 상이하다. 본 소절에서는 생성이 분석과 어떠한 면에서 차이를 보이는가를 우선 정리하고자 한다. 생성은 대부분의 연구자들에게 익숙한 접근법인 분석과는 다른 부분이 있어 이 부분에 대한 언급이 선행되어야 이후의 논의를 이끌어 갈 수 있기 때문이다.

가장 명시적인 차이점은 입력력이 서로 반대라는 것이다. 분석은 자연어 즉, 중의성을 지니는 단어의 집합을 그 입력으로 받아 이를 해석하고 이해하여 중의성을 제거한 논리 형태로 변환하는 절차를 의미한다. 생성의 입력은 이와 반대로 그 자체로 중의성이 없는 논리 형태이다. 이 논리 형태를 각 매개 언어의 규칙에 입각하여 중의성을 지닐 수 있는 하나 이상의 자연어 문장으로 산출해 내는 과정이 생성이다⁴⁾.

4) 일반적 의미로 자연언어생성이라 하면, 준비된 문장 (canned text), 문틀 기반 생성 (template filtering) 까지도



[그림1] 분석 및 생성의 입출력

입출력의 대상이 다르기 때문에, 당연히 분석과 생성에서 주요하게 고려되어야 할 요소 또한 다르다. 분석에서는 중의성을 제거하고 모호한 입력으로부터 최대한의 정보를 도출하는 것이 중요한 반면에 (김영택 외, 2001:311), 생성에서는 내용, 어휘, 문장 및 담화 구조 등에 대한 선택이 중요한 화두가 된다 (Jurafsky and Martin, 2000:766). 이는 더 나아가 생성에서는 주어진 정보 자체뿐만 아니라 그 주어진 정보를 전달하는 방식까지도 고려되어야 함⁵⁾을 의미한다. 다시 말해, 문 자체의 내포적 의미뿐만 아니라 발화 상황에 대한 정보까지도 고려의 대상에 속한다. 예컨대, 발화의 상대가 누구인가에 따라 존칭법의 사용 유무가 결정되어야 하고, 또는 발화자와 청자의 관계가 어떠한가에 따라 사용되는 어휘나 태, 법 등과 같은 통사적 요소까지도 달리 선택될 수 있다.⁶⁾

끝으로 구현상에서 차이점을 밝히자면, 분석은 가능한 한 견고하게 이루어 져야 하지만, 생성은 가능한 한 엄격하게 이루어져야 한다 (Bond et al., 2008)⁷⁾. 이상의 비교점을 정리하자면 아래 표와 같다.

포괄한다 (Jurafsky and Martin, 2000:763). 또한, 최근의 통계기반 기계번역에서 쓰이는 방법론 역시 넓은 범위에서 보면 생성의 일환이라 할 수 있다. 그러나 이는 기계적 또는 통계적 조작에 의한 것일 뿐 인간의 문법 모형에 근간을 둔 것은 아니다. 더하여, 생성 가능한 문장의 집합 차수 역시 무한할 수 없다는 점에서 엄밀한 의미의 생성은 아니다. 논리 형태를 통로로 하여 이루어지는 분석과 생성의 유기적 과정이야 말로 촛스키 이후 현대 언어학의 관점에서 가장 이상적인 것이며 (Allen 1995:290), 무한한 인간의 사고과정을 풍부하고 역동적으로 표현할 수 있다는 점에서 우월한 방법이다 (Oepen et al., 2004).

- 5) 자연어 생성의 단계를 단순화 시켜 표현하자면, (1) “what to say”, (2) “how to say”, 그리고 (3) “say it”으로 구분할 수 있을 것이다 (Banik, 2009). “what to say”의 정보는 논리 형태에서 가져온다고 가정하지만, “how to say”의 경우에는 생성 시스템이 결정해 주어야만 한다. 본고에서의 “what to say”에 해당하는 논리 형태는 MRS이며, “how to say”는 뒤에서 언급될 STYLE 자질 구조가 담당할 것이다. 끝으로 “say it”은 LKB가 역할을 할 것이다.
- 6) 이러한 까닭에, 자연어 생성은 체계 문법을 오랜 기간 참조하여 왔다. 이 문법은 주제성, 정보성 등과 같은 상황 맥락에 대한 의존을 크게 강조하고 있기 때문이다 (Nagao, 1996).
- 7) “This means that robust rules (for example a rule to allow verbs not to agree in number with their subject) will apply in parsing but not in generation. The grammar will thus parse *The dog bark* or *The dog barks* but only generate *The dog barks*.” (Bond et al., 2008:152)

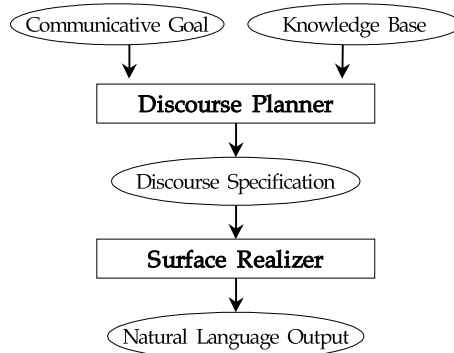
	분석	생성
입력	자연어	논리 형태
출력	논리 형태	자연어
고려사항	중의성 해소	선택
발화상황	거의 상관없음	매우 중요
방식	견고하게	엄격하게

[표1] 분석과 생성의 차이점

여기까지만 보면 분석과 생성은 전혀 별개의 과정으로 비취질 수 있을 것이다. 그러나 그러한 것은 아니다. 분석이 선행되지 않으면 생성의 입력값이 되는 논리 형태가 존재할 수 없기 때문에, 실제 시스템에서 생성은 결국 분석에 의존하여 기능한다. 더군다나 좋은 생성 출력이 가능하려면 우선 좋은 분석 문법이 전제되어야 한다. 분석 문법이 낙후한데도 생성이 원활하게 이루어지는 경우는 이론적으로 말해 존재할 수 없기 때문이다. 역으로 생성의 결과를 가지고 분석 문법의 오류를 검침하여 문법의 효율성을 제고시키는 방법론 또한 가능하다 (Goodman and Bond, 2009). 서론에서 밝힌 바와 같이, 생성은 이론적, 구현적 관점에서 문법의 포괄범위를 밝히는 척도가 될 수 있기 때문이다. 정리하자면, 분석과 생성은 상보적 관계라 하겠다.

3. 생성의 모형

Jurafsky and Martin (2000)에서는 자연어 생성의 주요 과정을 표층 실현부 (surface realizer)와 담화 기획부 (discourse planner)로 양분하고 있다.



[그림2] 자연어 생성 시스템 (Jurafsky and Martin, 2000:767)

담화 기획부는 앞 소절에서 설명한 ‘선택’의 역할을 담당한다. 주어진 발화상황에 알맞은

내용, 어휘, 표현 양태 등을 선택하여 맥락에 가장 근접한 대상을 만들어 낸다. 발화상황에 따른 고려 요소 역시 매개 언어마다 다를 수 있다. 예컨대, 한국어에서는 발화자-청자 관계에 따른 존칭법과 상황에 따른 격식체-비격식체의 선택 등이 있을 수 있다.

마찬가지 맥락에서 Nagao (1996)에서도 문생성을 결정하는 요인으로서 아래와 같은 총 10개 범주를 상정하고 있다. 문을 생성하기에 앞서 고려되어야할 요소가 의외로 간단치 않다는 점을 알 수 있다.

	기 호	
문형	A	(1)평서문, (2)비교문, (3)의문문, (4)감탄문, (5)조건문, (6)질문문
	B	(1)긍정문, (2)부정문
	C	(1)선언, (2)서술, (3)의뢰, (4)원망, (5)의무, (6)가능, (7)당연, (8)가정, (9)상상, (10)권유, (11)공동
	D	(1)현재, (2)과거, (3)미래, (4)무시제
	E	(1)default, (2)진행중, (3)반복, (4)완료, (5)개시, (6)상태 계속
	F	(1)절대시간
발화자의 태도	I	(1)보통, (2)정중, (3)난폭, (4)default
	J	화자자, (1)남, (2)여, (3)어린이, (4)default
	K	청자자, (1)남, (2)여, (3)어린이, (4)default
	L	청자에게, (1)존경, (2)명령, (3)비하, (4)대등, (5)default

[표2] 발화의 조건으로서의 인자 (Nagao, 1996:284)⁸⁾

다음으로 표층 실현부는 각 매개 언어의 어휘 및 통사 규칙에 입각하여 입력된 논리 형태의 진리조건을 전달할 수 있는 자연어 문장 집합을 구성한다. 동일한 진리조건을 담는 문장이 여러 쌍이 될 수 있기 때문에 결과로서 제시되는 문장 역시 복수가 될 수 있다. 이 복수의 문장에서 최종 결과로서 선택되는 항목을 추리는 것은 담화 기획부 단계에서 결정된 선택값에 근거한다. 다시 말해, 표층 실현부는 논리 형태로 ‘추상화’되어 있는 대상을 위 표2에서 제시된 발화인자를 참조하여 언어 표현으로 ‘구체화’시키는 과정을 담당한다.

이러한 표층 실현부의 구현을 위해 그간 사용되어 왔던 문법 모형은 크게 두 가지를 들 수 있다.⁹⁾ 첫째는 체계 문법 (systemic grammar)이다 (Winograd, 1983). 이는 문맥 정보를 중요시하는 문법이 담화 기획부에서 결정된 사항을 표층에서 능동적으로 실현하기에 유용하다는 데 기인한다. 그러나 동시에 분명한 한계를 지닌다. 발화외적 요소에 대한 고려가 없는 촘스키 문법의 난점을 극복한 것은 긍정적이지만, 반대로 문의 구성에 대한 세밀한 부분까지는 설명하지 못하는 단점이 존재하기 때문이다. Nagao (1996:265)에서는 이 양자를 균형 있

8) 이러한 맥락에서 김영택 외 (2001:313)에서는 문생성을 의사 결정 문제의 일종으로 보고 있다.

9) 여기에 언급된 것 이외에, ATN(Augmented Transition Network)도 자연어 생성에 적용된 바 있다 (김영택 외, 2001:323).

게 모색할 수 있는 모형의 필요성을 언급하고 있는데, 이러한 요구에 잘 부응하고 있는 대안이 두 번째인 이른바 통합기반 문법이다. 전통적으로는 FUG(Functional Unification Grammar)가 자연어 생성의 문법으로 이용되었으며, 그 후 자질 구조, 통합, 어휘중심 등의 특성을 공유하는 LFG, HPSG 등의 전산 문법이 그 역할을 계승하고 있다. 이는 언어 처리의 각 단계를 분할하는 것이 아니라, ‘음운-형태-통사-의미-화용’의 각 단계를 총괄적으로 파악하고 처리한다는 점에서 체계 문법과도 근본적인 특성을 공유한다 (Jurafsky and Martin 2000:776). 생성을 언어의 단계마다 선택을 반복하는 과정으로 단순화시켜 정의한다면, 언어의 처리를 총괄적으로 진행하는 방식은 효과적인 생성에 큰 장점이 있을 것이다.

이러한 일련의 처리 흐름은 본고가 제안하고자 하는 모형과도 일치한다.

3.1 제안하는 모형

앞의 내용을 요약하면, 생성의 절차는 담화 기획부와 표층 실현부로 양분된다. 담화 기획부에서는 문의 종류나 시제 등을 담는 문형 정보와 발화자의 태도에 대한 정보가 결정되어야 한다. 표층 실현부를 위한 가장 합리적인 대안은 통합기반의 전산 문법이라 할 수 있다.

3.1.1 표층 실현부 : HPSG, MRS, PET, LKB

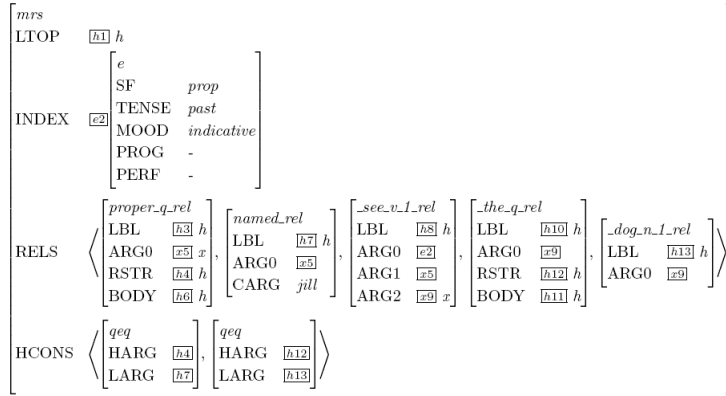
본고에서 이용하는 통합기반 문법은 HPSG이며 구체적으로는 KRG를 대상으로 한다. 핵어를 문법 관계의 중심축으로 삼는 HPSG는 그 전산적 구현을 거듭하여 왔으며, 실제로 문법을 구축하고 응용하기 위한 통합 시스템을 제공하고 있다. 이상의 환경은 HPSG를 한국어 생성 연구에 접목하기 위한 우수한 조건을 제공한다.

생성을 위한 근간 모형으로 HPSG를 사용하는 것은 두 가지 측면에서 장점을 지닌다. 우선, 앞서 설명한 바와 같이 기 연구 과정을 통해 생성에 적합한 문법 모형으로 어느 정도 검증받은 접근법이라는 점에서 안정적 연구의 초석이 마련된다. 둘째로, 분석용 문법과 생성용 문법을 별도로 구축하지 않아도 된다는 것이다. 동일한 문장에 분석용·생성용 문법을 따로 적용하는 것은 생성문법의 정신에도 맞지 않을 뿐 더러, 총괄적 처리를 지향하는 통합 기반 문법의 접근과도 사뭇 다른 것이다. 즉, 하나의 문법을 가지고 분석과 생성을 모두 할 수 있어야 하며, 이는 영어, 일본어, 독일어 등의 다른 언어 HPSG 연구를 통해 그 실현 가능성이 이미 입증되었다 (Oepen et al., 2007).

HPSG가 채택하고 있는 의미 표상 방법론인 MRS는 전산적 처리의 능동성을 염두에 두고 제시된 바, 분석과 생성의 가교 역할을 수행하는 데 있어 최적이다. 실제로 논리 형태를 입력력으로 한 생성 방법론에 대해서는 Allen (1995:286) 등에서도 제시한 바 있는데, 아래 (2)는 문장 (1)에 대한 논리 형태에 해당한다.

- (1) Jill saw the dog.
 (2) (NP **SEM** ?semsubj)
 (VP **SUBJ** ?semsubj)
SEM (<PAST SEES1> **s1** (NAME **j1** "Jill") <THE **d1** (DOG1 **d1**)>))

MRS가 표상하는 논리 형태¹⁰⁾는 (2)보다 세 가지 측면에서 발전된 것이다. 첫째로는 RELS 목록과 같이 평면적 구조를 택함으로써 필요한 연산량을 감소시켰다는 점이며, 두 번째는 handle이라는 기제를 통해 각 요소의 관계를 보다 긴밀하고도 직관적으로 연결시켜 놓았다는 점이다. 끝으로 HCONS 등을 통해 문장의 중의성을 능동적으로 파악할 수 있는 장점을 가진다 (Copestake *et al.*, 2005). 그림3은 (1)에 대한 MRS 구조이다.



[그림3] MRS 예시 (ERG: English Resource Grammar)

이러한 이론적 체계와는 별도로 기 구축된 HPSG 기반 한국어 분석문법이 존재한다는 점도 중요한 이점이 된다. 본고에서 사용된 문법은 KRG를 발전시켜 생성까지 가능하도록 재구성된 것이다.¹¹⁾ 더하여 자연어 문장에서 MRS 표상을 이끌어 낼 수 있는 PET 분석기와 MRS 구조를 입력으로 하여 다시 자연어 문장을 생성해 낼 수 있는 LKB 환경이 구축되어 있

10) MRS 의미 표상은 기본적으로 LTOP (local top), INDEX, RELS (relations), HCONS (handle constraints)로 구성된다. LTOP은 주어진 의미관계를 지시할 수 있는 지표인 handle을 담고, INDEX는 해당 의미 관계가 명사성에 가까운 개체 (individual)인지 아니면 동사성에 가까운 사건 (event)인지를 표시한다. RELS는 낱말의 술어관계를 모아 둔 목록이며, HCONS는 작용역 (scope)을 드러낸다. MRS 일반에 대한 보다 자세한 설명은 Copestake *et al.* (2005)를, 한국어 MRS에 관련하여서는 Kim (2006)을 참조하길 바란다.

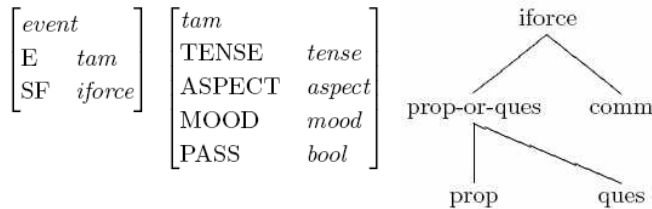
11) 분석을 주된 목적으로 개발된 문법을 생성까지 가능하도록 발전시키기 위해서는 몇 가지 추가적인 작업이 필요하다. 예컨대, 어휘 형성의 위계 구조를 단순화시키고, MRS 처리부를 효율성을 중심으로 재편하며, 문법의 전체적 구조를 표준화하는 등의 수정 과정이 요구된다. 이는 그 자체로 독립된 연구 주제이며, HPSG 기반 한국어 문법의 생성 가능성을 타진하고 그 실례를 입증하고자 하는 본고의 목적에서는 벗어나므로 이후 별도의 논문에서 제시하고자 한다.

으며, 이 모듈들이 무료로 공개되어 활용할 수 있다는 점도 본 연구의 큰 출발점이다.

3.1.2 담화 기획부 : MRS, STYLE

앞 표2를 기준으로 하였을 때, 발화상황에 관여한 인자는 두 가지 축으로 나누어진다. 하나는 문형이다. 예를 들어 시제가 과거인지 현재인지 또는 문의 양상이 어떠한지에 관련된 항목으로서, 이는 생성 처리의 입력 자료로 들어오는 MRS가 대부분을 포괄한다. 구체적으로는 event의 두 하위 범주인 TAM(Tense, Aspect, Mood) 자질 구조와 SF(Sentence Force)가 담당한다. 다른 하나는 발화와 관련된 외적 환경에 해당한다. 예컨대, 장르에 따른 문체의 구분을 들 수 있다. 이는 각 기호의 직접 하종 (subtype)으로 처리되는 STYLE 자질 구조가 담당을 한다.

먼저 문형 정보를 취득하여 이를 생성결과에 반영하는 MRS 항목에 대해 살펴보자. 대표적으로 위 (2)에서 시제 정보 'PAST'는 그림3에서 INDEX 아래 event 관계의 TENSE의 자질값으로 명세 되어 있다. MRS는 여기에 더하여 문장의 발화 강도를 나타내는 SF, 법, 진행상 및 완료상 등의 상적 속성까지도 망라한다. 한국어와 관련한 자질 및 위계 구조는 아래와 같다.



[그림4] 문형 정보 MRS

위 표2에서 언급된 내용과 비교하여 처리부를 구분하면 아래 표와 같다.12)

12) 종래의 KRG 구조에서는 이 항목들이 head 자질의 하종 (subtype)으로 상정되어 있었다. 그러나 이 자질들은 생성 단계에 있어 중요하게 참조되어야 할 사항이므로, MRS로 이전되는 것이 바람직하다. 이때, [PASS bool]과 같은 태 정보가 TAM 자질 구조의 하나로 놓일 것인가 아닌가에 대해 논란의 소지가 있을 수 있는데, Allen(1995:290) 및 Jacy (일본어 HPSG 문법)를 근거로 하여 TAM의 하나로 가정한다. 무엇보다, 수동태의 실현이 개별언어에 따라 상이한 까닭에, 태 정보 역시 선택의 문제로 환원되기 때문이다.

기호	정보	MRS
A	(1)평서문, (2)비교문, (3)의문문, (4)감탄문, (5)조건문, (6)질문문	SF
B	(1)긍정문, (2)부정문	
C	(1)선언, (2)서술, (3)의뢰, (4)원망, (5)의무, (6)가능, (7)당연, (8)가정, (9)상상, (10)권유, (11)공동	MOOD
D	(1)현재, (2)과거, (3)미래, (4)무시제	TENSE
E	(1)default, (2)진행중, (3)반복, (4)완료, (5)개시, (6)상태 계속	ASPECT
F	(1)절대시간	×

[표3] 발화인자 (Nagao, 1996:284)와 MRS

즉, MRS가 담화 차원의 결정 사항을 포함한 채로 표층을 구성하는 매개체 역할을 하는 것이다.¹³⁾ 결과적으로 논리 형태로 MRS를 이용하는 것은 문장을 구성하는 각 과정마다 MRS가 선택의 지침으로 활용되기 때문에 일관된 생성 패턴을 형성할 수 있다는 장점을 낳는다.

다음으로 STYLE 자질 구조의 역할에 대해서 살펴보자. STYLE 자질 구조는 발화외적 상황에 대한 통제를 하는 것으로 쉽게 말해 어떠한 표현의 문체적 특성에 관한 표지이다. 중의성을 지니는 자연어 문장이 복수의 논리 형태로 분석될 수 있는 것처럼, 각 논리 형태에 해당 하는 자연어 문장 역시 하나 이상이 될 가능성이 있다. 이러한 관계를 가장 잘 보여주는 예가 이른바 바뀌쓰기 연구이다. Bond *et al.* (2008)에서는 (3)과 같은 바뀌쓰기의 의미적, 문법적 가능성을 제시하였으며, (4a)의 문장이 영어 문생성을 통해 그 이하의 문장으로 바뀌쓰기 되는 과정을 제시한다.

- (3) a. 어순 : Kim sometimes goes $\equiv$ Kim goes sometimes
 b. 어휘 대체 : everyone $\equiv$ everybody
 c. 축약 : going to $\equiv$ gonna
 d. 수정 : the the $\rightarrow$ the
- (4) a. Everybody often goes to the the movies.
 b. Everyone often goes to the movies.
 c. Everybody often goes to the movies.
 d. Everyone goes often to the movies.
 e. Everybody goes often to the movies.
 f. Everyone goes to the movies often.
 g. Everybody goes to the movies often.

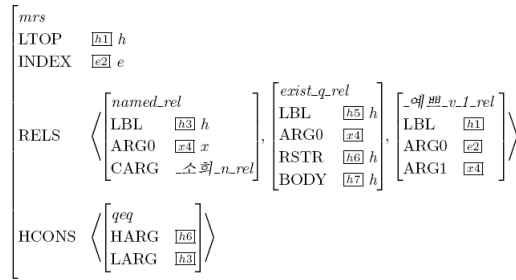
13) 이는 체계 문법이 주요 생성 모형으로 활용된 이유와 일맥상통한다.

(3d) 및 (4a)의 ‘the the’는 완전히 문법적인 표현은 아니다. 그러나 이는 앞에서 설명한 ‘분석은 견고하게’, ‘생성은 엄격하게’라는 원칙에 따라 분석 가능해야 한다. 즉, 자연어 문장에서 약간의 오류가 있더라도 분석 시스템은 이를 유연하게 처리할 수 있어야 한다. 반면, 생성은 최대한 사용자에게 가장 정확한 문장을 전달하는 것을 목표로 해야 한다. 따라서, ‘the the’와 같은 다소 비문법적인 입력이라 하더라도 이를 유연하게 분석하여 ‘the’로 수정한 후 이를 최종 생성 결과에 반영한다. (4b-g)는 생성된 결과인데, ‘the the’와 같은 유형이 출현하지 않는 것에서 확인할 수 있다. 이때, ‘the the’와 같은 표현이 출현하는 (4a)의 STYLE은 ‘robust’에 해당한다. (3a-c)의 경우에도 이러한 구분이 있을 수 있다. (3a)에서 ‘sometimes’는 일반동사의 앞에 오는 것이 일반적이므로 후자가 전자보다 STYLE 면에서는 보다 견고하게 처리되어야 한다. (3c)의 축약도 마찬가지로 후자가 전자보다 더 특화된 STYLE 특성을 지닌다. 끝으로 (3b)의 어휘 대체의 경우에 위에서는 그 정도의 차이가 거의 없으나, 동일한 의미를 지시하는 명사라 하더라도 STYLE 상의 차이가 존재할 수 있다.

- (5) a. 예쁘다 ≡ 이쁘다
- b. 양동이 ≡ 바께쓰
- c. 부추 ≡ 정구지

(5a)의 경우에는 맞춤법상의 오류에 속하며, (5b)는 외래어 남용에 의한 것으로 이들은 흔히 말하는 언어 순화의 대상이다. (5c)의 경우에는 나쁜 말은 아니나 지역 방언에 해당한다. 사용자의 실제 언어생활을 있는 그대로 중요시 하는 자연어 처리 분야는 규범 문법과 그 관점이 같을 수 없다. 사용자가 ‘바께쓰’와 같은 단어를 입력했다 하더라도 잘못된 표현이라는 이유로 분석 결과를 제시하지 않을 수는 없기 때문이다. 다시 말해, (5)와 같은 단어 쌍의 관계는 옳고 그름보다는 STYLE상의 차이로 받아들이는 것이 타당하다.

이러한 개별 표현상의 차이에 의한 STYLE 구분 외에 장르 특성에 따른 차이도 존재한다 (강범모, 1999). 문어체와 구어체의 구분이 비교적 뚜렷한 한국어에서는 같은 의미를 전달한다고 하더라도 문어체인지, 구어체인지, 또는 방송대본과 같은 준구어체인지에 따라 그 표현이 달라진다. 높임법이 매우 발달한 한국어에서는 사용하는 문체가 격식체인지 또는 그렇지 않은지에 따른 차이도 존재한다. 한편, 텍스트의 유형, 예컨대 소설, 신문, 인터넷 글 등과 같은 구분도 해당 언어의 표현적 특성에 적지 않은 영향을 준다. (6)은 모두 MRS 논리 형태 그림5에 해당하는 언어 표현들이다.



[그림5] MRS : 소희가 예쁘다

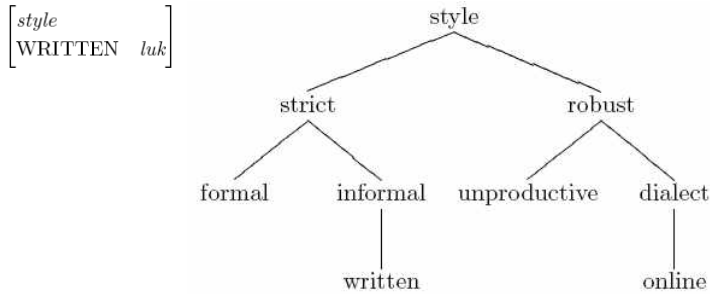
- (6) a. 소희가 예쁩니다 (격식체)
- b. 소희가 예쁘다 (비격식체)
- c. 소희가 이쁘다 (맞춤법 오류)
- d. 소희가 예쁘네 (구어체)
- e. 소희가 예뻐 (인터넷 글)

‘분석’의 관점에서만 보면 통사적 판별에 큰 영향을 주지 않는다는 점에서 STYLE 자질은 그다지 유의미한 정보라고 할 수 없다. 그러나 ‘생성’의 관점에서 STYLE 구분은 매우 중요한 요소가 된다.¹⁴⁾ 앞서 설명한 담화 기획부의 ‘선택’의 문제가 되기 때문이다. 즉, 그림5의 논리 형태를 (6)의 어떠한 문장으로 생성해 낼 것인가는 철저하게 생성 과정의 선택에 따른 것이다.

STYLE 자질 구조는 종래의 KRG에서도 일부 처리하고 있었다. 종래 구조의 문제점은 두 가지인데, 하나는 격식체 (formality) 또는 비격식체(informality)로의 구분만이 가능하다는 점이다. 이는 위에서 살핀 바와 같이 실제 문의 생성에 관련되는 STYLE의 다양한 층위를 다 포괄하지 못하는 문제점이 있다. 두 번째는 STYLE이 핵 자질 구조의 하종 (subtype)으로 상정되어 있었다는 점이다. 이는 STYLE이 어휘에서부터 담화에 이르기까지 언어의 다양한 층위에 의해 영향을 받는다는 점을 반영할 수 없으며, 선택의 기준이 되는 담화 기획의 요소가 아니라 표층 실현의 인자로 상정된다는 점에서도 수정되어야 한다.

본고에서는 이 자질의 구조를 생성에 효율적인 형태로 개선하고 그 위치를 이동시킨다. 우선 위치와 관련하여서는 HPSG 문법의 근간이 되는 기호의 직접 하종 (subtype)으로 두어, 모든 단계에서 접근이 가능하도록 한다.

14) 이는 굳이 HPSG 기반 문생성 뿐만 아니라 넓은 범위의 생성 분야 연구 전반에 해당하는 것이다. 대표적으로 통계기반 기계번역의 경우 기반이 되는 언어 자원의 장르 특성에 크게 영향을 받는 것으로 알려져 있다.

[그림6] STYLE 자질 및 위계 구조¹⁵⁾

다음으로 구조는 위 그림6과 같은 자질 구조 및 위계를 형성한다. 먼저 구어·문어의 구분을 위하여 각 STYLE 구조는 [WRITTEN *luk*]를 명세한다. 여기서 '*luk*'는 이른바 '*boolean*' 자질과 유사한데, *boolean*이 +, -, 그리고 미명세의 ±를 사용하는 데 비해, 삼원구조를 보다 명시적으로 사용할 수 있다는 장점이 있다. 이는 문어, 구어, 준구어의 삼원 구별이 존재하는 한국어 문체 구분에 보다 유용하다 (강범모, 1999). 위계 구조에서 각 절점에 대한 설명을 하면 아래와 같다.

- (7) a. strict : 일반적인 학교 문법에 맞는 표현
- b. formal : 격식체 (소회가 예뻐니다)
- c. informal : 비격식체 (소회가 예쁘다)
- d. written : 신문, 논문 등의 공식적 유형의 텍스트에서 출현하는 표현
- e. robust : 학교 문법에는 완전히 맞지 않으나 이해 가능한 표현
- f. unproductive : 비생산적이나 가능한 표현 (소회 예쁘다)
- g. dialect : 방언 등의 비규범적 표현 (소회가 예쁘데이)
- h. online : 인터넷 등의 이른바 통신체 표현 (소회가 예뻐)

위에서 written은 가장 엄격한 문체에 속한다. 본고에서 생성하고자 하는 문의 유형 또한 이 단계를 기본으로 한다. unproductive는 대표적으로 격조사가 생략된 경우를 들 수 있다. 실제 한국어 화행에서는 격조사 생략은 매우 빈번하게 나타나지만, 신문기사 등과 같은 문체적 제약이 있는 텍스트를 생성할 때에는 가급적 모든 문법 관계를 명시적으로 제시하는 것이 바람직하다. 다음으로 dialect는 표준어법이 아닌 경우를 말하는 데, 실제로 소설류의 텍스트에서는 이러한 형태가 빈번히 출현한다. 소설류에 해당하는 영어 또는 일본어 문장을 한국어로 변환하여 생성할 때에는 이 유형을 선택하는 것이 자연스러운 기계번역을 위해 필요할 것이다. 끝으로 online은 대표적으로 이모티콘을 예로 들 수 있다. 이는 최근 주목을 받고 있는

15) 이는 주로 ERG의 형태를 참조한 것이다.

분야로서, 전자우편, 블로그, 위키피디아와 같은 인터넷 게시물의 경우 그 장르의 특성에 맞는 분석과 생성이 요구된다.

3.2 요약

이상에서 논의된 분석과 생성, 그리고 문법의 관계를 도식화하면 아래 그림7과 같다.

분석	형태소 분석 (어휘)			
	구문 분석 (통사)			
	의미 해석 (의미-화용)			
KRG	<u>HPSG</u>	<u>MRS</u>		<u>STYLE</u>
		LTOP, RELS, HCONS	SF, TAM	
생성	<u>표층 실현</u>		<u>담화 기획</u>	
	어휘 선택		문형	발화양식
	형태 합성			
	통사 구성			

[그림7] 분석, 생성 및 문법의 연계

이어지는 각 장에서는 위의 제안이 실제로 어떻게 구현되는지를 실례를 통해 살펴볼 것이다. 4장에서는 ‘개가 짖다’라는 문장을 분석의 입력열로 하여 이것이 논리 형태로 변환되고 생성되는 과정을 따라가 볼 것이다. 5장에서는 ‘소희가 예쁘다’라는 문장의 분석을 기반으로 생성 결과의 선택에 대해 살펴볼 것이다.

4. MRS를 통한 표층 문의 생성

분석의 대상으로 삼는 한국어 문장은 (8)이다.

(8) 개가 짖다

문의 길이가 짧을지라도 전형적인 자연어 문장이기 때문에, (8)은 몇 가지 중의적 의미로 해석될 수 있다. 우선 첫 번째로는 명사구 ‘개’의 해석에 관한 것이다. 일반적으로 알려진 바와 같이 한국어에서 복수 표지 ‘-들’의 삽입은 선택적이다. 따라서 ‘개’는 두 가지 해석이 가능하다. 하나는 단수의 의미로 ‘개 한 마리’를 지칭한다. 반면 이것이 표지 ‘-들’이 유표적으로 붙지 않은 복수 의미라고 가정하면 ‘무표 복수’에 해당한다 (강범모, 2007). 두 번째로는 술어

‘짓다’의 층위 문제이다. ‘짓다’의 경우 보통 장면 층위 술어(stage-level predicate)로 판별되지만, 전영철 (1997)에서 지적된 바와 같이 한국어에서 장면 층위 술어와 개체 층위 술어(individual-level predicate)를 구분하는 일은 단순하지 않다. 실제로 전영철 (1997)에서는 위 (8)의 변형인 아래 (9)가 특정문과 총칭문으로 모두 해석 가능하다는 입장을 내놓고 있다.¹⁶⁾

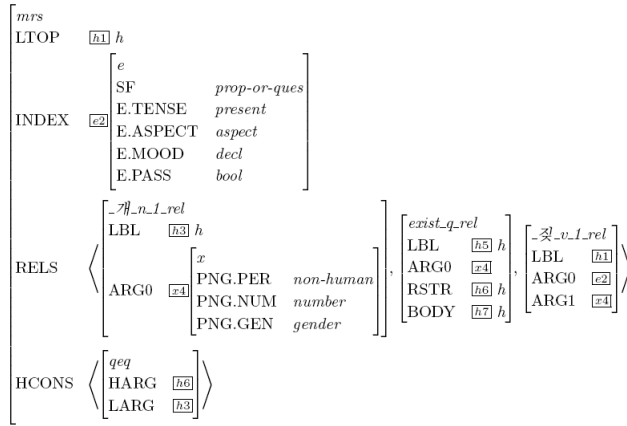
(9) 개가 짓는다

(8)과 (9)를 구분하자면, (8)은 다소 중립적인 문으로서 특정문으로의 해석이 지배적이다. 반면 (9)은 현재형 표시 ‘-는’이 삽입되어 상태성이 보다 강조되며 총칭문으로의 해석이 강화된다. 즉, (9)는 (8)에서 도출될 수 있는 바꿔쓰기 문장의 하나인 것이다. 이상을 정리하면, (8)은 아래와 같이 바꿔쓰기 될 수 있어야 할 것이다.

- (10) a. 개가 짓다
b. 개들이 짓다
c. 개가 짓는다
d. 개들이 짓는다

아래에 제시된 논리 형태를 살펴보자. 첫 번째 중의적 요소, 즉 ‘개’의 복수 해석과 관련하여서는 ‘_개_n_1_rel’의 ARG0 항목의 PNG(Person, Number, Gender) 항목 가운데 PNG.NUM이 관여한다. 여기에서는 [PNG.NUM *number*]로 미명세된 값이다. 즉, ‘개’의 단수 또는 복수 여부가 결정되지 않은 것이며, 이는 표층에서 단수나 복수 모두로 해석될 수 있다. 두 번째 중의적 요소인 ‘짓다’의 해석에 관해서는 INDEX의 각 자질값이 관여한다. [E.TENSE *present*] 및 [E.MOOD *decl*]가 주어져 있고, 상태성이나 총칭성에 대한 추가적인 정보는 제시되지 않았기 때문에 가능한 형태는 ‘짓다’와 ‘짓는다’ 두 가지로 결정된다. 논리적으로 따졌을 때, 결국 가능한 생성 후보군은 2×2로 4개가 될 것이다.

16) (8) 또는 (9)의 문 해석과 관련하여서는 여러 논란이 있을 수 있을 것으로 본다. 한국어 명사구와 술어, 그리고 문의 유형을 판별하는 것은 매우 중요한 연구 과제임은 분명하나, 이는 엄밀한 의미에서 형식의미론의 과제이며, 논리 형태에서 자연어 생성을 도출하는 방법론에 대한 본고의 주된 연구 대상은 아니다. 본고에서는 우선 (8)의 문장이 복수성, 술어의 속성에 따라 중의적 해석을 가진다는 가정 하에 논의를 계속하기로 한다.



[그림8] MRS : 개가 짖다

주지한 바와 같이 HPSC 기반 문생성은 위 MRS 구조를 입력의 기본 형태로 삼는다. 일반적인 형태의 분석 및 생성의 형태는 아래 그림9의 흐름을 기본으로 한다 (Bond *et al.*, 2005).



[그림9] 분석-생성 흐름도

여기서 ‘Source Text’는 (8)의 문장이 되며, ‘Source Analysis’는 PET 분석기를 통해 이루어진다. 이 결과 만들어 진 ‘MRS_s’는 위 그림8과 같다. 다음 과정은 연구의 대상에 따라 달라지는데, 생성을 활용하는 주요한 방법으로 기계번역과 바꿔쓰기를 예로 들 수 있다. 즉, ‘Semantic Transfer’의 과정을 영어나 일본어와 같은 타 언어를 대상으로 하여 해당 언어의 ‘MRS_r’로 변환한다면, 이는 기계번역의 과정이다. 이 때 ‘Target Generation’은 LKB 상에서 ERG 혹은 Jacy 문법을 통해 표층 실현이 될 것이며, 그 결과로 제시되는 ‘Target Text’는 아래 (11) 또는 (12)가 될 것이다.

- (11) a. The dog barks.
 b. The dogs bark.
 c. A dog barks.
 d. Dogs bark.
- (12) a. 犬が吠える。
 b. 犬がほえる。

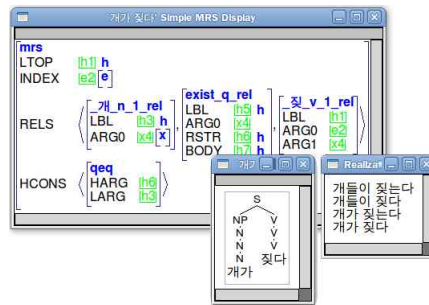
- c. 이ぬが吠える。
d. 이ぬがほえる。

반면에 ‘MRS_S’를 그대로 ‘MRS_T’로 활용한다면 이는 바뀌쓰기에 해당한다. 이 경우에는 ‘Target Generation’ 역시 KRG가 된다. 그 최종 생성 결과는 (10)과 같아야 할 것이다.

한영 또는 한일 기계번역의 과제는 독립적인 연구 주제이기 때문에 여기서는 우선 간단한 바뀌쓰기 과정을 보도록 한다. 아래 그림10은 PET 분석기를 이용하여 도출된 MRS이며¹⁷⁾, 그림11은 같은 문장을 LKB¹⁸⁾에서 분석 및 생성 처리한 예시이다.¹⁹⁾

```
[ LTOP: h1
INDEX: e2 [ e E.ASPECT: ASPECT E.MOOD: DECL E.PASS: BOOL E.TENSE: PRESENT SF: PROP-OR-QUES ]
RELS: <
  [ "_개_n_1_rel"
    LBL: h3
    ARG0: x4 [ x PNG.NUM: NUMBER PNG.GEN: GENDER PNG.PER: NON-HUMAN ] ]
  [ "exist_q_rel"
    LBL: h5
    ARG0: x4
    RSTR: h6
    BODY: h7 ]
  [ "_짓_v_1_rel"
    LBL: h1
    ARG0: e2
    ARG1: x4 ] >
HCONS: < h6 qeq h3 > ]
```

[그림10] PET 분석 결과 : MRSS & MRST



[그림11] LKB 분석 및 생성 결과 : 개가 짖다

17) PET 분석기는 리눅스 환경에서 작동하며, C언어로 구현되어 있어 실행 속도가 빠르다. 문법 파일을 만드는 모듈은 ‘flop’이며, 분석을 하는 모듈은 ‘cheap’이다. PET 구동을 위한 환경 설정 파일이 ‘pet.tdl’에 명세 되어 있다고 가정할 때, 일반적으로 그 실행은 리눅스 커널에서 다음과 같은 명령행으로 이루어진다.

- 바이너리 문법 파일 작성 : flop pet.tdl (‘pet.grm’이라는 바이너리 파일이 생성됨)
- MRS 문 분석 : cheap -mrs pet.grm < 입력파일경로 &> 출력파일경로

18) PET와 LKB를 연동하는 방법에는 여러 가지가 있을 수 있으나, 여기에서는 별도로 작성된 리눅스 셸 스크립트를 활용하였다. 이 셸 스크립트의 주된 역할은 분석과정은 PET의 ‘cheap’ 모듈을 이용하여 처리하고, 그 결과를 LKB 상에 로드된 한국어 문법 KRG로 전달하여 생성을 명령하는 것이다.

19) LKB는 분석의 용도로도 사용될 수 있기 때문에, 이 과정 역시 PET를 사용하지 않고 LKB만으로도 사용 가능하다. 단, PET가 LKB에 비해 처리 속도가 훨씬 빠른 데다, 분석과 생성을 분리하여 처리하는 것이 가능함을 보이기 위해 두 과정을 나누어 제시하였다.

5. STYLE 자질 구조와 선택의 문제

위 (10) 또는 그림11에서 제시된 문장은 가장 엄격한 형태인 ‘비격식체+ 문어체’를 취한 결과로서, 이는 신문기사 등에 적용되는 문체이다. 실제 자연언어의 표현방식은 위보다 훨씬 다양하기 때문에, 아래 (13)과 같은 문장 역시 생성 가능하다.

- (13) a. 개 짓다
b. 개가 짓네
c. 개가 짓습니다
d. 개가 짓어요

자연어 생성은 주지한 바와 같이 ‘선택’의 문제이다. 따라서 자연어 생성 시스템은 위 (8)의 입력을 가지고도 (13)의 문장을 출력할 수 있어야 하며, 거꾸로 (13)의 문장을 가지고도 기본형인 (8)의 문장을 만들 수 있어야 한다. 이는 표층 실현과는 별도로 STYLE 상의 선택 값에 따라 결정되는 것이다.²⁰⁾ STYLE 자질값의 차이를 야기하는 표현상의 요소는 크게 ‘어휘적 선택’, ‘형태적 굴절’, ‘통사적 구성’ 등의 세 가지로 정리할 수 있다. 이어지는 소절에서는 각각의 처리가 어떻게 되는지를 구체적인 예를 통해 살펴본다.

5.1 어휘적 선택

첫 번째로는 어휘상의 차이에 따른 것으로, 앞서 살핀 (5)의 내용과 같다.

- (14) a. 소희가 이쁘다
b. 소희가 예쁘다

‘분석은 견고하게’ 그리고 ‘생성은 엄격하게’라는 원칙을 지키려면, (14a)의 문장은 분석이 가능하여야 한다.²¹⁾ 반면, 생성의 단계에서는 (14b)만 선택되는 것이 바람직하다. 이 경우, ‘이쁘.’는 ‘예쁘.’의 이형태의 한 가지이므로 술어 이름은 그림12에 나타난 것처럼 ‘예쁘.’와 동일하여야 한다.

20) 통상, 이 선택은 HPSG 문법 구조에서 시작 노드를 결정하는 ‘roots.tdl’ 파일과 사용자 환경 변수를 저장하는 ‘user-pref.lsp’ 파일에 명시된다.

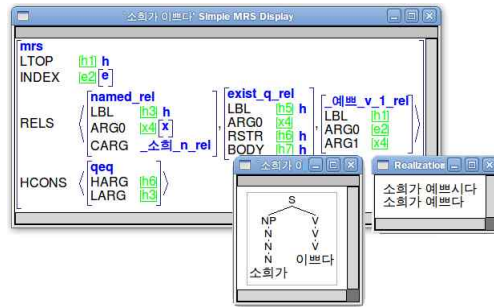
21) 보통 ‘이쁘.’는 ‘예쁘.’의 잘못된 형태라고 하나, 실제 세종 형태분석 코퍼스 1250만 어절 규모에서 찾아보면 ‘이쁘./VA’는 총 84회 출현하여 ‘VA’표지를 단 전체 형용사 어휘 1630개 전체 가운데 288위를 차지한다. 즉, 비규범적이기는 하나 실제 언어가 빈번하게 사용하는 형태에 해당한다.

```

이쁘 := a-intr &
[ STEM < "이쁘" >,
  STYLE dialect,
  SYNSEM.LKEYS.KEYREL.PRED "_예쁘_v_1_rel" ].
    
```

[그림12] ‘이쁘.’의 TDL 기술

결과적으로 그림12와 같은 구조는 앞에서 제시된 그림5와 동일한 논리 형태로 귀결된다. 즉, 분석의 과정에서는 아무 문제가 없으며, ‘예쁘.’와 동일한 어휘처럼 취급되는 것이다. 단, [STYLE *dialect*]로 명세되어 있다는 점에서 생성의 과정에서는 차이가 있다. 이 STYLE 자질을 담화 기획 단계에서 선택의 대상에 포함시키지 않는 이상, 이 형태는 생성결과에서 제외된다. STYLE 자질값으로 *written*을 선택한 결과는 그림13에 제시되어 있다.



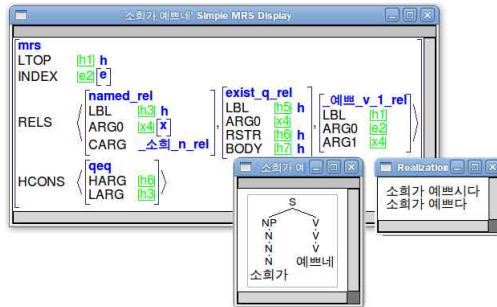
[그림13] LKB 분석 및 생성 결과 : 소희가 이쁘다

5.2 형태적 굴절

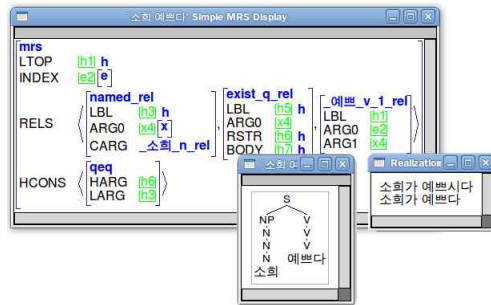
명사구 또는 동사구의 굴절에 관계되는 조사 혹은 (선)어말어미가 어떠하냐에 따라 STYLE 자질 역시 결정되는 바가 크다.

- (15) a. 소희가 예쁘네
 b. 소희 예쁘다
 c. 소희가 예쁘다

(15a)는 (15c)의 구어체에 해당하여 [STYLE *informal* [WRITTEN -]]을 STYLE 값으로 가지게 된다. 또한 (15b)는 (15c)에서 주격조사 ‘-가’가 생략된 형태로 이는 [STYLE *robust*]에 속한다. 이들 역시 견고하게 분석이 되어 그림5와 동일한 논리 형태를 만들어 낼 수 있어야 한다. PET 분석기가 이 부분을 담당한다. 다음으로 형성된 논리 형태에 따라 [STYLE *written*]으로 생성한 결과는 위 그림13과 동일하다.



[그림14] LKB 분석 및 생성 결과 : 소희가 예쁘네



[그림15] LKB 분석 및 생성 결과 : 소희 예쁘다

중요한 점은 위 선택을 거꾸로 할 수 있다는 점이다. 분석을 (15c)의 기본 문형을 가지고 하더라도 논리 형태는 동일하기 때문에, 생성 시 STYLE 자질 값을 [STYLE *informal*] [WRITTEN -] 또는 [STYLE *robust*]로 설정하면 거꾸로 (15a-b)의 문장을 결과로 얻을 수 있다.

5.3 통사적 구성

끝으로 통사 구성상의 특징 역시 STYLE의 차이를 낳는다. 아래 (16a)는 보조용언 구성에서 본용언에 격조가 '-가'가 붙은 경우이다.²²⁾ 이 경우에 (16b)가 보다 일반적이고 자연스러운 형태임은 분명하다.

- (16) a. 소희가 예쁘고가 싶다
b. 소희가 예쁘고 싶다

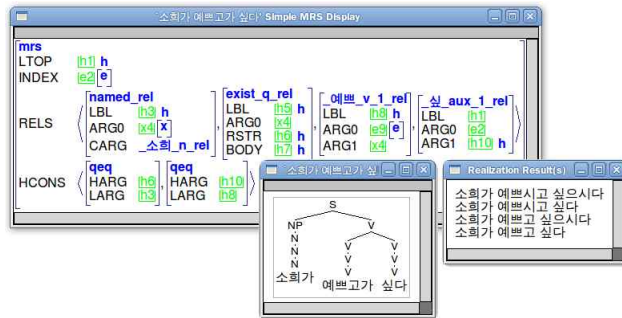
22) 해당 문법에 대한 구체적인 설명은 김종복 (2004)를 참조하길 바란다.

아래는 (16a)에 대한 PET의 분석 결과로서, (16a-b)는 진리조건이 같은 것으로 가정되어 동일한 MRS를 가진다. 다만, (16a)는 [STYLE *unproductive*]로서 (16b)와 STYLE 상의 차이만을 가진다.

```
[ LTOP: h1
  INDEX: e2 [ e SF: PROP-OR-QUES E.PASS: - E.MOOD: DECL E.ASPECT: ASPECT E.TENSE: PRESENT ]
  RELS: <
    [ "named_rel"
      LBL: h3
      ARG0: x4 [ x PNG.PER: 3RD PNG.NUM: NON-PL PNG.GEN: GENDER ]
      CARG: "_소_적_n_rel" ]
    [ "exist_q_rel"
      LBL: h5
      ARG0: x4
      RSTR: h6
      BODY: h7 ]
    [ "_예쁘_v_1_rel"
      LBL: h8
      ARG0: e9 [ e E.ASPECT: ASPECT E.MOOD: NO_MOOD E.PASS: BOOL E.TENSE: PRESENT SF: IFORCE ]
      ARG1: x4 ]
    [ "_싶_aux_1_rel"
      LBL: h1
      ARG0: e2
      ARG1: h10 ] >
  HCONS: < h6 qeq h3 h10 qeq h8 > ]
```

[그림16] PET 분석 결과 : 소희가 예쁘고가 싶다

따라서 위 구조를 가장 엄격한 선택인 [STYLE *written*]을 기준으로 LKB변환 및 생성을 거치면 그림17과 같은 결과를 얻게 된다.



[그림17] LKB 분석 및 생성 결과 : 소희가 예쁘고가 싶다

6. 결론 및 향후 과제

이상의 논의에서 PET 분석기를 통해 한국어 문장을 MRS로 표상하고 그 결과를 LKB에서 입력으로 받아, 다시 한국어 문장을 생성하는 과정을 제시하였다. 생성에는 두 가지 측면, 즉 표층 실현부와 담화 기획부가 작용을 한다고 가정을 하면, 본고에서 제시한 모형에서는 표층 실현부는 HPSG 문법 체계와 MRS 일부가 담당을 하고 담화 기획부는 MRS의 나머지 부

분과 STYLE 자질 구조가 그 역할을 한다. MRS에서 표층 실현과 관련되는 요소는 LTOP, RELS, HCONS 등이며, 담화 기획 층위와 관련되는 것은 SF와 TAM 등이다. 본고에서는 한국어의 다양한 장르별, 문제별 특성을 반영하기 위해 STYLE 자질 구조를 위계화 하였으며 실제 시스템에서는 그에 따라 생성 결과를 선택할 수 있도록 하였다.

이후의 추가적인 연구로서는 크게 세 가지를 들 수 있겠다. 첫째는 어휘 선택을 보다 정교화 하는 일이다 (Jurafsky and Martin 2000:790). 예컨대, 영어에서 ‘rather’와 ‘fairly’는 모두 ‘꽤’로 번역이 가능하지만, 전자는 주로 부정문에 후자는 주로 긍정문에서 출현한다. 이러한 이른바 의미 운율을 생성 과정에 반영하여 기계번역 등에서 보다 매끄러운 문생성을 유도하기 위한 장치가 요구된다. 둘째는 생성 결과를 바탕으로 분석 문법의 논리적 오류를 발견하는 연구이다. 생성 결과에서 언어 직관에 모순되는 문장이 발견된다면, 이는 분석 문법 차원에서 기인하였을 가능성이 매우 크다. 이와 관련하여서는 최근 Goodman and Bond (2009) 등의 연구가 제시되고 있으며, 생성 분야로의 확대를 막 시작한 KRG의 지속적 발전을 위해서 반드시 거쳐야할 순서라 하겠다. 끝으로 한-영, 한-일 기계번역을 위한 토대 연구이다. 본고는 주로 한국어 문장을 다시 한국어 문장으로 바꾸어 쓰는 방법을 제시하였으며 이것 역시 나름의 의의를 지닌 사실이나 (Bond *et al.*, 2008), HPSG 기반 문생성의 궁극적 목적은 다국어 기계번역이라 하여도 과언이 아니다 (Oepen *et al.*, 2007). 이를 위해서는 한국어 MRS를 타 언어 MRS로 변환하는 의미 변환 규칙에 대한 체계적인 연구가 있어야 할 것이다.

참고문헌

- 강범모. 1999. 『한국어의 텍스트 장르와 언어 특성』. 서울: 고려대학교 출판부.
- 강범모. 2007. “복수성과 복수 표지: “들”을 중심으로,” 『언어학』 47, pp.3-31.
- 김영택 외. 2001. 『자연언어처리』. 서울: 생능출판사.
- 김종복. 2004. 『한국어 구구조 문법』. 서울: 한국문화사.
- 전영철. 1997. “한국어 총칭문의 유형,” 『언어학』 21, pp.289-303.
- Allen, James. 1995. Natural Language Understanding. Redwood, CA: The Benjamin/Cummings Publishing Company, Inc.
- Banik, Eva. 2009. "Parenthetical Constructions - an Argument against Modularity," *Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks, ACL-IJCNLP 2009*, pp.46-53.
- Bond, Francis, Stephan Oepen, Melanie Siegel, Ann Copestake, and Dan Flickinger. 2005. "Open Source r.chine Translation with DELPH-IN," *Proceedings of Open-Source r.chine Translation: Workshop at MT Summit X*, pp.15-22.
- Bond, Francis, Eric Nichols, Darren S. Appling, and Michael Paul. 2008. "Improving

- Statistical Machine Translation by Paraphrasing the Training Data," *Proceedings of the 5th International Workshop on Spoken Language Translation*, pp.150-157.
- Callmeier, Ulrich. 2000. "PET - a Platform for Experimentation with Efficient HPSG Processing Techniques," *Natural Language Engineering* 6, pp.99-108.
- Copestake, Ann. 2002. Implementing Typed Feature Structure Grammars. Stanford: CSLI Publications.
- Copestake, Ann, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. "Minimal Recursion Semantics: An Introduction," *Research on Language and Computation* 3, pp.281-332.
- Goodman, Michael W., Francis Bond. 2009. "Using Generation for Grammar Analysis and Error Detection," *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp.109-112.
- Jurafsky, Daniel and James H. Martin. 2000. Speech and Language Processing. Upper Saddle River, NJ: Prentice Hall.
- Kim, Jong-Bok, and Jaehyung Yang. 2003. "Korean Phrase Structure Grammar and Its Implementations into the LKB System," *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation*, pp.88-97.
- Kim, Jong-Bok. 2006. "Minimal Recursion Semantics: An Application into Korean," *The Linguistic Association of Korea Journal* 14, pp.59-85.
- Oepen, Stephan, Dan Flickinger, Kristina Toutanova, and Christopher D. Manning. 2004. "LinGO Redwoods - A Rich and Dynamic Treebank for HPSG," *Research on Language and Computation* 2, pp.575-596.
- Oepen, Stephan, Erik Velldal, Jan T. Lønning, Paul Meurer, and Victoria Rosén, and Dan Flickinger. 2007. "Towards Hybrid Quality-Oriented Machine Translation - On Linguistics and Probabilities in MT -," *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp.144-153.
- Nagao, Makoto. 1996. 自然言語處理. 日本: 岩波書店 (황도삼 외 옮김, 1998, 『자연언어처리』 서울: 홍릉과학출판사).
- Winograd, Terry. 1983. Language as a Cognitive Process: Volume 1 Syntax. MA: Addison-Wesley.

Sanghoun Song

1008 B North 45th Street

Seattle, WA, USA

98108

E-mail: sanghounsong@gmail.com

접수일자: 2009. 09. 28

수정일자: 2009. 12. 12

게재일자: 2009. 12. 24