통합기반 다국어 자동번역 시스템에서의 한국어 분석과 변환

최 승권　　박 동인
시스템공학연구소 자연어정보처리연구부

Korean Analysis and Transfer in Unification-based Multilingual Machine Translation System

Sung-Kwon Choi, Dong-In Park
Natural Language Information Processing Department, Systems Engineering Research Institute

요　　약

다국어 자동번역이란 2 개국어 이상 언어들간의 번역을 말한다. 기존의 다국어 자동번역 시스템은 크게 변환기반 transfer-based 방식과 피봇방식으로 분류될 수 있는데 변환기반 다국어 자동번역 시스템에서는 각 언어의 분석과 생성 규칙이 상이하게 작성됨으로써 언어들간의 공통성이 수용되지 못하였고 그로 인해 전체 번역 메모리의 크기가 증가하는 결과를 초래하였었다. 또한 기존의 피봇방식에서는 다국어에 적용될 수 있는 언어학적 보편성 모델을 구현하는 어려움이 있었다. 이러한 기존의 다국어 자동번역 시스템의 단점들을 극복하기 위해 본 논문에서는 언어들간의 공통성을 수용하며 또한 여러 언어에서 공유될 수 있는 공통 규칙에 의한 다국어 자동번역 시스템을 제안하고자 한다. 공통 규칙의 장점은 전산학적으로는 여러 언어에서 단지 한번 load 되기 때문에 전체 번역 메모리의 크기를 줄일 수 있다는 것과 언어학적으로는 문법 정보의 작성.수정.관리의 일관성을 유지할 수 있다는 것이다.[1]

1. 서론

기존의 변환기반 다국어 자동번역 시스템들(SYSTRAN, EUROTRA, METAL, LOGOS, GETA 등)에서는 번역될 언어쌍에 따라 한 언어의 분석과 생성 규칙이 독립적이면서 상이하게 작성되었었다.[Hutchins 1992] 이러한 모습은 언어에 따라 공통성이 있음에도 불구하고 기존의 다국어 자동번역 시스템들이 이를 인정하지 못하게 하였다. 이때문에 기존의 다국어 자동번역 시스템들은 단순히 양언어간 자동번역 시스템을 묶어놓은 모습을 가지게 되었으며 전체 시스템의 크기를 증가시키는 결과를 초래하게 되었다. 분석과 생성 규칙 이외에도 다국어 자동번역에서 변환 과정의 수를 줄이기 위해 피봇방식을 추구하는 다국어 자동번역 시스템이 있지만(CETA, SALAT, DLT, KANT 등) 언어학적 보편성 모델을 완성하기 어렵기 때문에 진정한 피봇방식의 자동번역 시스템 구현이 이루어지지 못하고 있다[Lewis 1992].

[1] 본 논문은 다국어 자동번역 시스템 EUROTRA 의 후속 프로젝트인 독일의 통합기반 다국어 자동번역 시스템 CAT2[Sharp 1994]에서의 실험 결과를 요약한 것이다. 이 시스템은 현재 Unix 워크스테이션 환경에서 작동하고 있으며 PROLOG 로 구현되어 있고 파서로는 Constraint Bottom-Up Chart 를 사용하고 있다. 한국어에서 번역되는 언어로는 영어와 독일어가 있으며 불어,중국어,러시아어,일본어를 목표언어로 하여 번역 시험 중이다.

이런 관점에서 본 논문에서는 기존의 다국어 자동번역 시스템들이 지녔던 개별 언어적인 모듈의 문제와 피봇방식의 어려움을 극복하기 위해 공통 규칙과 제어 규칙에 의한 다국어 자동번역을 제안하고자 한다. 여기서 공통 규칙이란 다국어에 의해 공통으로 공유되는 규칙을 말한다. 공통 규칙의 장점은 단지 한번 호출(load)되어 여러 언어에서 사용되기 때문에 전산학적으로는 메모리 크기를 줄일 수 있다는 것과 언어학적으로는 문법 규칙이나 사전의 정보 구조를 관리, 작성, 수정하는데 일관성을 유지할 수 있다는 것을 들수 있다. 또한 시스템에 새로운 언어를 첨가하여 기존의 언어들과 번역하고자 할 때에도 새로운 문법 규칙을 작성할 필요가 없다는 장점을 가진다. 제어 규칙이란 각 개별언어에 존재하는 개별 언어적인 특성을 제어하는 규칙을 말한다.

본 논문은 다음과 같이 구성된다. 2 장에서는 전체 시스템 구성을 소개하며 3 장에서는 공통 규칙을 구성하는 공통 문법 규칙, 공통 사전 정보 구조, 공통 구조 변환 규칙, 공통 정보 변환 규칙을 기술하고자 한다. 4 장에서는 3 장에서 소개된 공통 규칙과 한국어와의 관계성을 살펴보고 한국어의 파라미터화된 공통규칙과 제어 규칙에 의한 한국어의 분석과 변환을 기술하겠다.

2. 전체 시스템 구성

공통 규칙과 제어 규칙에 의한 다국어 자동번역 시스템의 전체적인 구성은 그림 1 과 같다:

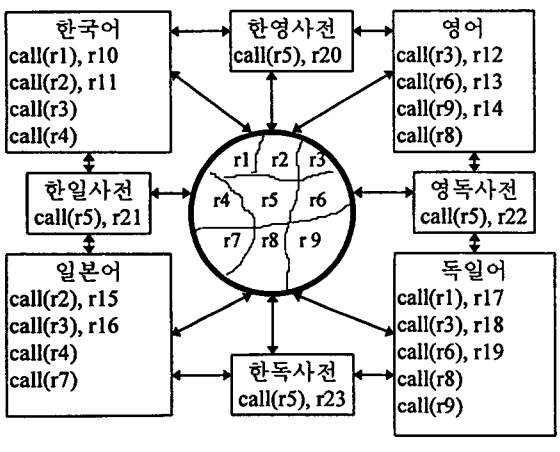


그림 1. 다국어 자동번역 시스템 전체 구성도

그림에서 굵은 선으로 된 원이 공통 모듈을 의미하며 공통 모듈 안의 r_n 는 공통 모듈을 구성하는 공통 규칙 파일들을 의미한다. 이러한 공통 규칙은 각 개별 언어에서 호출(call)됨으로써 개별 언어내에 들어 있는 제어 규칙들과 더불어 개별 언어의 모듈을 구성하게 된다. 예를 들어 그림 1 에서 한국어는 일본어, 영어, 독일어와 함께 r3 라는 규칙 파일을 공유하는 것을 알 수 있으며 영어나 독일어와는 달리 언어 친족성에 의해 일본어와 더 많은 규칙 파일인 r2, r4 를 공유하는 것을 알 수 있다.

3. 공통 규칙

본 장에서는 공통 규칙의 구성에 대해서 알아보고자 한다. 분석을 위한 공통 규칙은 공통 문법 규칙과 공통 사전 정보 구조로 이루어지며 변환을 위한 공통 규칙은 공통 구조 변환 규칙과 공통 정보 변환 규칙으로 이루어진다.

3.1. 공통 문법 규칙

다국어 자동번역 시스템에서 다국어를 처리하기 위한 공통 문법 규칙은 무엇보다 각국 언어들의 언어현상을 가능한 한 많이 수용할 수 있어야 한다. 영어와 같은 구성적(configurational) 언어뿐만 아니라 어순이 비교적 자유로운 한국어나 일본어, 독일어와 같은 비구성적(nonconfigurational) 언어들의 언어 현상을 문법에 의해 설명할 수 있기 위해 X-bar 통사 이론[Jackendoff 1977]과 HPSG[Pollard 1994]를 혼합한 새로운 문법규칙을 작성하게 되었다. 이러한 새로운 문법 규칙은 삼분법인 접속어구 규칙이외에는 모두 이분법 구조의 규칙으로 이루어졌다. 공통 문법 규칙에서 개별 언어들의 어순을 고려한 모습은 구의 머리와 비머리어와의 관계에 의해 도표 1 처럼 보인다:

구의 머리가 오른쪽인 구조(head_final_structure)
1. 인수-술어(argument-predicate) 구조
2. 수식어-피수식어(modifier-modifee) 구조
3. 인수-기능어(argument-functional word) 구조
구의 머리가 왼쪽인 구조(head_first_structure)
1. 술어-인수(predicate-argument) 구조
2. 피수식어-수식어(modifee-modifier) 구조
3. 기능어-인수(functional word-argument) 구조

1. 인수 1-접속어-인수 2(argument1-coordinator-argument2)구조

도표 1. 공통 문법 규칙들

도표 1의 다국어 공통 문법 규칙들의 CAT2 표기법에 의한 기술은 별첨 1과 같다.

3.2. 공통 사전 정보 구조

다국어 자동번역용 사전의 정보를 좀더 일관되게 입력.관리. 수정하기 위해 다국어용 사전정보구조를 작성하는 것이 필요하다. 사전의 정보구조는 가능한 한 모든 언어학적 정보를 표현해야 하며 이러한 정보를 수월하게 이동시키기 위해 단층이 아닌 다층적인 구조를 형성하는 것이 바람직하다. 이런 이유에서 다국어용 사전정보구조는 자질구조(feature structure)를 채택하였으며 속성은 여러언어에서 동일하도록 정의하였다. 다국어용 사전정보구조의 예로 중요 정보만을 나열한 예가 별첨 2와 같다.

3.3. 공통 구조 변환 규칙

변환과정도 다국어가 공유할 수 있는 부분이 있다. 즉 출발언어의 분석 출력 구조와 목표 언어의 생성 입력 구조가 동일할 때 해당 노드를 그대로 목표 언어의 노드로 복사할 수 있는 조합적 변환이 이에 해당한다. 다국어 자동번역에서 언어들간의 상이한 구조를 가능한 한 조합적으로 변환시키기 위한 방법으로써 분석과정에서 개별언어의 기능범주를 삭제하고 술어-인수-수식어의 트리배열을 갖추도록 변형시키는 방법이 이용되었다.

공통구조변환규칙에 의해 잘못 변환되어질 수 있는 비조합적 변환은 어휘에 의존하기 때문에 변환사전에 등재하였다. 변환의 순서는 우선성을 두어 변환사전의 비조합적 변환이 적용된 후 공통구조변환규칙이 적용되고 어휘변환사전을 참조하는 수순으로 이루어진다.

다국어용 공통구조변환규칙에 의한 조합 변환은 다음과 같은 모습을 가진다:

(1) 공통구조변환규칙 = { }.[+ node] <=> { }.[+ node].

위의 규칙은 Top-down, Depth-first로 적용되며 비조합적인 구조를 제외한 모든 조합적인 트리인 '+ node'를 출발언어에서 목표언어로 변화없이 변환시키는 것을 의미한다.

3.4. 공통 정보 변환 규칙

다국어에서 변환과정을 단순화하기 위한 노력은 구조변환과 정보변환을 분리함으로써도 이루어질 수 있다. 기존의 변환기반 자동번역시스템들이 구조변환에 정보변환을 포함시킴으로써 정보의 중복과 기억용량의 증대를 가져온 반면 구조변환과 정보변환을 분리한 결과는 정보의 중복과 기억용량의 크기를 감소시킬 수 있다는 장점을 가진다. 이런 의미에서 공통정보변환규칙이라는 것은 여러 언어에서 공통으로 사용할 수 있는 정보변환규칙을 의미하며 의미와 관련된 부분만을 변화없이 출발언어에서 목표언어로 복사해주는 규칙이다. 의미 정보는 개별언어를 분석할 때 형태와 의미의 함수에 의해 만들어지고 있다. 다음과 같은 규칙들이 공통정보변환 규칙으로 사용되고 있다: (CAT2의 표기법을 사용한다)

(2) 공통정보변환 규칙들

- 어휘의미의 변환

  {head: {ehead: {sem: SEM}}}.[*] <=>

  {head: {ehead: {sem: SEM}}}.[*]

- 의미격의 변환

  {role: ROLE}.[*] <=> {role: ROLE}.[*].

어휘의미의 변환규칙은 출발언어의 어휘의미정보를 노드에 상관없이 목표언어의 어휘의미정보로 복사한다는 것을 의미하며 양방향 '<=>'으로 사용될 수 있음을 나타낸다. 의미격의 변환 또한 출발언어의 의미격을 목표언어의 의미격에 복사한다는 것을 의미한다.

4. 파라미터화된 공통규칙과 제어규칙에 의한 한국어의 분석과 변환

개별언어의 문법은 보편적인 규칙과 보편적인 규칙에 대한 파라미터로 구성된다[Chomsky 1981]고 할 수 있으며 파라미터에 따라 언어유형을 분류할 수도 있다[Greenberg 1963]. 자동번역에서 보편규칙과 파라미터를 사용한 예로는 [Dorr 1993]가 있다.

Greenberg 의 파라미터화된 어순의 보편성에 의하면 한국어는 다른 외국어들과 비교하여 다음과 같은 기본 어순을 가지고 있다

(3) 한국어의 기본 어순

- SOV
- 수-명사
- 지시어-명사
- 형용사-명사
- 소유격 대명사-명사
- 관계절-명사

이러한 기본 어순은 개별언어의 파라미터를 만들 수 있는 실마리를 제공한다. 다음절에서는 한국어를 위한 파라미터화된 공통문법규칙을 살펴보고자 한다.

4.1. 파라미터화된 공통문법규칙에 의한 한국어 분석

한국어의 기본 어순에 의하면 한국어에서는 머리어가 항상 인수나 수식어의 뒤에 위치한다. 이에 따라 도표 1 의 다국어 공통문법규칙중에 머리어 후위 공통규칙만을 선택할 수 있다. 또한 정보의 이동은 HPSG 의 머리자질상승원리(Head Feature Principle)에 따라 하위구의 머리어 정보가 상위구의 정보가 된다는 것을 알 수 있다. 이에 따라 도표 1 의 다국어 공통문법규칙 부분중에서 Head-final 부분과 머리정보 상승은 다음과 같이 기술된다: (한국어의 접속어구도 '인수-기능어 구조'의 일종으로 보았다. 분석의 효율상 삼분구조를 유지하였다)

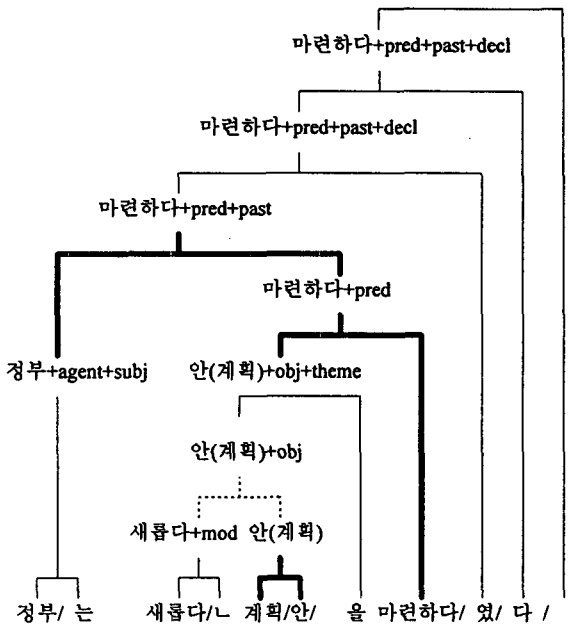
구의 머리가 오른쪽인 구조(head_final_structure)
1. 인수-술어(argument-predicate) 구조
2. 수식어-피수식어(modifier-modified) 구조
3. 인수-기능어(argument-functional word) 구조
구의 머리가 가운데인 구조(head_middle_structure)
1. 인수 1-접속어-인수 2(argument1-coordinator-argument2)구조

도표 2. 한국어를 위한 파라미터화된 공통문법규칙

(4) 머리정보 상승= {head: HEAD}.[{}, {head: HEAD}].

한국어의 파라미터화된 공통문법규칙과 (4)의 머리정보의 상승으로 한국어의 한 예문을 분석한 결과는 다음과 같다:

(5) 정부는 새로운 계획안을 마련하였다.



(5)에서 가는 실선은 '인수-기능어 구조'의 적용을, 점선은 '수식어-피수식어 구조'의 적용을, 굵은 실선은 '인수-술어 구조'의 적용을 나타낸다.

4.2. 문법 제어규칙에 의한 한국어 분석

자동번역에서 한국어를 분석할 때 특별히 고려하여야 할 사항으로는 다음과 같은 것을 들 수 있다 [오길록 1994].

(6) 한국어 분석시 고려할 사항들

- 형태음운적 특성

  형태소의 음가에 의해 후위 형태소의 형태가 결정됨

  예) 소년-이, 소녀-가
- 이중목적어의 중출

  한 문장에 목적어가 두번 나타나는 경우

  예) 그는 서울을 여행을 하였다.
- 경어법과 일치의 현상

  대화참여자의 사회적 관계에 따른 (선)어말어미와 주어

와의 관계.

예) 교수님께서 오십니다.

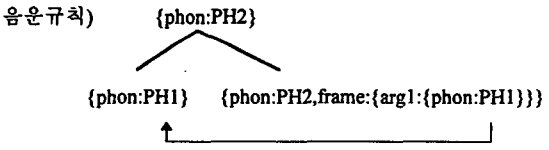
이상의 한국어 특수성들은 모두 공통 문법 규칙이나 공통 사전 정보 구조와 관련한 한국어 제어규칙에 의해 설명될 수 있다.

[illegible]	[illegible]	[illegible]
형태음운적 특성	인수-기능어 구조	음운규칙
이중목적어의 중출	인수-술어 구조	인수교환
경어법과 일치의 현상	머리정보 상승	문맥정보

도표 3. 공통 규칙과 제어 규칙

● 음운 규칙

사전의 모든 형태소에 종성음을 기입하고 기능어는 인수가 될 형태소의 종성음을 예측하게 함.



예) 소년[phon:con] - 이[phon:voc, frame:{arg1:{phon:con}}]

● 인수 교환

'하다'와 '술어명사'의 하위범주화구조를 사전에서 교환시킨다.

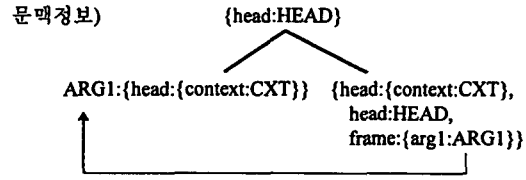
'하다'의 사전)

lex	하다			
frame	arg1	ARG1		
	arg2	ARG2		
	arg3	cat	noun	
		frame	arg1	ARG1
			arg2	ARG2

예)그는(arg1) 서울을(arg2) 여행을(arg3) 하(arg1,arg2,arg3)였다.

● 문맥 정보

문장에 나타나는 주어 명사구의 문맥정보를 동사구의 문맥정보와 일치시킨다.

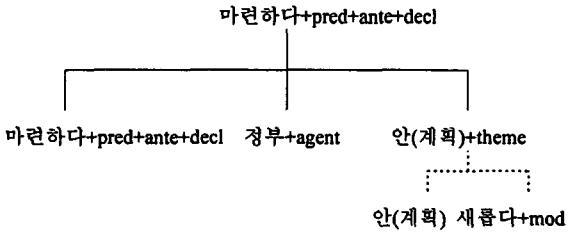


예) 교수님께서(context:honor) 오시(context:honor)ㅂ니다.

4.3. 변환 제어규칙

한국어의 통사분석 트리는 변형을 거쳐 한국어의 의미 분석 트리로 만들어 진다. 의미 분석 트리의 각 노드는 술어-인수-수식어의 어순으로 배열되며 술어-인수-수식어 노드의 술어머리 정보가 머리정보 상승규칙에 의해 상위 노드의 정보로 이동한다. 한국어 통사분석 트리 (5)는 변형에 의해 다음과 같은 의미 분석 트리로 만들어 진다.

(7) 정부는 새로운 계획안을 마련하였다.



이상의 의미 분석 트리가 변환의 입력이 되며 조합적 변환의 경우는 모두 '공통구조변환규칙'과 '공통정보변환규칙'에 의해 목표언어들로 변환된다. 하지만 조합적 변환이라도 이러한 공통정보변환규칙에 적용이 되지 않는 경우가 있다. 이것의 예로는 '하다'나 '되다'와 같은 기능동사가 붙은 한국어 관용어구를 들 수 있다. '하다' 관용어를 해결하기 위해 '하다'를 탈락시키고 '하다'의 정보를 '하다'와 연결된 '술어명사(predicate noun)'의 '기능동사(functional verb)' 자질에 복사함으로써 전체 문장의 술어가 '술어명사'가 되도록 하고 있다. 하지만 술어명사가 외국어에서는 술어명사가 될 경우도 있고 일반동사나 형용사가 될 경우가 있기 때문에 이를 제어해 줄 필요가 있다. 현재 공통정보변환규칙의 제어규칙으로 사용하는 한국어의 변환규칙

으로는 '술어명사'의 제어규칙이 있다.

(8) 술어명사의 제어규칙

● 술어명사-술어명사 제어규칙

한국어의 '하다'-술어명사에 대응되는 목표언어의 어휘가 '하다'와 같은 기능동사를 가질 경우 출발언어의 '하다'의 기능동사정보를 목표언어의 기능동사에 복사한다.

예) 산보를 하다 => take a walk, einen Spaziergan machen

散歩をする

일을 하다 => しごとをする

● 술어명사-비술어명사 제어규칙

한국어의 '하다'-술어명사에 대응되는 목표언어의 어휘가 '하다'와 같은 기능동사를 가지지 않을 경우 출발언어의 '하다'의 기능동사정보를 목표언어의 어휘에 복사한다.

예) 일을 하다 => work, arbeiten

5. 결 론

언어간 공통성을 수용함으로써 다국어 자동번역 시스템의 전체 번역 메모리를 줄이며 번역과정을 단순화시킬 수 있는 새로운 다국어 자동번역의 철학을 본 논문에서 선보였다. 이러한 철학은 다국어 자동번역시스템에서 다국어 처리를 위한 공통규칙과 제어규칙이라는 개념에 의해 설명되었다. 한국어 분석은 파라미터화된 공통문법규칙과 문법제어규칙에 의해 설명되었으며 변환과정은 공통구조변환규칙과 공통정보변환규칙, 그리고 변환제어규칙에 의해 설명되었다.

한국어의 분석과 변환을 위해 분석에서 구축된 파라미터화된 공통문법규칙과 문법제어규칙 그리고 변환에서 구축된 공통변환규칙과 변환제어규칙의 수는 다음과 같았다:

통사분석		통사-의미변형		의미분석		변환	
공통	제어	공통	제어	공통	제어	공통	제어
9	55	20	33	39	8	43	3

● 문제점 파악 및 향후 연구 방향

공통규칙과 제어규칙에 의한 다국어 자동번역 시스템에서도 아직 해결할 것이 많이 있다. 이들을 정리하면 다음과 같다:

- 파스트리 수의 축소
- 기존 사전정보와 새로운 사전정보와의 충돌
- 관용어구의 미숙한 처리
- 다의어의 처리

위의 문제들을 해결하기 위해 우리는 앞으로 다음과 같은 해결책을 고려하고 있다:

- 통계기법의 수용
- 정보유형의 다중유전(multiple inheritance)기법
- 파싱전 관용구 인식기 구현
- Domain 에 따른 역어정의

참고문헌

[오길록 1994] 오길록, 최기선, 박세영 (1994). 한글공학. 대영사.

[Chomsky 1981] Chomsky, N. (1981). Lectures on Government and Binding. The Pisa Lectures. Studies in Generative Grammar 9. Foris Publication, Doordrecht Holland & Cinnaminson U.S.A.

[Greenberg 1963] Greenberg, J.H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In: Joseph H.Greenberg(ed.,) Universals of Language. The M.I.T.Press, Cambridge, Massachusets, 2 edition.

[Dorr 1993] Dorr, B.J. (1993). Machine Translation: A View from the Lexicon. MIT Press, Cambridge, Massachusetts. London, England.

[Hutchins 1992] Hutchins, W.J. & H.L.Somers (1992). An Introduction to Machine Translation. Academic Press.

[Jackendoff 1977] Jackendoff, R.S. (1977). X-bar Syntax: A Study of Phrase Structure. Cambridge: MIT Press.

[Lewis 1992] Lewis, D. (1992). Computers and Translation. In: Christopher Butler(ed.) Computers and written Texts. Blackwell, 75-114.

[Pollard 1994] Pollard, C. and I. Sag (1994). Head-Driven Phrase Structure Grammar. Studies in Contemporary Linguistics. The University of Chicago Press, Chicago & London.

[Sharp 1994] Sharp, R. (1994). CAT2 Reference Manual Version 3.6. IAI Working Papers N.27. Saarbruecken, Germany.

별첨 1. CAT2 표기법으로 작성된 다국어 공통문법규칙

(대문자는 자질구조를 의미하며 ';'는 or를 의미하며 '>>'는 if-then을 의미한다.
'role'은 의미격을 의미하며 'frame'은 하위범주화구조를 의미한다)

구의 머리가 오른쪽인 구조(head_final_structure)
1. 인수-술어(argument-predicate) 구조= {head:HEAD}.[ARG, {head: HEAD, frame: ({arg1:ARG}; {arg2:ARG}; {arg3:ARG};{ arg4:ARG})}]. 2. 수식어-피수식어(modifier-modified) 구조= {head:HEAD}.[{role: mod, head: {restr: RESTR}}, {head: HEAD} >> RESTR]. 3. 인수-기능어(argument-functional word) 구조= {head:HEAD}.[ARG, {head: HEAD, frame: {arg1:ARG}}].
구의 머리가 왼쪽인 구조(head_first_structure)
1. 술어-인수(predicate-argument) 구조= {head:HEAD}.[{head: HEAD, frame: ({arg1:ARG}; {arg2:ARG}; {arg3:ARG};{ arg4:ARG})}, ARG]. 2. 피수식어-수식어(modified-modifier) 구조= {head:HEAD}.[{head: HEAD} >> RESTR, {role: mod, head: {restr: RESTR}}]. 3. 기능어-인수(functional word-argument) 구조= {head:HEAD}.[{head: HEAD, frame: {arg1:ARG}}, ARG].
구의 머리가 가운데인 구조(head_middle_structure)
1. 인수 1-접속어-인수 2(argument1-coordinator-argument2)구조= {head:HEAD}.[ARG1, {head: HEAD, frame: {arg1:ARG1, arg2:ARG2}}, ARG2].

별첨 2. 다국어 사전정보구조

속성	값				설명
string	STRING				문자열
lex	LEX				기본형태
first	yes / no				형태소의 첫자리 차지 가능성
last	yes / no				형태소의 끝자리 차지 가능성
pos	left / right / middle				어순(왼쪽 / 중간 / 오른쪽)
role	agent / theme / goal / location / direction / phen / ...				의미격
head	cat	명사, 동사, 부사, 조사, 선어말어미, ...			머리의 품사
	scase	주격, 목적격, 대격			표층격
	restr	RESTR			피수식어 정보
	chead	cat	명사, 동사, 부사		확장된 머리의 범주
		num	단수, 복수		수
		pform	에, 에서, 로, 부터, ...		조사 혹은 전치사의 형태
		tense	현재, 과거, 미래		표층적 시제
		aspect	완료, 진행, 완료진행		표층적 상
		modal	능력, 허락, ...		표층적 양상
		type	주절, 종속절, 관계절, 총칭,...		품사별 유형
		sem	stense	현시,과거시,미래시	시제의 의미
			saspect	일시, 유지, 종결	상의 의미
			smodal	능력, 허락, 추측,..	양상의 의미
			anim	인간, 동물, 식물	어휘 의미
frame	FRAME				인수들의 하위범주화 정보