『영어영문학』 제37권 2호
(1991) 531—557

영한 기계번역을 위한 어휘부의 모형 수립

朴秉洙 · 安相哲 · 姜汎模

1. 서 론

기계번역 시스템에 활용할 사전을 어떻게 구성할 것인가?* 우리는 이 논문에서 그러한 사전의 구성 원리에 관하여 새로운 제안을 하려고 한다. 주로 영한 기계번역을 예로 하여 여러 가지 문제들을 검토할 것이며, Pollard와 Sag (1987)의 핵어 중심 구구조 문법(Head-driven Phrase Structure Grammar: 이하 HPSG로 약칭)에 이론적 토대를 두고 사전 구성의 원리를 생각해 보고자 한다.

먼저 이와 같은 사전에 관한 우리의 논의에 전제가 되는 기계번역 시스템의 성격을 규정해 둘 필요가 있다. 우리의 기계번역 시스템의 특징은 다음과 같이 요약할 수 있다.

(1) a. 직접 변환(direct transfer) 방식을 채택한다.[1]
b. Pollard와 Sag(1987 : 191~218)에서 정의한 유형(type)에 토대를 두는 유형 구동(type-driven)의 체계이다.
c. 번역 장치는 양방향(bidirectional)으로 움직인다.
d. 전반적인 언어이론의 경향은 초어휘주의(superlexicalism)를 지향한다.

제 2절에서 (1)에서 본 우리의 번역 시스템의 특징, 그 중에서 특히 유형

*이 연구는 1988년도 문교부 학술연구조성비(대학부설연구소)의 지원을 받았음.

[1]직접변환(direct transfer) 방식은 오늘날 많은 기계번역 시스템들이 채택하고 있는 방식으로서 번역대상 언어의 문장을 중간 매개 언어의 개입 없이 직접 번역목표 언어의 문장으로 변환하는 방법이다. 이와 대조적으로 보편적 의미 중심의 중간적인 인공 언어로의 변환 과정을 거쳐서 번역을 수행하는 방식이 있는데 이를 중간언어 방식(interlingua method)이라고 부른다. 70년대 이후 대체로 전자의 방식이 기계번역 연구의 주종을 이루어 왔으나 최근에 와서 다시 중간언어 방식에 대한 관심이 높아지고 있다. 이러한 경향은 아마도 이상적인 기계번역의 방식은 중간언어 방식이지만 그 연구개발의 어려움으로 현실성이 아직도 희박하고 직접 변환 방식은 이론적으로 무리가 있으나 제한된 범위내에서 비교적 용이한 실현 가능성이 있다는 차이를 반영한다고 보여진다. 우리가 이 연구에서 직접변환 방식을 채택한 것은 이론적 고려에서라기 보다는 논의의 편의성에 있음을 밝혀 둔다.

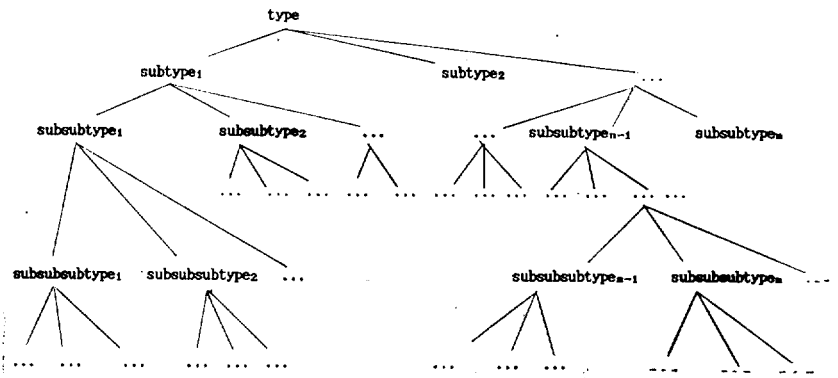
의 개념에 대하여 중점적으로 논의하고, 3절에서는 번역의 과정과 변환사전(Transfer Dictionary)의 운용과의 관계에 대하여 생각해 본다. 그리고 제4절에서 기계번역 시스템 안에서의 어휘부(Lexicon)의 모형을 제시하기로 한다.

2. 유형(Type)의 개념과 변환사전(Transfer Dictionary)의 구성

유형(type)의 개념에 대하여서는 Pollard와 Sag (1987)의 마지막 장에서 상세한 설명이 제공되어 있다. 우리는 이 유형의 개념이 어휘부(lexicon)를 조직하는데 어떠한 역할을 하는지를 알아보고자 한다.

우리가 알기로는 Pollard와 Sag이 이 개념을 최초로 언어학에 도입하였지만 이것은 컴퓨터 과학에서는 오래전부터 하나의 상식으로 통용되어 왔을 뿐만 아니라 일상적 의미에서도 하나의 상식이라 할 수 있을 것이다. 보통 사람들이 사물을 분류할 때, 일반적으로 이 개념에 바탕을 두고 분류하는 것을 흔히 볼 수 있다. 가령, 동물이나 식물과 같은 자연적인 범주나 탈 것, 집, 노래, 문학작품 등과 같은 인공적인 산물을 분류할 때 대개 이들을 하나의 위계 관계(hierarchy)에 위치시킨다. 가장 일반적인 범주를 정상에 놓고 가장 구체적인 것들을 맨 아래에 놓은 다음 그 사이에 여러 가지의 중간적인 범주들을 늘어 놓으며 이 모든 범주들을 꼭대기에서부터 밑바닥에 이르기까지 서로 연결되도록 배치한다. 이와 같은 체계가 곧 유형의 위계 관계이다. 어휘들도 이러한 방식으로 분류할 수 있다. 먼저 어휘들을 어떤 유

〈그림 1〉 포섭도의 유형



형(type)으로 분류하고 다음에는 이들을 다시 하위유형(subtype)으로 세분하고 또다시 그보다 더 작은 하위하위유형(subsubtype)으로 분류하여 마지막에는 매우 구체적인, 작은 유형으로 분류하는 데까지 도달할 수 있을 것이다. 이와 같은 분류방법을 앞의 〈그림 1〉로 나타낼 수 있는데 이와 같은 형태의 수형도를 포섭 수형도(subsumption tree) 또는 포섭도(subsubsumption graph)라고 부른다.

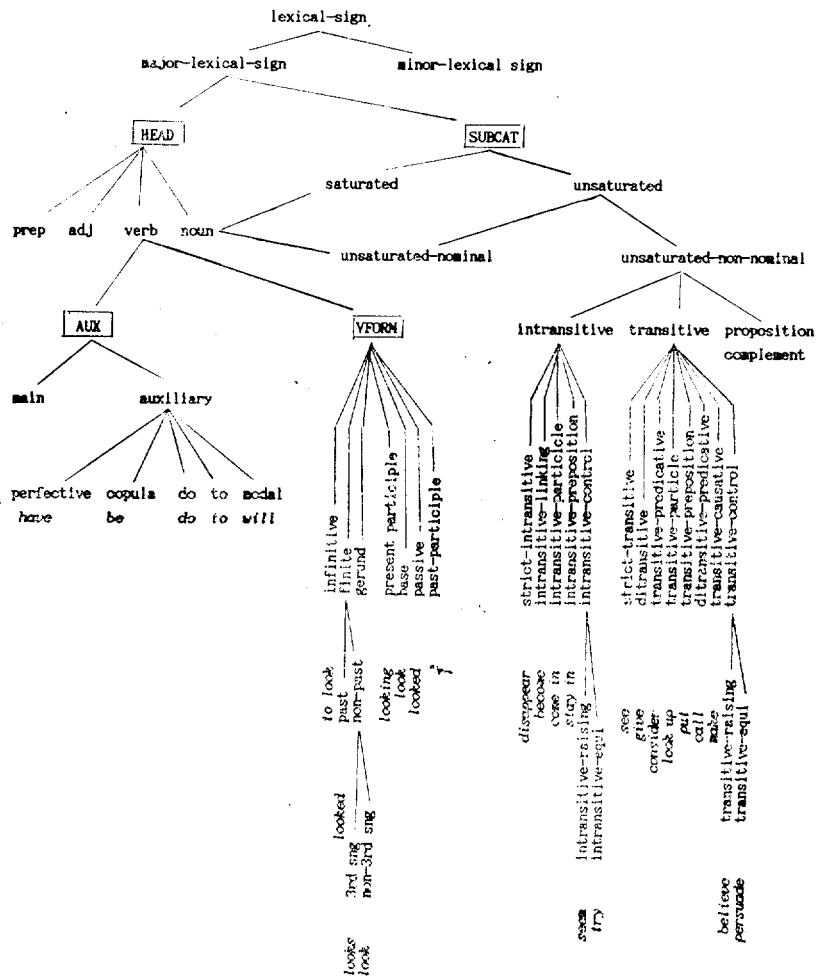
포섭은 정보 내용의 양에 의하여 결정되는 관계이다. 어떤 요소가 정보 내용의 양이 많으면 많을수록 구체적인 요소가 되고 반대로 적으면 적을 수록 일반적인 요소가 된다. 가령 「명사」라는 정보 내용만 가진 요소는 이 「명사」 정보 이외에 「단수」라는 정보를 하나 더 가진 요소보다 더 일반적인 요소이다. 이때에 전자가 후자를 포섭한다고 말한다. 다시 말하면 일반적인 정보 내용을 가진 요소가 그보다 구체적인 정보 내용을 가진 요소를 포섭한다. 그러므로 일반적으로 말해서 A가 B를 포섭하면 B는 A가 가진 모든 정보를 가지고 거기에 추가해서 A에 없는 새로운 정보를 하나 이상 가진다. 〈그림 1〉에서 type은 subtype를 포섭하고 subtype은 subsubtype을 포섭하며 다시 그것은 subsubsubtype을 포섭한다.

영어 동사를 이와 같은 방식으로 분류할 수 있다. 즉 하위분류범주화(subcategorization) 요건을 기준으로 하여 동사들을 여러 하위 집단으로 나눌 수 있다. 이 포섭도의 정상에 놓일 유형은 동사가 될 것이다. 그 아래에 자동사와 타동사가 그 하위유형으로 놓이게 될 것이다. 그리고 다시 그 아래로 동사가 요구하는 보어(complement)와 수식어(modifier)의 수와 종류에 따라 여러 가지 하위유형들로 계속 세분되어 나갈 것이다. 아래 〈그림 2〉에서 이를 포섭도로 표현해 보았다. 이것은 Pollard와 Sag(1987)의 것을 일부 수정하거나 확장하여 만든 것이다.

최상위 유형이 lexical-sign(어휘기호)이다. 이는 모든 어휘를 포함한다. 그 하위유형이 major-lexical-sign과 minor-lexical-sign인데 전자는 무슨 기준에 의하여서든지 더욱 세분할 필요가 있는 유형이고 후자는 그렇게 할 필요가 없는 유형이다. 그래서 전자를 두 가지 기준으로 분류한다. 왼쪽으로 머리자질(head feature)에 의한 하위유형이 나타나고 오른쪽으로는 하위범주화(SUBCAT)에 의한 하위유형들이 나타난다. 동사는 AUX 기준에 의하여 주동사 유형(main)과 조동사 유형(auxiliary)으로 분류되고 이와 동시에 형태적 기준(VFORM=verb form)에 의하여 infinitive, finite 등 일곱가지 유형으로 분류된다.

하위범주화 기준에 의하여 모든 주요 어휘 유형은 포화(saturated) 유형과 불포화(unsaturated) 유형으로 2대별된다. 포화 유형은 보어를 필요로 하

〈그림 2〉 영어 주요어휘의 포섭도(동사)



지 않는 것이고 불포화 유형은 보어를 요구하는 것이다. 불포화 유형은 다시 명사적인 것과 비명사적인 것(unsaturated-non-nominal)으로 나누어지고 후자는 자동사 유형과 타동사 유형과 전치사 보어 유형으로 나누어지며, 이들은 다시 보어의 종류에 따라 여러 가지 유형으로 나누어진다.

이러한 포섭도에서 얻는 정보 중 가장 중요한 것은 포섭 관계이다. 즉 상위유형(supertype)과 하위유형(subtype) 사이에 포섭 관계가 성립한다는 점이다. 이와 같이 포섭관계가 성립하면 하위유형은 상위유형이 지니는 모든

정보를 물려 받게 된다. 다시 말하면, 포섭되는 하위유형은 포섭하는 상위유형에 담겨 있는 모든 정보를 모두 그대로 지니면서 추가적으로 새로운 정보를 지니게 된다.

Pollard와 Sag(1987 : 8장)이 지적하는 바와 같이 포섭 관계와 정보 상속 장치를 활용하면 어휘 엔트리안에 동일한 정보가 계속해서 등장하는 중복의 문제점을 완전히 제거할 수 있는 매우 경제적인 사전을 조직할 수 있다고 본다. 예를 들면 영어 동사 persuaded는 〈그림 2〉 포섭도에 의하여 [main], [verb], [major-lexical-sign], [lexical-sign] 등의 유형이 됨과 동시에 [past]와 [finite] 유형이라는 정보를 얻고, 또 한편으로는 [transitive-equi], [transitive-control], [transitive], [unsaturated-non-nominal], [unsaturated] 등 유형이 된다는 정보를 얻는다.

그러나 persuaded의 엔트리에는 이들 모든 유형에 관한 정보를 다 나타낼 필요가 없다. 필요한 것은 [main], [past], [transitive-equi] 등 포섭도 최하단에 나타나는 세 가지 유형만 밝혀주면 된다. 이 세 가지를 포섭하는 상위의 유형들은 모두 생략해도 좋다. 이러한 방법으로 그 단어가 가지고 있는 여러 가지 정보 중에 많은 다른 단어들도 공유하는 정보는 그 단어의 사전 엔트리에 일절 포함시키지 않아도 된다. 이러한 정보는 포섭도의 정보상속장치에 의하여 자동적으로 공급될 수 있는 것이다. 그러한 장치는 일반적으로 다음과 같은 형식으로 되어 있다.

(2) 만약 어떤 어휘가 A유형이면, 그것은 또한 B유형이다. 단 포섭도에서 B가 A를 포섭해야 한다.

예컨대 persuaded가 [main], [past], [transitive-equi]이면 그것은 또한 [verb], [major-lexical-sign], [transitive], [unsaturated-non-nominal], [unsaturated]이기도 하다. 이에 따라 persuaded는 앞의 세 유형이 갖는 정보 뿐만 아니라 후자의 모든 유형이 지니는 정보를 전부 지니게 된다.

우리는 이 두 개의 서로 연관된 개념, 즉 포섭과 정보 상속이 기계번역에 있어서 중요한 역할을 하게 되는 것을 보여줄 것이다. 특히 번역 대상 언어(source language)와 번역 언어(target language) 사이에 등가어(equivalent word)를 병치시키는 작업에 결정적인 역할을 하는 것을 보게 될 것이다.

유형 분류에서 또 한가지 유의할 사항은 유형 분류는 일반적으로 교차 분류(cross-classification)의 성질을 띠운다는 점이다. 즉 하나의 요소가 관련된 분류에는 대체로 몇 개의 분류 기준이 동시에 연관되어 있다는 점이다. 이 점은 persuaded의 예에서 이미 확인한 바 있다. 이 동사의 유형 분류에

세 가지 분류 기준이 동시에 연관되어 있었다. AUX 기준에 의하여 main 유형, VFORM 기준에 의하여 finite 유형, 그리고 SUBCAT 기준에 의하여 transitive 유형으로 분류된다. 유형 분류의 이러한 성질도 우리의 유형 구동의 기계번역 시스템과 밀접한 관련이 있음을 보게 될 것이다.

이제 유형 구동(type-driven)의 기계번역 시스템의 기본 원리에 대하여 좀 더 자세히 설명할 단계가 되었다. 먼저 유형 구동이란 무엇인가? 앞에서 우리는 유형의 개념에 대하여 자세히 논의하였는데, 유형 구동의 과정이란 번역 대상 언어와 번역 언어의 유형들을 대조시키는 과정을 말한다. 이와 같은 유형대조는 양 언어의 변환사전(transfer dictionary)에 포함되는 정보이다. 그리하여 변환사전은 두 가지의 중요한 대조 또는 대비에 관한 정보를 제공한다. 하나는 양 언어 사이의 유형 대비에 관한 정보이고 다른 하나는 어휘형태에 관한 정보이다.

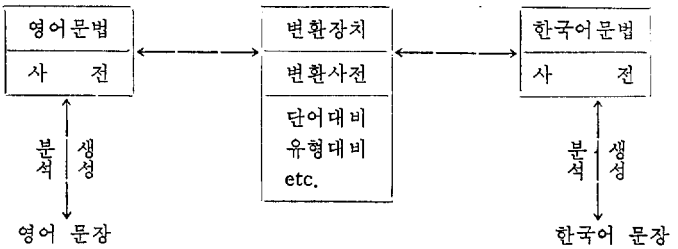
그리고 우리의 기계 번역 시스템에서는 번역 대상 언어의 문법과 번역 언어의 문법은 완전히 독립적인 체제를 유지하는 것으로 한다. 두 가지 의미에서 독립적이다. 첫째, 그 두 문법은 서로 독립적이다. 한 언어의 문법이 다른 언어의 문법을 간섭하지도 않고 하나가 다른 것에 의존하지도 않는다. 둘째, 언어 문법은 번역 시스템으로부터 독립적이다. 문법이 번역 시스템의 장치에 전혀 구애 받지 않는 것을 이상적으로 한다.

예를 들면 영한 기계번역 시스템에서 우리는 영어 문법과 한국어 문법을 따로 따로 구성하며 각기 그 자체로서 자급자족하고 서로 영향을 주고 받지 않도록 한다. 그리하여 이상적으로 말하면, 전혀 번역에 구애됨이 없이 언어학적으로 가장 훌륭한 영어 문법과 가장 훌륭한 한국어 문법을 구축하려고 노력한다. 말하자면 언어학적으로 가장 좋은 문법이 기계번역에 있어서도 가장 좋은 문법이 되어야 한다는 입장이다.

다음으로 기존의 전환 방식 시스템과는 달리, 분석(parsing) 과정에 사용하는 문법과 생성(production) 과정에 사용하는 문법이 동일한 것으로 본다. 그리고 변환 장치도 동일한 하나의 장치만으로 되어 있고 이것이 영한 번역도 수행하고 한영번역도 수행할 수 있도록 한다. 이것이 곧 양방향 번역 시스템이다. 우리가 가정하는 번역 시스템의 조직은 대강 아래 〈그림 3〉으로 나타낼 수 있다.

지금까지 우리의 기계번역 시스템의 세 가지 특성에 대하여 논의하였다. 유형구동의 개념에 대해서는 다소 길게 논하였고, 양방향의 특징과 변환 방식의 특징에 대하여서는 비교적 간단히 언급하였다. 나머지 두 특징, 즉 통합적 토대와 초어휘론적 성격은 앞으로 논의 과정에서 구체적으로 밝혀질 것이다. 이하에서 우리는 하나의 실제 문장을 예로 하여 그 번역의 실례를

〈그림 3〉



검토함으로써 유형 구동의 번역 시스템의 윤곽을 설명하려고 한다.

3. 번역의 과정과 변환사전의 운용

다음 영어 문장 (3)을 한국어 문장 (4)로 번역하고 다음에는 역으로 (4)를 (3)으로 번역하려고 한다.

(3) John persuaded Mary to write two letters to Bill after we left.
(4) 우리가 떠난 뒤에 존이 메리에게 두 통의 편지를 빌에게 쓰도록 설득하였다.[2]

(3)을 (4)로 번역하기 위하여 번역 시스템은 적어도 다음 사항을 인지해야 한다. John이 주어라는 것. Mary와 to write two letters to Bill이 주동사가 하위범주화하는 두 보어라는 것. 두번째 보어내에서 to write는 write의 부정사 형태이며, two letters와 to Bill은 동사 write의 보어라는 점. 그리고 after we left는 종속절이며 그것은 주절의 행동이 발생한 시점을 가리킨다는 점도 인지해야 한다.

그러한 정보를 처리하는데 필요한 원리와 규칙을 Generalized Phrase Structure Grammar(GPSG)와 HPSG에서부터 자유롭게 도입한다. 그리하여 영어 문법의 어휘부에는 persuade와 write와 같은 동사의 하위범주화 정보가 다음과 같은 형식으로 표현된다.

(5) a. persuade 유형 : [SUBCAT 〈VP[INF], NP[ACC], NP[NOM]〉]
b. write 유형 : [SUBCAT 〈PP[*to*], NP[ACC], NP[NOM]〉]

INF=infinitive, ACC=accusative case, NOM=nominative case

[2]이 한국어 문장이 (3)을 아주 자연스럽고 훌륭하게 번역한 것이라고 생각되지 않지만 논의의 필요에 따라 적정한 번역문으로 가정한다.

주어를 포함하여 모든 보어들을 개별 동사의 어휘적 속성으로 보고 이들을 이와 같은 사격성(obliqueness)의 순서에[3] 따라 배열한다. 이 점은 HPSG의 방식을 따른 것이고, VFORM, PFORM, CASE 등 보어의 형태를 밝혀주는 데 필요한 통사 자질들을 첨가해 주는 것은 원래 GPSG의 방식이다.

우리의 문법은 HPSG에서와 같이 문법규칙을 포함한다. HPSG의 문법규칙은 극도로 일반적이다. GPSG의 경우처럼 개별적인 동사나 명사를 실제로 도입하는 구구조 규칙과는 전혀 다른 규칙이다. 이런 의미에서 HPSG의 문법규칙은 오히려 원리(principle)에 가깝다.

(6) 문법규칙 1

어휘 핵은 그 보어(들)와 결합하여 불포화 상태의 구를 이룬다. 단 이 구는 단 하나의 보어를 제공하면 포화상태에 도달하는 구이다.

(7) 문법규칙 2

비어휘 핵은 그 보어와 결합하여 포화상태의 구를 이룬다.

(8) 문법규칙 3

비어휘 핵은 그 수식어와 결합하여 원래의 핵과 동일한 범주의 구를 이룬다.

문법규칙 1과 2는 HPSG의 규칙 그대로이고 문법규칙 3은 GPSG로부터 채택한 것이다. 문법규칙 3을 수립함으로써 수식어는 어휘 핵이 요구하는 요소가 아니라 구 핵의 자매 성분(sister)으로 분석하기로 한다.

문법규칙과 더불어 원리들이 있다. 이 원리들은 성분들의 결합 가능성에 제약을 가하는 역할을 한다. HPSG에서는 이를 위하여 두 개의 보편 원리를 수립한다. 하위범주화의 원리(Subcategorization Principle)와 핵자질의 원리(Head Feature Principle)가 그것이다.

---

[3]사격성 위계(obliqueness hierarchy)는 HPSG 이론에서 핵심적 위치를 차지하는 개념이다. 원칙적으로 더 이상 정의할 수 없는 원시 개념이지만 상식적으로 이해할 수도 있는 개념이다. 의미적으로 동사와의 관련성이 더 밀접한 요소가 덜 밀접한 요소보다 사격성이 더 크다고 말한다. 다른 말로 하면, 사격성 위계란 문장의 구성요소들의 동사 의존성의 정도의 차이라고 볼 수 있다. 동사에 의미적으로 의존하는 정도가 크면 클수록 사격성이 크다고 한다. 가령 John put the vase on the table에서 전치사구 on the table이 가장 동사 의존성이 높고 주어 John이 가장 낮다고 보며 목적어 the vase가 그 중간이라고 본다. 그래서 사격성의 큰 것부터 차례로 놓으면 전치사구—목적어—주어의 순서가 된다. 이와 같은 사격성 위계는 어순을 결정하는 문제, 표현되지 않은 주어를 찾는 통제(control)의 문제, 조응사의 선행사를 정해주는 결속(binding)의 문제 등을 다루는 데 핵심적인 역할을 한다. Pollard Sag (1987, 1989, 1990) 참조.

하위범주화의 원리는 불포화 상태의 어휘핵이 포화상태로 진전되는 과정에서 제공되는 보어와 새로 형성되는 구의 SUBCAT 리스트와의 관계가 어떻게 되는가를 정의하는 원리이다.

(9) 하위범주화의 원리
핵이 있는 구의 SUBCAT 값은 그 구의 핵 딸 성분(head daughter)의 SUBCAT 리스트에서 제공된 보어 성분(들)을 삭감한 나머지와 같다.

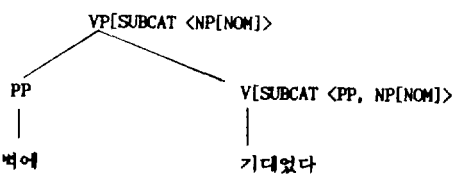
이 원리와 문법규칙의 작용으로 문장의 성분구조를 분석할 수 있다. 종전의 생성문법에서 구구조규칙이 하는 일을 HPSG에서는 보편적인 문법규칙과 원리로써 하게 되는 것이다. 그러나 어휘들이 갖는 고유의 특성으로서 SUBCAT 값이 미리 어휘부에 명세화되어 있어야 한다.

예를 들어 한국어 동사 '기대다'는 원칙적으로 '-에'와 같은 장소를 뜻하는 후치사구(postpositional phrase)를 요구하므로 이 점을 이 동사의 SUBCAT 리스트에 반영해 두어야 한다.

(10) 기대다 : [SUBCAT 〈PP, NP[NOM]〉]

이 어휘범주는 문법규칙 1에 의하여 보어 PP와 결합하여 불포화 상태의 구를 이루게 된다. 이렇게 구성된 구의 하위범주화 값은 하위범주화의 원리에 따라 결정된다. 즉 어휘범주 '기대다'의 SUBCAT 리스트 〈PP, NP[NOM]〉에서 제공된 보어 PP를 삭감하면 〈NP[NOM]〉이 남게 되는데 이것은 그 불포화 상태의 구의 하위범주화 값이 된다.[4]

(11)

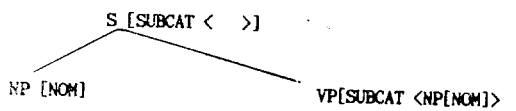


다음에는 문법규칙 2에 의하여 비어휘 범주는 NP[NOM] 보어와 결합하여 포화상태에 도달한다. 그리고 그것의 하위범주화 값은 NP[NOM]을 삭

[4]이하에 사용하는 수형도는 전통적인 변형문법에서 보는 수형도와는 다르다. 이것은 GPSG와 HPSG의 방법을 적절히 절충한 수형도라고 할 수 있다. 설명의 편의상 HPSG의 자질구조를 GPSG식으로 풀어서 나타내었다.

감한 나머지인 〈 〉이 된다.

(12)



그리고 (11)에서 V가 PP가 결합하여 VP를 이루고 (12)에서 VP가 NP와 결합하여 S를 이루는데 이것은 무엇으로 보장되는가? 이것이 곧 핵자질의 원리에 의한 결과이다.

(13) 핵자질의 원리 (Head Feature Principle)
구가 핵성분을 가질 때 그 구와 핵성분은 동일한 핵자질을 공유한다.

(11)의 모범주가 VP인 것은 그것의 핵구성성분(head daughter)이 V이기 때문이며, (12)의 모범주가 S인 것은 그것의 핵구성성분이 VP이기 때문이다.[5] 이러한 관계를 핵자질의 원리가 포착하는 것이다.

여기서 HPSG의 이론 장치를 자세히 논의할 수도 없고 그렇게 할 필요도 없다. 다만 영한 기계번역의 과정, 그 중에서도 특히 사전 구성에 관련되는 범위 내에서 HPSG의 이론을 간략히 검토하는 것으로 충분하다.

지금까지 주로 영어 문법의 영역을 살펴보았다. 한국어 문법의 경우에도 근본적으로 영어의 경우와 같거나 비슷한 장치가 필요하다. 앞에서 설명한 문법규칙은 한국어의 경우에도 그대로 적용된다고 보여진다. (그러나 그 세 가지 문법규칙이 필요한 모든 경우를 다 포괄한다고 주장하는 것은 아니다. 개별적인 사례를 처리하려면 문법규칙의 수가 다소 늘어날 것으로 전망된다.)

사전의 구성도 원칙적으로 영어와 한국어가 같은 방법으로 처리된다. 영어 동사 persuade와 write에 해당하는 한국어 동사 '설득하다'와 '쓰다'의 어휘 속성, 특히 하위범주화에 관한 정보가 아래 (14)와 같이 표시된다.

(14) a. 설득하다 : [SUBCAT 〈VP[도록], NP[DAT], NP[NOM]〉]
b. 쓰다 : [SUBCAT 〈NP[DAT], NP[ACC], NP[NOM]〉]

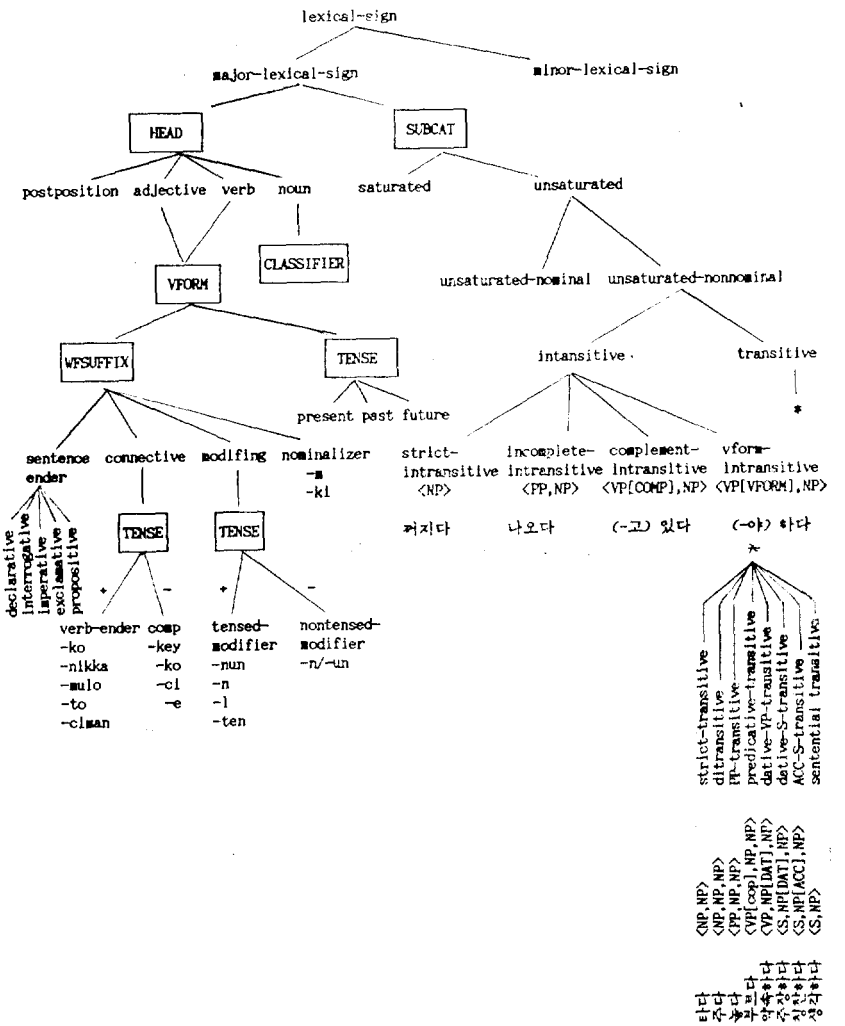
하위범주화 요구 조건은 영어의 경우와 비교할 때 세부 사항에 있어서는

[5] V와 VP와 S는 모두 핵자질(cat verb)을 공유한다. V는 [LEX+]이고 VP와 S는 [LEX-]이며, VP는 [SUBCAT 〈NP〉] 이고 S는 [SUBCAT〈 〉]이다.

차이가 있으나 전체적인 윤곽은 거의 동일하다는 점에 유의할 필요가 있다. 따라서 한영 번역을 수행할 때 격이라든가 보문소(complementizer)등 한영간의 차이점이 드러나는 세부사항에 특히 유의하여야 한다.

영어 문장 (3)을 한국어 문장 (4)로 번역하는 과정을 설명하는 데 있어 부사 부가어 (즉 after we left)는 당분간 보류해 두고 주절에만 주의를 기울

〈그림 4〉 한국어 주요어휘의 포섭도(동사와 형용사)



이기로 한다.

우리의 유형구동의 MT 체계의 중심부문이 되는 것은 변환사전(transfer dictionary)이다. 이것은 영한 변환사전과 한영 변환사전을 포함한다. (이제부터 "E-K 사전", "K-E 사전"이라고 명명한다.) 변환사전은 기본적으로 두 가지 정보를 담고 있다. 첫째, 영어 어휘에 해당하는 한국어 어휘가 무엇인가를 알려 준다. 둘째, 한 언어의 유형이 다른쪽 언어의 무슨 유형과 일치되는가를 알려 준다. 간단히 말하면, 변환사전은 영어와 한국어 사이에 단어의 형태와 유형에 있어서의 대비관계가 어떻게 되는가에 관한 상세한 정보를 제공한다.

영한 유형대비를 구체화하기 위하여 앞의 〈그림 2〉와 같은 한국어 어휘를 위한 유형 포섭도를 수립해야 한다. 우선 동사와 형용사에 촛점을 두고 〈그림 4〉의 유형 포섭도를 설정해 보았다.

VFORM의 기준을 2등분하여 WFSUFFIX(=word formation suffix)와 TENSE로 나눈다. 전자는 전통적인 한국어 문법에서 활용어미에 해당하는 요소를 말하는 것으로서 종결어미(sentence ender), 연결어미(connective), 수식어미(modifying), 명사형어미(nominalizer) 등 네 가지 유형으로 나뉜다. 연결어미는 시제 형태소를 동반할 수 있는가 없는가에 따라 verb ender (동사 종결형)와 보문소(complementizer)로 나눈다. 예컨대, '먹고, 먹겠고, 먹었고' 등처럼 '-고'는 시제형태소를 수반할 수 있다. 그러나 '-게, -고, -지, -어/아' 등은 시제형태소를 수반할 수 없다. '먹게 되었다'는 좋으나 '*먹었게 되었다'는 불가능하다. 수식어미도 시제가 있는 것과 시제가 없는 것으로 나누어진다. '먹는(현재), 먹은(과거), 먹을(미래), 먹던(과거회상)' 등과 같이 시제가 달라지면 어미가 달라진다.

수식어미 역시 시제 동반가능성의 유무에 따라 두 가지 종류로 분류된다. 동사는 시제를 동반하는 반면에 형용사는 시제를 동반할 수 없다. '먹-는'(현재), '먹-은'(과거), '먹을'(미래), '먹-던'(회상) 등과 같이 시제에 따라 다른 수식어미가 붙는다. 그러나 형용사는 시제와 무관한 '-ㄴ/-은'이 있을 뿐이다. '나쁘-ㄴ', '희-ㄴ', '검-은', '젊-은'과 같이 형용사에 붙는 '-ㄴ/은' 수식어미는 동사에 붙는 '-ㄴ/은'과 구별해야 한다. 후자는 과거시제를 나타내는데 반하여 전자는 과거시제와 관계없다. 형용사의 시제는 수식어미에 나타나지 않고 문장의 다른 부분, 주로 주 동사에 나타나는 시제에 의하여 결정된다.

(15) a. 젊은 학자이다. (현재)
b. 젊은 학자이었다. (과거)
c. 젊은 학자일 것이다. (미래)

형용사 어간 다음에 오는 어미 '-은'은 시제와 무관한 것으로 보아야 한다. 이것은 동사어간 다음에 오는 '-은'과 구별되어야 한다. 후자는 과거시제를 나타내지만 전자는 시제에 관한 한 중립이다.

명사화 어미에는 '-ㅁ/음'과 '-기'가 있다. ('먹음/봄/', '먹기', '좋음/큼', '크기' 등등) 동사어미와 형용사 어미를 구별하지 않고 이 두 어미가 붙는다.

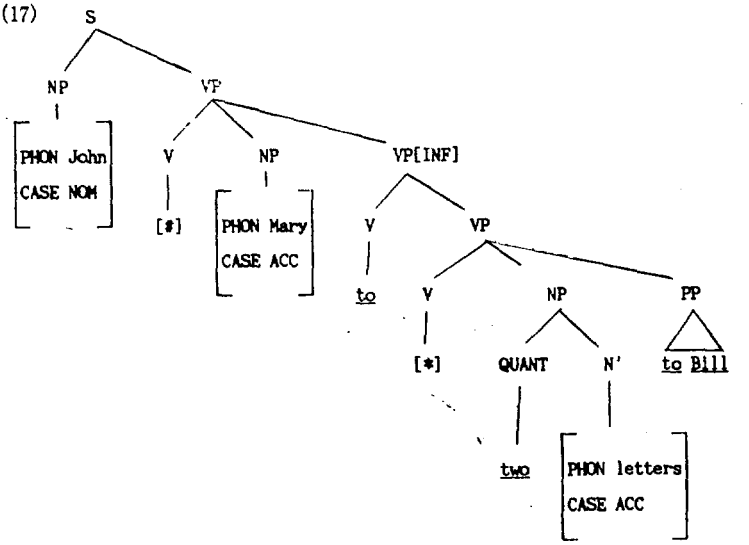
SUBCAT을 기준으로 하는 분류에 있어서는 보어를 필요로 하는 불포화 유형(unsaturated)이 문제가 된다. 이 유형은 명사류(unsaturated-nominal)와 비명사류(unsaturated-nonnominal)로 나누어진다. 〈그림 4〉는 불포화 비명사류에 치중하여 분류작업을 수행한 것이다. 완전자동사류(strict-intransitive)는 주어 NP만을 요구하는 동사들이다. 물론 이것이 가장 단순한 동사이다. 이 완전자동사류를 포함하여 열 두 종류의 동사-형용사류를 분류하였다. 아래에 그 예와 예문을 제시한다.

(16) a. 완전자동사류(strict-intransitive)
'불이 꺼졌다.'
b. 불완전자동사류(incomplete-intransitive)
'아이들이 밖으로 나왔다.'
c. 보어자동사류(complement-intransitive)
'아이들이 놀고 있다.'
d. 동사형태자동사류(vform-intransitive)
'아이들은 잘 뛰어 놀아야 한다.'
e. 완전타동사류(strict-transitive)
'아이들이 그네를 탄다.'
f. 중목적어타동사류(ditransitive)
'어른들이 아이들에게 선물을 준다.'
g. 후치사구타동사류(PP-transitive)
'아이들이 그릇을 밥상에 놓았다.'
h. 서술타동사류(predicative-transitive)
'아이들이 그를 아저씨라고 부른다.'
i. 여격-동사구타동사류(dative-VP-transitive)
'우리는 아이들에게 다시 올것을 약속했다.'
j. 여격-문장타동사류(dative-S-transitive)
'우리는 그들에게 그것이 옳다고 주장했다.'
k. 대격-문장타동사류(accusative-S-transitive)
'우리는 그들을 잘 했다고 칭찬했다.'
l. 문장-타동사류
'우리는 그들이 정직하다고 생각한다.'

이제 다시 문장 (3)에 돌아가서 그것의 번역 과정을 검토하기로 한다.

영어 문장 (3)을 parsing한 결과로서 우리는 다음 수형도가 나타내는 정보를 얻게 된다.

(17)



(#)
PHON	persuaded
INV	−
MAIN	+
FORM	PAST
SUBCAT	⟨VP[INF], NP[ACC], NP[NOM]⟩

(*)
PHON	write
VFORM	BASE
SUBCAT	⟨PP, NP[ACC], NP[NOM]⟩

그리고 영한 변환사전에서 단어와 유형에 관한 다음의 정보를 얻는다.

(18) a. 단어 변환 : (영) persuade- (한) 설득하다.
b. 유형 변환 :
(i) (영) PAST - (한) PAST
(ii) (영) uninverted-main - (한) declarative
(iii) (영) transitive-equi - (한) transitive-dative-VP 또는 transitive-accusative-VP

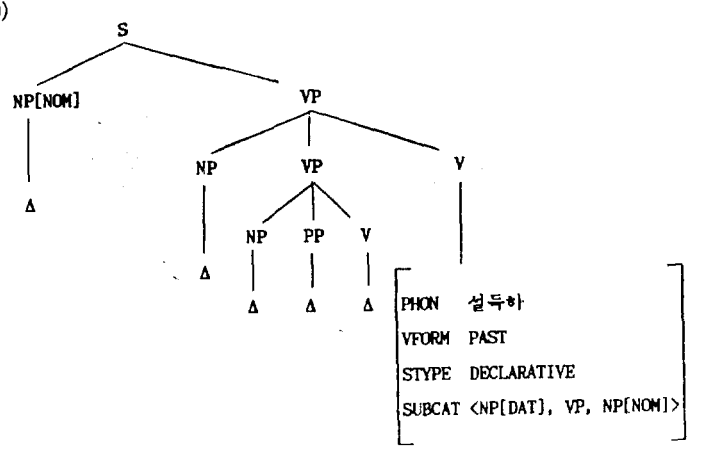
한국어 어휘부에는 하위범주화 유형으로 보면 적어도 세 가지 다른 '설득

하다' 동사가 있는 셈이다. 첫째, NP 하나만 요구하는 순수한 타동사, 둘째, VFORM '고'를 수반하는 VP와 대격 명사구를 요구하는 타동사, 셋째 VFORM '도록'을 수반하는 VP와 여격 명사구를 요구하는 타동사가 그 셋이다.

(19) a. 존이 메리를 설득했다.
b. 존이 편지를 쓰라고 메리를 설득했다.
c. 존이 메리에게 편지를 쓰도록 설득했다.

(18b)에서 밝힌 유형 변환에 관한 상세화 중의 세번째 정보에 의하여 MT 체계는 (19b) 또는 (19c)를 선택하게 된다. 일단 목표 문장 (즉, 한국어 번역문)의 주동사의 유형이 결정되면 한국어 문법규칙과 하위범주화 원리에 의하여, 그 구조의 골격이 형성될 수 있다. 아래 수형도가 바로 그 구조가 되는데 여기에는 아직 채워지지 않은 곳이 여러 군데가 있다.

(20)



빈 자리를 채우는 과정은 전통적인 변형생성문법에서의 어휘삽입과정과 유사한 과정이라고 볼 수 있다. 영한 변환사전과 한국어 사전에서 적절한 단어를 선택하여 빈자리를 채우게 되는데, 이때 원어문의 구조 (즉 (17))와 목표문의 구조 (즉 (20))를 비교함과 동시에 영어 persuade와 한국어 '설득하다'의 하위범주화 요구사항들을 대비함으로써 빈 자리 채우기 과정이 진행될 수 있다.

그런데 유형포섭도에 보면 각 동사 유형마다 하위범주화 리스트가 있는데

거기에는 VFORM에 관한 자세한 정보가 없다. 따라서 변환 사전에도 V-FORM에 관한 정보가 없다. 그러한 정보는 변환사전에 있는 것이 아니고 한국어 사전에 담아두는 것으로 한다. 영한 변환사전에는 가능한 한 일반적이고 규칙적인 정보를 담도록 하고, 불규칙적인 정보나 기타 산발적인 정보는 한국어 사전에 등재하기로 한다.

이러한 전략에 의하여 (20)의 주동사의 SUBCAT 값 중의 VFORM 값은 한국어 사전의 '설득하다' 엔트리를 찾아봄으로써 '도록'으로 판명된다.

(21) '설득하' [SUBCAT 〈NP[DAT], VP[VFORM 도록], NP[NOM]〉]

다음에 NP[DAT]가 지배하는 자리는 영어 대격 명사구 Mary의 번역으로 채워지는데 이는 해당되는 하위범주화 요구사항 (즉, '설득하다'에 의하여 요구되는 사항)에서 오는 국부수형도내에서 Mary가 유일한 NP이기 때문이다. 같은 이유로 VP[도록]은 영어 부정사 VP 'to write two letters to Bill'의 번역으로 채워진다. 마지막으로 주어 NP도 같은 방법으로 결정된다.

지금까지의 논의를 정리하면 다음과 같이 요약할 수 있다. 목표문장의 구조는 문법규칙과 하위범주화 원리에 의하여 결정되고 필요한 어휘들은 원문의 구조에서 얻는 정보와 목표언어의 사전과 변환사전 등 3자의 상호 작용에서 제공된다.

목표문의 구조를 통합해 내는 과정은 사실상 성분들과 그 정보의 통합(unify) 과정이다. 통합이 성공적으로 진행되는 범위 내에서 구조가 형성될 수 있고 통합이 진행되지 못하면 구조는 형성될 수 없다. 예컨대 세 개의 성분 '메리에게'와 '빌에게 편지를 쓰도록'과 '설득하였다'가 결합하여 하나의 성분구조, 즉 VP를 형성할 수 있는 것은 문법규칙 2와 하위범주화의 원리 그리고 변환사전과 개별언어사전에서 얻은 여러가지 관련 정보에 따라서 그 세 개의 성분에 포함된 모든 정보 내용들이 통합될 수 있기 때문이다.

유형 변환의 과정이 사전과 어떻게 상호 작용하는가를 좀 더 자세히 살펴보기 위하여 영어의 상승동사 believe와 want가 관련된 번역과정을 검토해 보기로 한다. 이 두 영어 동사에 해당하는 한국어 동사를 각각 '생각하다'와 '원하다'로 본다. 아래에 해당 예문을 든다.

(22) a. believe John to be honest.
b. want John to be honest.

(23) a. 존이 정직하다고 생각한다.
b. 존이 정직하기를 원한다.

영한 변환사전에서 얻을 수 있는 유형 대비는 아래와 같이 될 것이다.

(24) (영) transitive-raising - (한) transitive-clausal

그리고 한국어 동사들이 취하는 보어절의 VFORM과 CASE에 관한 정보는 한국어 사전에 아래와 같이 상세화 되어 있다.

(25) a. 생각하다 : [SUBCAT ⟨S[DECL, 고], NP[NOM]⟩]
b. 원하다 : [SUBCAT ⟨S[NOMINAL, ACC], NP[NOM]⟩]

MT 체계는 변환사전에서 정확한 역어와 정확한 유형이 무엇인지 알아낼 수 있다. 그러나 거기에는 아직 보어절의 VFORM 값에 대한 정보가 없다. 그러한 정보는 한국어 사전에서 구할 수 있음을 앞에서도 지적한 바 있다. (25)의 SUBCAT 정보는 그와 같은 과정을 거쳐 얻은 것이다. 이 정보가 확보됨으로써 다음과 같은 비문을 차단할 수 있게 된다.

(26) a. *존이 정직하기를 생각한다.
b. *존이 정직하다고 원한다.

다음에는 two books와 같은 명사구의 구조에 관심을 돌려 변환사전과 개별 언어 사전에 이러한 명사구의 번역 과정을 위하여 어떠한 정보를 상세화해야 할 것인지 논의하기로 한다. Two books를 '두 권의 책'으로 번역하고 흔히 분류어(classifier)라고 불리는 '권'과 같은 표현의 처리 문제에 주목하기로 한다.

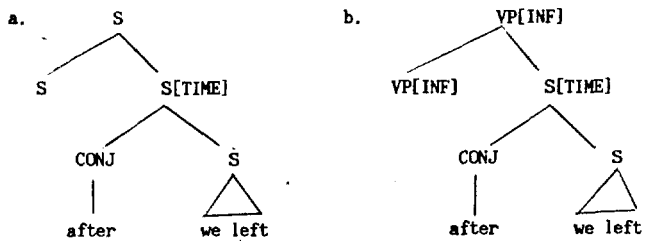
수사와 같이 나타나는 이 분류어는 사물의 성질에 따라 달라진다. 그런데 이러한 분류어는 영어에는 특별한 경우를 제외하고는 안 나타나는 경우가 많으므로 변환사전에서는 분류어에 관한 정보를 충분히 얻을 수 없다. (전혀 얻을 수 없는 경우도 많다.) 변환사전에서는 대부분의 경우에 있어서 단순히 영어의 영(zero) 형태와 한국어 분류어 사이의 대비가 있을 뿐이다.

(27) a. (영) zero - (한) 통 (편지, 서류 등)
b. (영) zero - (한) 개 (사고, 공동)
c. (영) zero - (한) 장 (종이, 담요 등)
d. (영) zero - (한) 사람 (아이, 여자 등)
e. etc.

이러한 이유로 분류어에 관한 정보는 전적으로 한국어 사전에서 얻는 것으로 하는 것이 개별 명사 경제적이라고 생각된다. 즉 개별 명사의 엔트리에 그것이 취하는 분류어가 무엇인지에 관한 정보를 상세화해야 한다. 가령 '편지'와 같은 명사가 수사와 같이 쓰이면 반드시 분류어 '통'을 택하도록 하는 장치를 마련해 두면 될 것이다.

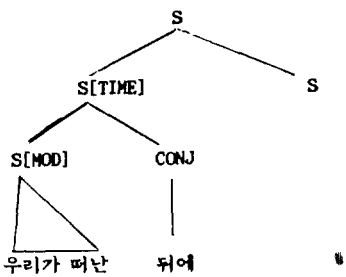
끝으로 (3)과 (4)에 나타나는 시간 부가어를 생각해 보자. (after we left-'우리가 떠난 뒤에') 이 시간 수식절 after we left는 (3)에서 문장 수식어로 해석될 수도 있고 VP 수식어로 해석될 수도 있다. 문장 부사로 해석되면 설득의 행위가 일어나 때를 가리키게 된다. 문법규칙 3에 의하여 문장 수식어의 경우는 아래 (28a)로, VP 수식어의 경우는 (28b)로 분석된다.

(28)



한국어 문장 (4)에 있어서는, 영어 접속사 after를 한국어 접속사 '뒤에'와 직접 대비되는 것으로 가정할 때, 이 접속사가 문장 보어를 요구하는 데 그것의 VFORM이 MODIFIER인 것으로 분석할 수 있다고 본다. ([VFORM] 자질로 말미암아 '떠난'의 '-ㄴ' 수식형이 이루어진다.) 이 분석을 아래 수형도로 나타낼 수 있다.

(29)



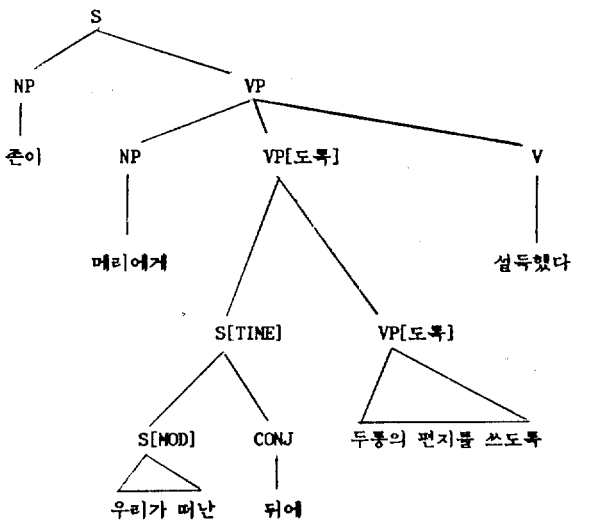
그런데 한국어 문장 (4)는 앞에 지적한 영어의 경우와는 달리 중의성이 없다. (4)의 시간 수식절은 문장 수식어로 (즉, 영어의 (28a)와 같이) 해석

될 뿐이다. 그러니까 (4)는 영어 문장 (3)의 두 가지 번역 중 하나에 해당할 뿐인 셈이다. (3)의 after we left가 VP 수식어로 해석될 경우 (즉 (28b)의 분석)를 번역하면 아래 문장 (30)가 될 것이다.

(30) 존이 메리에게 우리가 떠난 뒤에 두 통의 편지를 빌에게 쓰도록 설득했다.

(4)에서는 '우리가 떠난 뒤에'가 문두에 왔으나 (30)에서는 그것의 위치가 '메리에게 두 통의 편지를 쓰도록'이라는 VP 내부가 되었다. 이와 같이 부사절 성분의 위치의 차이에 따라 해석이 달라지게 된 것이다. (30)의 구조는 다음과 같이 되겠다.

(31)



그리하여 중의적인 영어 문장 (3)의 한 가지 의미는 한국어 문장 (4)로 번역되고 또 하나는 (30)으로 번역되어야 하는데 이것을 보장하는 방법이 무엇인가가 문제이다. 우리는 여기에서 기계번역에 있어 가장 어려운 문제로 알려져 있는 중의성의 문제에 봉착하였는데 우리가 여기에서 이 난제를 해결할 새로운 방법을 제시할 준비가 되어 있는 것은 결코 아니다. 다만 현재로서는 이러한 중의성의 문제가 기계번역과정에 인간의 개입이 불가피한 부분이 있다는 점을 상기시켜 준다는 점을 강조할 뿐이다. 이러한 의미에서 지금 단계에 우리가 제시하는 MT 체계는 "인간 보조의 MT(human-aided MT)"에서 벗어날 수 없는 것이라고 말할 수 있다.

4. 어휘부의 모형

영어 사전(E-Lexicon)의 어휘항목은 HPSG의 어휘이론에 따라 고유정보(specific information)와 유형정보(type information)로 구성된다.

고유정보에 명시해야 할 사항은 대강 다음과 같다.

(32) SPECIFIC
- (i) SUBCATEGORIZATION
- (ii) MORPHOLOGY
- (iii) ADJUNCTS
- (iv) SENSE
- (v) SEMANTIC PECULIAR

통사정보로서는 하위범주화(SUBCAT)로서 충분하고, 형태적 정보에는 형태적 특수사항을 포함한다. 그리고 수식어(ADJTNCTS)에 관한 특수성이 있을 경우에 이를 기록해야 한다. 의미정보로서는 Katz와 Fordor식의 semantic distinguisher에 해당하는 어휘풀이(유의어)가 있어야 할 것이고 영한대조에 활용할 수 있는 의미적 특수성을 또한 기록해야 할 것이다. Semantic peculiar란에는 예컨대 wear (또는 put on)와 같은 영어어휘가 그 직접목적어의 성질에 따라 "입다, 쓰다, 신다, 매다" 등으로 달리 번역되는 경우를 처리하기 위한 정보를 상세히 제공해 주어야 한다.

유형정보에 포함할 사항은 앞 절에서 설명한 HPSG의 방식과 동일하다. 즉 유형 포섭 도표에서 마지막 절점에 해당하는 유형만을 기록해 주면 된다.

앞에서 언급한 바와 같이 한국어 사전은 문장의 종합, 생성에 사용되는 사전이다. 잘 알려진 바와 같이 한국어의 한가지 중요한 특징은 형태소의 연결방식의 다양성이다. 특히 동사와 형용사의 어근에 여러 가지 어미가 첨가되어 단어가 형성되는 과정이 매우 다양하다. 이러한 어휘형성 과정을 어휘 규칙으로 형식화하여 한국어 사전(K-Lexicon)에 포함시켜야 한다. 따라서 이 방면에 있어서는 어휘부의 수평적 잉여성을 처리하기 위한 HPSG의 방법을 그대로 적용할 수 있다고 보여진다.

그렇다면 한국어 사전의 편찬에 있어서 다루어야 할 중요한 문제 중의 하나는 영어 사전의 경우와는 달리 어휘규칙을 여하히 수립할 것인가 하는 것이 되겠으며, 고유정보의 표시 방법이나 어휘유형의 표기방법 등은 영어의 경우와 다를 바가 없을 것이다.

어휘규칙이 하는 일은 어떠한 어간에 어떠한 어미가 첨가되며, 어간과 어

미 그리고 어미 사이의 순서가 어떻게 되는가를 규정하는 것이다. 몇 가지만 예를 들어 살펴 보기로 한다.

동사어근에는 전형적으로 시제어미가 첨가되고 그 다음에 종결 어미가 온다.

(33) a. 잡다/보다/먹다/쓰다

b. 잡았다/보았다/먹었다/썼다
c. *잡다았/*보다았/*먹다었/*쓰다었
d. *다잡았/*다보았/*다먹었/*다쓰었

위의 예는 동사 어근에 과거어미 '-았/었-'과 서술종결어미 '-다'가 첨가되는 현상을 보여준다.

먼저 과거어미가 첨가되는 어휘규칙을 아래와 같이 설정할 수 있을 것이다.

(34) 과거어미규칙

$$_{\text{base}}\begin{bmatrix}\text{PHON } [1]\\ \text{PAST } [2]\\ \text{SYN } \ \text{PAST}\end{bmatrix} \longrightarrow {}_{\text{past}}\begin{bmatrix}\text{PHON } f_{\text{PAST}}([1],\ [2])\\ \vdots\end{bmatrix}$$

함수 f_{PAST}의 내용은 어간(base form)에 '-았/었-'을 첨가하는 것이다. 그리고 동사 어근의 음성형태의 따라, 즉 그것의 마지막 음절의 모음이 'ㅏ' 또는 'ㅗ'이면 '-았-'이 붙고 그 외의 경우에는 '-었-'이 붙도록 한다. 그리고 이 형태소의 위치는 항상 어간의 직후가 된다. 과거시제어미 앞에 올 수 있는 어미로서 존칭어미 '-시/으시'가 있는데, 이 어미의 위치는 반드시 동사 어근 직후이다. 그러므로 과거어미는 존칭어미가 동사 어근에 첨가된 상태를 어간으로 간주하여 거기에 첨가되는 것으로 보아야 한다. 따라서 우리는 다음과 같은 존칭어미를 어간에 포함시키는 어휘규칙 필요하다.

(35) 존칭어미규칙

$$_{\text{base}}\begin{bmatrix}\text{PHON } [1]\\ \text{SYN } \ \text{HON}\end{bmatrix} \longrightarrow {}_{\text{base}}\begin{bmatrix}\text{PHON } f_{\text{PAST}}\ ([1],\ [2])\\ \vdots\end{bmatrix}$$

이렇게 되면 동사 어근에 존칭 어미가 붙어서 만들어진 형태가 과거어미 규칙에 대하여 어간으로 간주되어서 가령 '잡'에 '-으시'가 붙어 '잡으시'라는 어간이 되고 여기에 '-었-'이 붙으면 '잡으시었'을 얻게 된다.

서술 종결어미의 특징은 그 위치가 반드시 최종위치라는 점이다. 이는 물론 의문, 명령, 청유 등 모든 종결어미의 공통된 특징이다. 이 점을 내용으로 하는 f_{DEC} 함수를 설정하고 다음 규칙을 수립할 수 있다.

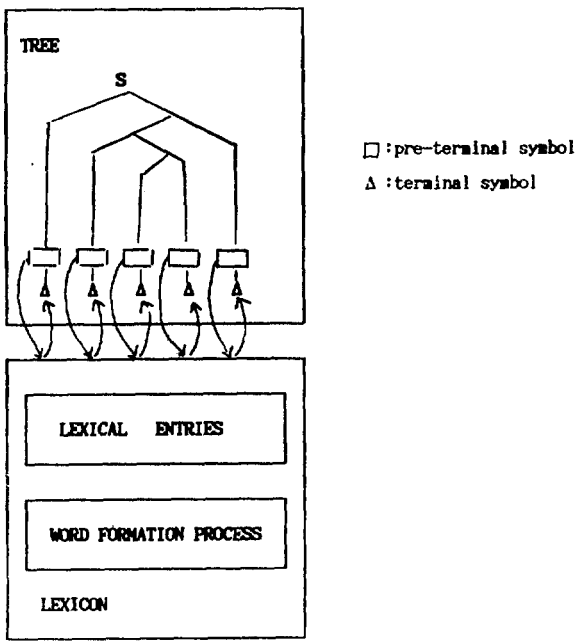
$$(36)\quad \underset{\text{tense}}{\begin{bmatrix}\text{PHON}\ [1]\\ \text{SYN}\quad \text{DECL}\end{bmatrix}} \longrightarrow \underset{\text{declaration}}{\begin{bmatrix}\text{PHON}\ f_{\text{PAST}}\ ([1],\ [2])\\ \vdots\end{bmatrix}}$$

불규칙동사를 처리하는 방법은 원칙적으로 HPSG의 방법을 따른다. 규칙동사의 어휘 항목에는 PAST 형태소의 속성가 (즉 (34)의 [2])가 주어지지 않고 불규칙동사의 어휘 항목에는 그것이 주어진다. 그러면 f_{PAST}는 [2]의 값이 명시되어 있으면 그것을 그대로 취하고 그것이 주어져 있지 않은 경우에는 규칙 (34)에 의하여 '-었/았-'을 첨가하게 한다. 가령 '굽다'의 어휘 항목을 다음과 같이 정해 두면 될 것이다.

$$(37)\quad \begin{bmatrix}\text{PHON}\quad \text{굽}\\ \text{PAST}\quad \text{구웠}\\ \vdots\end{bmatrix}$$

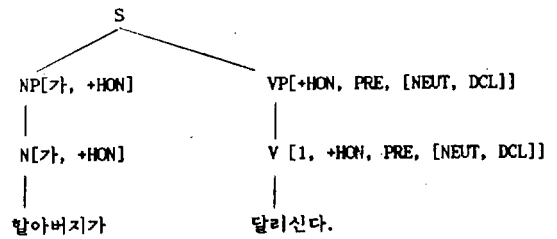
수형도와 어휘부의 관계는 다음 도표 (38)과 같이 나타낼 수 있다.

(38)



예비 종점 기호(pre-terminal symbols)는 모두 어휘 범주들이다. 여기에는 통사 구조의 성립 과정에서 부여된 각종 통사(또는 의미) 정보가 자질로 표현되어 포함되어 있다. 종점 기호(terminal symbol)는 마지막으로 실제의 어휘가 삽입될 곳이다. 이리하여 각 예비 종점 기호는 자신이 가진 통사 정보와 함께 어휘부에 들어가서 어휘 목록(lexical entries)에서 해당어휘를 선택하고 어휘 형성 과정(word formation process)에서 통사 정보가 지시하는 바에 따라 형태 규칙의 적용을 받아서 어휘의 형태가 결정된다. 그후 수형도로 돌아가서 종점 기호에 삽입된다. 이러한 과정을 아래 (39)의 예로써 (40)에서 보이는 것과 같은 과정으로 설명할 수 있다.

(39)



(40) "할아버지가"의 어휘삽입과정 :

(i) N[[+HON, [CASE NOM]]]이 수형도의 예비 종점 기호가 된다.
(ii) 이것이 어휘부로 들어간다.
(iii) N[+HON]에 맞는 "할아버지"를 어휘 목록에서 선택한다.
(iv) [CASE NOM] 정보가 있으므로 어휘 형성 과정으로 들어가 형태 규칙의 적용을 받는다.
(v) 형태 규칙 N[CASE NOM]->N+"-이/가"에 따라 "할아버지"에 "가"를 첨가하여 "할아버지가"를 얻는다.
(vi) 수형도로 돌아와서 종점 기호 △에 "할아버지가"를 삽입한다.

(41) "달리신다"의 어휘 삽입 과정 :

(i) V[1, HON, PRE, [NEUT, DCL]]이 어휘부에 들어간다.
(ii) V[1]에 맞는 "달리"를 선택한다.
(iii) [HON]과 [VFORM] 자질이 있으므로 어휘 형성 과정으로 들어간다.
(iv) 세 개의 형태 규칙에 의하여 "시", "-ㄴ", "-다" 등 세 개의 접미사가 순서대로 첨가되어 "달리신다"를 얻는다.
(v) 수형도로 돌아와서 종점 기호 △에 "달리신다"를 삽입한다.

이와 같이 어휘부가 효율적으로 이용되기 위해서는 어휘 형성과정에서의 형태 규칙과 어휘 규칙 등을 상세하고도 완벽하게 수립하는 것이 중요하다.

우리는 이들에 관하여서는 위에서 몇 가지 실례를 들어서 설명하였다.

5. 결 론

기계번역 시스템의 일부가 되는 사전을 구성하는 데 있어서 GPSG와 HPSG 이론의 어휘부 모형을 이용하였는데 크게 보아 세 가지 장점이 부각되었다고 생각된다.

첫째, 자질 이론을 최대한 활용하여 어휘가 가지는 모든 정보를 속성-속성가 행렬(attribute-value matrix)로 표현하고 이를 토대로 자질들의 영한 대응관계를 적절히 정의함으로써 번역 과정을 효과적으로 처리할 수 있게 해 준다.

둘째, HPSG의 유형(type)의 개념은 사전에 포함될 엄청나게 많은 양의 정보를 조직적으로, 잉여성과 반복성을 최대한 줄이는 방향으로 사전에 담을 수 있게 해 준다.

셋째, 문장의 구조는 그 문장을 이루는 성분들의 어휘적 속성의 발현 양식에 의하여 결정되는 것이라고 주장하는 극단적 어휘주의(Radical Lexicalism)의 이상을 실현시킬 가능성을 제시해 준다.

그러나 우리는 이 글에서 기계번역을 위한 사전 편찬의 원리와 원칙의 윤곽을 세우기 위한 극히 기초적인 고려 사항들을 논의하였을 뿐이다. 구체적이고 실제적인 여러 중요한 사항들은 앞으로 계속 더 연구하여 보완해야 할 우리의 과제이다.

〈경 희 대〉

References

박병수(1988). '영한기계 번역을 위한 일반 구구조 규칙의 수립,' 언어 13.1, 한국언어학회.

_____(1988). '영한기계 번역을 위한 형태 통사론적 토대의 연구. GPSG 모델의 개발.「자동번역 시스템개발기술에 관한 연구」, 한국과학기술원시스템 공학센터.

_____(1989). '통합문법에 있어서의 어휘부의 모형 : 영한 기계번역의 경우',「어학연구」25.1, Language Research Institute, Seoul National University, Seoul.

_____(1989). '기계 번역에서 본 한국어의 특징',「정보과학학회지」제 7권 제 6호. pp. 31-39.

장석진(1989). '자연언어처리를 위한 통합 문법 : 하위 범주화와 어휘적 잉여성', 이혜숙 교수 정년 기념 논문집.

Ahn, S.-C. (1985). *The Interplay of Phonology and Morphology in Korean*, Doctoral dissertation, University of Illinois at Urbana-Champaign.

Bar-Hillel, Y. (1958). *A demonstratton of the non-feasibility of Fully Automatic*

High Quality Translation. Language and Information, Reading, Mass.: Addion-Wesley.

Barwise, Jon. (1989). *The Situation in Logic.* CSLI Lecture Notes No. 17, CSLI, Stanford University.

Chomsky, N. (1980a). *Rules and Representations.* New York: Columbia University.

_____(1980b). *Lectures on Government and Binding.* Dordrecht: Foris Publications.

Gazdar, Gerald, Ewan Klein, Geoffrey Pullum, and Ivan Sag(1985). *Generalized Phrase Structure Grammar*, Basil Blackwell, Oxford.

Gunji, T. (1987). *Japanese Phrase Structure Grammar: A Unification-Based Approach.* Dordrecht: D. Reidel.

Karttunen, Lauri(1985). 'Radical Lexicalism,' in Shieber and Wasow(1987), *The Processing of Linguistic Structure*, Course material for the 1987 Linguistic Institute of the Linguistic Society of America.

Lehrberger, John and Laurent Bourbeau(1988). *Machine Translation: Linguistci Characteristics of MT Systerms and General Methoology of Evaluation*, John Benjamins Publishing Co., Amsterdam.

Park, B.-S. (1985). Some control agreement problems in Korean: a GPSG analysis of honorific expressions. in *Proceeding of '84 Matsuyama Workshop on Formal Grammar*, ed. by S. Kubo(1985)

_____(1988). Sentential' predicates in Generalized Phrase Structure Grammar: an analysis of Korean double subject constructions. Korean Linguistics, Vol. 5, 59~74 International Circle of Korean Linguistics.

_____(1989). 'How to Handle Korean Verbal Suffixes: a Unification Grammar Approach,' *Korean Linguistics* 6, International Circle of Korean Linguistics.

_____(1989). 'Development of a Unification Grammar lexicon for English-Korean machine translation.' *Language Research* 25, SNU. pp. 11-29.

_____(1989). *Construction of a Lexicon-Driven Grammar Model and Its Application for English-Korean Machine Translation.* (In Korean) Report for Systems Engineering Research Institute, KAIST. (Co-work with K. Lee and H.B. Im).

Pereira, F. and S. Shieber(1987). *PROLOG and Natural-Language Analysis*, CSLI Lecture Notes No. 10, CSLI, Stanford University.

Peters, S. and R. Ritchie(1973). On the Generative Power of Transformational Grammars. *Information Sciences* 6, 49~83.

Pollard, Carl and Ivan A. Sag(1987). *Information-Based Syntax and Semantics, Volume I; Fundamentals*, CSLI, Stanford University.

_____(1988). 'An Information-Based Theory of Agreement,' *CSLI Report No.* CSLI-88-132, Stanford University.

_____(1990). *Information-Based-Syntax and Semantics, Volume II: Topics in Control, Bining, and Agreement*, ms. Stanford University.

Reyle, U. and C. Rohrer, ed. (1988). *Natural Language Parsing and Linguistic Theories*, Studies in Linguistics and Philosophy vol. 35, D. Reidel Publishing Co., Dordrecht.

Sells, P. (1985). *Lectures on Contemporary Syntactic Theories: An Introduction to Government-Binding Theory, Generalized Phrase Structure Grammar, and*

Lexical-Functional Grammar. CSLI, Stanford U.
Slocum, Jonathan (1985). *Machine Translation Systems*, Cambridge UP.
_____ (1985). 'A Survey of Machine Translation: Its History, Current Status, and Future Prospects,' in Slocum (1985).
Shieber, Stuart M. (1986). *An Intoduction to Unification-Based Approches to Grammar*, CSLI, Stanford University.
_____ (1987). Separating Linguistic Analysis from Linguistic Theories, in Whitelock et al, (1987).
Whitelock, P., M.M. Wood, H.L. Somers, R. Johnson, and P. Bennett, eds. (1987). *Linguistic Theory and Computer Applications*, Academic Press, London.
Wilks, Y.A. (1983). Machine translation and the artificial paradigm of language. In *Computers in Language Research*, 2, 61~105, eds. by W.A. Sedelow, Jr. and S.Y. Sedelow.

On the model of the lexicon for English-Korean machine translation

Byung-Soo Park
Sang-Cheol Ahn
Beom-mo Kang

Abstract

The purpose of this paper is to propose several principles for constructing a model of the lexicon for English-Korean machine translation. For this purpose, we employ the frameworks of Generalized Phrase Structure Grammar (GPSG) and Head-driven Phrase Structure Grammar (HPSG). The machine-translation system we are assuming in this paper is a direct-transfer, type-driven, and bidirectional one which is dependent on superlexicalism. In order to justify the validity of our proposal, we discuss the characteristics of our translation system and the transfer dictionary, as well as the procedure of translation.

More specifically, it will be shown that our model contributes to machine-translation in several ways. First, it makes the translation procedure more effective as it represents all of the lexical information as attribute-value matrices. Second, the notion of 'type' enables us to arrange a substantial amount of lexical information systematically

by **minim**izing redundancy. Third, it shows various advantages of superlexicalism as the structure of a sentence is determined by lexical properties in this framework.