어휘기능문법(Lexical-Functional Grammar)에 근거한 한-영 양방향 기계 번역기의 언어학적 구성

김정렬*
한국교원대

Jeong-ryeol Kim. 1999. Linguistic design of a bidirectional Korean-English machine translation system based on Lexical-Functional Grammar. *Language and Information 3.1*, 65–82. The interests in Machine Translation (MT) have gotten revitalized lately with the rapid expansion of internet users. MT technology has gone through several different stages of development, but the longest surviving methods usually maintains the following characteristics: the expandability and flexibility based on proved linguistic formalism, the transfer method of translation, the continued efforts of systematic updates being made into the system. This paper introduces one such system, L&H Korean-English bidirectional MT system. This system uses Lexical-Functional Grammar as its linguistic framework. It also adopts the transfer method of MT and has been around on the market for over 10 years for other language pairs. Currently, the system covers over 10 different languages including Chinese, Japanese and Arabic, in addition to European languages. This paper will review the system in its core and discuss related tools and resources being used to enhance the quality of translation. **(Korea National University of Education)**

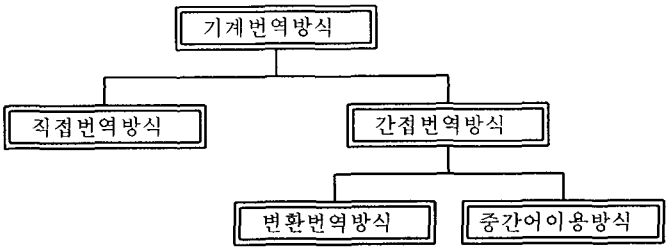
1. 머리말

컴퓨터를 이용한 인간 언어의 번역은 하드웨어가 점점 고속화되고 소프트웨어 기술이 발달하면서 과거 어느 때 보다도 실현 가능성이 높아졌다. 최근 인터넷을 이용하여 각종 외국어로 된 정보의 양이 엄청나게 불어나면서 기계 번역의 필요성은 점차로 고조되고 있고 이러한 필요성은 기계 번역기 시장을 국내에서도 서서히 만들어 가고 있다. 컴퓨터는 인간이 일상적으로 의사소통 하는 방식에 가장 가까울 때 일반인들이 사용하기 용이해진다. 다시말해, 컴퓨터의 보급과 사용은 인간 언어를 컴퓨터가 얼마나 잘 이해할 수 있느냐 즉 얼마나 컴퓨터와 인간이 쉽게 의사소통을 할 수 있느냐 하는 데 달려있다. 컴퓨터가 인간의 음성 언어를 얼마나 잘 인지할 수 있느냐를 연구하는 음성 인식, 인간이 알아들을 수 있는 언어를 생성해 내는 음성 생성과 더불어서 인간이 쓰는 언어를 이해하고 인간의 욕구대로 다른 언어로 번역해 주는 기계 번역 기술은 앞으로 컴퓨터의 지능을 가늠할 중요한 요소들 중의 하나이다.

이 글에서는 음성인식, 음성생성 및 기계번역 중에서 기계번역을 중심으로 논의를 하겠다. 기계번역 방법은 전통적으로 직접 번역 방법과 간접 번역 방법으로 나누어졌고, 간접 번역 방법은 다시 변환 번역 방법 및 중간어 이용 번역 방법으로 66페이지의 그림과 같이 대별해 볼 수 있다.

직접 번역 방법은 ALPAC 보고서 후 70년대 초에 나온 기계 번역기인 Systran이나 Georgetown의 GAT 등을 들 수 있고 국내에서는 구조적으로 유사한 한국어-일본어의 번

* 363-791 충북 청원군 강내면 다락리 한국교원대학교 영어교육과, E-mail : jrkim@cc.knue.ac.kr



[그림 1] 기계 번역 방식

역에서 채택되어서 사용되고 있다. 이 방식은 어휘를 바꾸고 형태소 분석 및 생성을 통해서 번역하는 방법이다. 이에 반해서 간접 번역 방식의 하나인 중간어 이용 번역 방법은 다국어 번역 방식에서 이들 언어들의 중간 언어를 설계하고 그 문법을 정리하여 모든 언어를 중간어까지 번역하고 생성해 놓으면 그 다음은 자유롭게 어느 방향이든지 번역해 갈 수 있다는 것이다. 그러나, 이 방식은 어디까지나 이론적으로 가능한 것이지 아직까지 실제로 검증 되어서 나온 실험실 밖의 시스템은 없는 실정이다. 따라서 현실적으로 한국어-영어와 같이 구조적으로 차이가 많이 나는 언어의 경우는 변환 번역 방법을 주로 쓰고 있다. 뒤에서 자세히 설명하겠지만, 분석이 끝난 입력문에서 목표어의 출력문 구조로 변환해 주는 과정을 분석 및 생성 외에 따로 두고 시스템을 설계하는 방법이다.

이 글에서 중점적으로 논의될 L&H 시스템도 마찬가지로 변환 번역 방법을 쓰고 있다. 변환시에 구조적인 차이의 해결 외에도 의미의 중의성을 해결할 수 있도록 구조적 맥락에 의존한 해석 등의 다양한 주변 자원들을 제공하고 있다. 그리고 본 연구의 중심 논의 사항이 되는 어휘-기능 문법(Kaplan & Bresnan 1982)을 이용해서 시스템을 구현하였기 때문에 언어학적으로 아주 확장성이 뛰어난 번역 시스템의 모델을 제시하고 있다. 어휘-기능 문법은 구성 성분의 전후 관계와 수직 관계를 나타내는 구성 성분 구조 (c-structure)와 이들 구성 성분간의 문법 기능을 밝혀 주는 기능 구조 (f-structure)로 나뉜다. 언어들 간에 구성 성분 구조는 아주 다를 수 있지만 기능 구조는 구성 성분 구조에 비해보면 아주 유사하기 때문에, 어휘-기능 문법을 변환 번역 방식의 문법 기반으로 사용할 때 그 만큼 구조적인 차이를 해결해야 하는 부담이 상대적으로 줄어들게 된다.

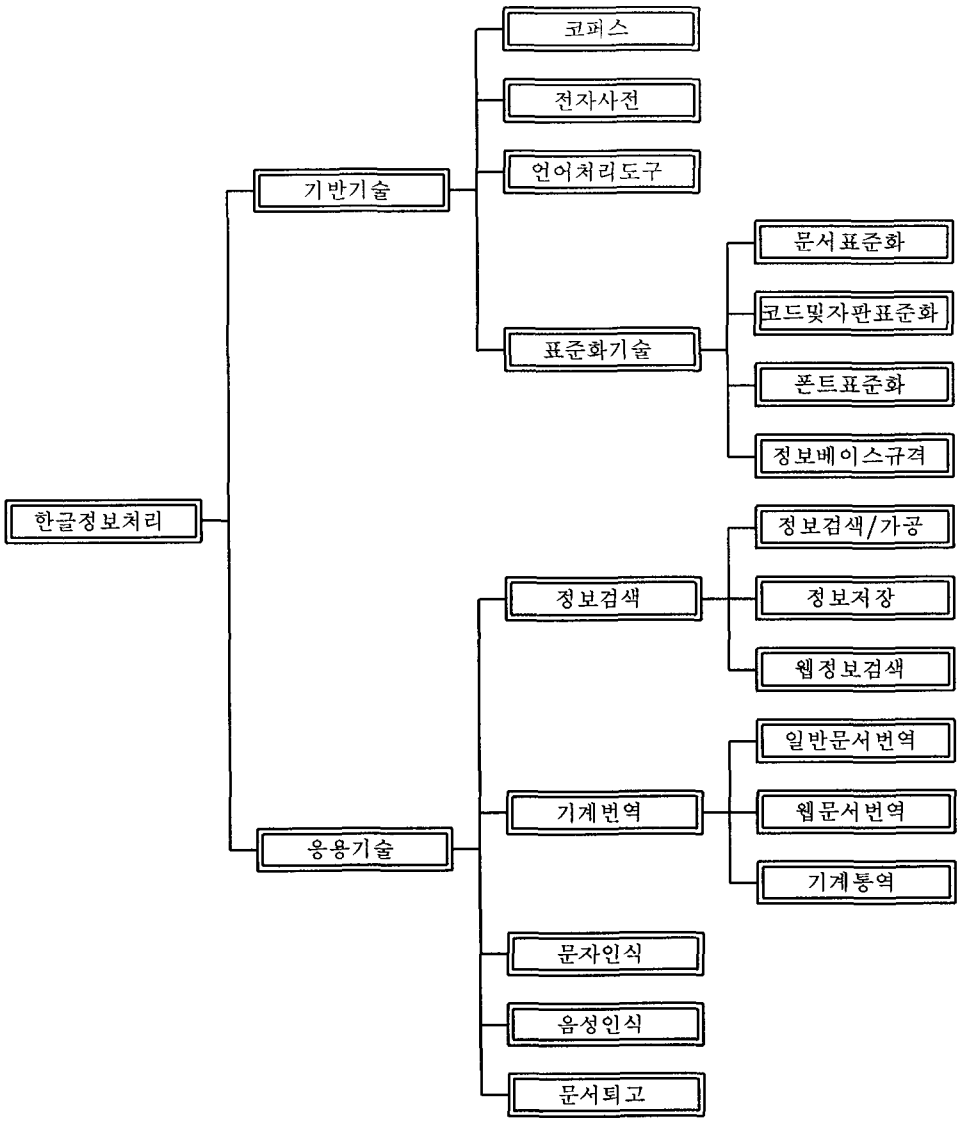
그리고, 기계번역의 발달을 살펴보면 크게 규칙기반 기계번역과 말뭉치 기반 기계번역으로 나누어 볼 수 있다 (최승권 (1999)). 규칙기반 기계번역은 통합문법 (Unification Grammar)인 어휘-기능 문법, HPSG 등의 통사 이론을 바탕으로 언어학적 정보를 이용한 구조 중심의 번역을 말한다. 이에 반해, 말뭉치 기반 기계번역은 통계 및 자료에 바탕을 둔 방법으로 대역 말뭉치를 사용한 예제를 중심으로 하여 번역하는 방법이다.

본 논문의 목적은 어휘-기능 문법에 기반을 둔 변환 방식 기계 번역 시스템의 하부 구성 모듈들을 설명하고, 출발어의 입력문에서 목표어의 출력문으로 내보내기까지 어떠한 과정을 거쳐서 번역이 이루어지는지 알아보는 것이다. 본 논문의 논의 순서는 우선 기계번역 시스템과 자연언어 처리와 관련된 최근 상용 시스템 구성 현황을 국내외적으로 알아보고, L&H 시스템의 하부 모듈들을 번역 순서대로 기술한다. 그리고, 마지막으로 논문의 초점이 되는 어휘-기능 문법을 기반으로 한 기계 번역 과정을 언어학적인 측면에서 살펴보기로 한 다.

2. 한국어 기계 번역의 현황

한국어 정보 처리는 구성 요소 기술의 특성을 기준으로 (시스템 공학 연구소, 1998) 67 페이지에서와 같이 나누고 있다.

67페이지의 [그림 2]에서 보는 것처럼 기계 번역은 한국어 정보처리 중에서 응용 기술에 속하고, 많은 부분 한국어 정보처리 기반 기술의 언어 자료를 이용해야 한다. 우리 나라의 언어 정보 처리 역사도 한글 코-드의 표준화가 70년대 말에 이루어진 이래 거의 20년 가까이 되



[그림 2] 한글정보처리 기술 분류

면서 많은 시행 착오 속에서도 중단 없는 발전을 거듭해 왔다. 그래서, 이제 표준화 된 태그 세트나 자연상태에서 채취된 많은 언어 자료를 담은 코퍼스들을 비교적 손쉽게 구할 수 있게 되었다.

특히 최근 들어 인터넷의 폭발적인 확산과 더불어 각종 외국어에 대한 수요가 늘어나면서 번역기가 얼마나 웹브라우저들과 결합해서 잘 운용되느냐 하는 것에 많은 관심을 가지고 움직이고 있다. 그리고 인터넷 사용자가 단일 언어만을 쓰는 사람이라고 할지라도 여러 가지 다른 외국어로 된 정보를 검색할 수 있는 다국어 정보 검색 시스템이 등장하여 야후 등에서 상용화되고 있다. 더불어 자동통역에 대한 관심이 높아지고 있는 데, 이는 우선 문자언어의 번역 기술 위에 음성 인식 기술과 음성 생성 기술이 결합되어야 가능한 것이다. 음성 언어에서 음성 언어로 번역하는 자동통역 기술은 L&H를 비롯해서 미국, 일본 등지에서 활발하게 연구되고 있다. 우리나라에서도 한국통신과 전자통신 연구소가 주축이 되어서 제한된 영역 (호텔예약)에

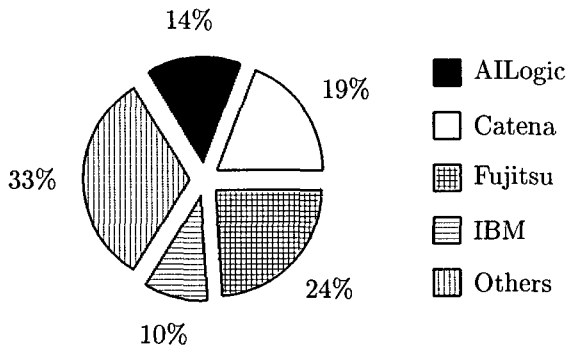
서 시제품을 개발한 것으로 알고 있다.

언어	이름	개발 기관	설명		
			대상	성능	기타
영-한	에서로/EK	SERI	범용문서, 웹문서	100,000단어	윈도우95
	번역마당	리틀컴퓨터	범용 문서	100,000단어	윈도우95
	앙코르	한국 IBM	범용 문서, 웹문서	200,000단어	윈도우95, OS2
	워드체인지 3.1	정소프트	범용 문서, 웹문서	100,000단어, 80,000숙어	윈도우95
	트레니98	언어공학연구소	범용문서, 웹문서	200,000단어	윈도우 95/98
	세종대왕 3.0	미래소프트웨어	범용 문서, 웹문서	실시간 번역	-
	L&H 영한 번역기	L&H Korea	범용 문서, 웹문서	실시간 번역	윈도우 95/98, UNIX
	E-Tran	서울대학교 자연언어처리실	범용 문서, 웹문서	실시간 번역	윈도우 95/98
	인가이드	엔엘아이소프트	범용 문서, 웹문서	실시간 번역	윈도우 95/98
	EZ Reader	언어와 컴퓨터	범용 문서, 웹문서	실시간 번역	윈도우 95/98
한-영	L&H 한영 번역기	L&H Korea	범용 문서, 웹문서	실시간 번역	윈도우 95/98
일-한	에서로/JK	SERI	범용 문서, 웹문서	200,000단어	윈도우95
	마이다스 1.0	이호정보기술	범용 문서, 웹문서	180,000단어	윈도우95
	바벨 2.0	유니소프트	범용 문서, 웹문서	120,000단어	윈도우95
	조선통신사	창신컴퓨터	범용 문서, 웹문서	150,000단어	윈도우95
	J-Seoul/JK	디코	범용 문서, 웹문서	전문가용 사전	윈도우95
한-일	드림 KJ	드림 C&C	범용 문서, 웹문서	실시간 번역	-
	조선통신사	창신컴퓨터	범용 문서, 웹문서	실시간 번역	전자우편 전송용 클라이언트

[표 1] 국내 번역 프로그램

국내에서 개발된 번역기를 살펴보면 우선 인터넷 번역을 위한 일한 번역 시스템으로 유니소프트의 바벨, 디코시스템의 I-Seoul/JK, 이호정보기술의 마이다스, 창신컴퓨터의 조선통신사 등이 있다. 이들은 모두 웹 상에서 실시간으로 일본어 홈페이지를 한국어 홈 페이지로 번역해 주는 기능을 가지고 있다. 영한 번역 시스템으로는 본 논문에서 구체적으로 소개할 L&H가 개발 중인 한국어-영어 양방향 번역기 외에 다니엘텍의 랑프리, IBM의 앙꼬르, 정소프트의 워드체인지, 언어공학연구소의 사이버트랜스 등이 있다. 그리고, 인터넷 번역 기능과 브라우저 기능을 통합한 영한 번역 시스템으로 성운시스템의 세계로96이 있다. 그리고, 최근에 많은 주목을 받고 있는 엘엔아이소프트의 인가이드, 앙코르의 후신이라고 할 수 있는 서울대학교 자연언어처리실의 E-TRAN 98, 언어와 컴퓨터사의 EZ READER등이 개발되어 속속 시장에 출시되었다. 시스템 공학연구소의 정보화 백서에 나온 내용을 좀 더 보강하여 이들을 도표로 소개하면 다음과 같다.

국외의 기계번역은 미국의 경우, Georgetown 대학의 초기 프로젝트로부터 시작되어 나중에 Systran 시스템에 포함된 기계번역(Machine Translation) 시스템으로 개발되었다. 최근들어 사용자 편집 도구, 다른 작업 환경으로의 포팅, 새로운 언어로의 확장성 및 워드프로세서나 출판 시스템들과의 인터페이스를 고려하는 방향으로 확장 발전되어 가고 있다. EC의 경우, EC 가맹국 9개 언어간의 다국어 번역기 개발을 위한 EUROPA, 다국어 정보 유통/통신 서비스 구축을 위한 INFO'2000, 독어, 영어, 일본어 휴대용 자동 통역기 개발을 위한 독일 중심의 Vermobil 프로젝트, 벨기에의 Flanders Language Valley에 위치한 L&H의 PowerTranslator 등이 있다. 그리고, 일본의 경우는 전체 기계 번역기 시장을 소위 Big 4에 해당하는 Fujitsu, Catena, AILogic, IBM이 70% 정도를 점하고 있다. 이들은 모두 영-일 번역 시스템을 상용화했고, AILogic과 같은 소수의 기업은 중국어-일본어까지 상용화시키고 있다.



[그림 3] 일본 기계 번역기의 시장 점유 규모

위에서 논의된 6개의 대표적인 국외 번역시스템을 열거해 보면 아래 [표 2]와 같다.

[표 2] 국외 번역 프로그램

제품	공급자	주요 적용 언어	적용 언어 쌍	작업 환경	번역 속도
Logos	Logos Corporation	독일어, 영어	7	Wang Minis, IBM	30,000 단어/시간
Metal	Siemens Nixdorf	영어, 독일어	5	Symbolics	11,000 단어/시간
Smart	Smart Comm.	영어	6	Sun workstation	3백만 단어/시간
Systran	Groupe Gachot	영어, 러시아어, 프랑스어, 독일어	15	IBM Mainframes	50,000 단어/시간
Power Translator	L&H	영어, 스페인어, 프랑스어, 독일어	7	Windows 95/98	100,000 단어/시간
Tovna	Tovna MTS	영어, 프랑스어, 러시아어	4	Sun workstation	3600 단어/시간

그리고 음성언어 번역은 궁극적으로 대화의 어휘 연속 음성을 화자 독립(speaker inde-

pendent)으로 인식하고 이해해야 하는 기술과, 실시간으로 목표어로 번역하고 다시 목표어의 자연스런 음성으로 생성해내는 기술들이 통합되어야 한다. 외국의 경우는 미국 MIT의 보이저(VOYAGER), 페가수스(PEGASUS), 일본 ATR의 아수라(ASURA) 시스템, 국제 공동연구 그룹인 C-STAR (Consortium for Speech Translation Advanced Research)의 음성언어 번역 시스템, 미국 CMU의 스피치트랜스(SpeechTrans) 시스템과 야누스(JANUS) 시스템, 그리고 독일의 버브모빌(Verbmobil) 시스템 개발 그리고 L&H의 VoiceXpress와 Real Speak 등이 있다.

국내 현황은 한국통신(KT)과 일본의 국제전신전화(KDD), 그리고 ETRI에서 호텔 예약과 교환 서비스에 관한 영역에서 한일간의 국제간 데모(낭독체 음성언어 번역)를 1995년 5월에 실시했고, 1996년 6월에 ETRI에서 C-STAR의 일원으로서 여행 계획 영역에서의 대화체 음성언어 번역의 시연이 있었다. 현재 멀티미디어 환경하에서의 음성 번역 통신 기술 개발에 관한 연구를 수행중에 있으며 1998년까지 3000 단어에 기반한 호텔 예약 영역에 대한 통역을 목표로 하고 있다.

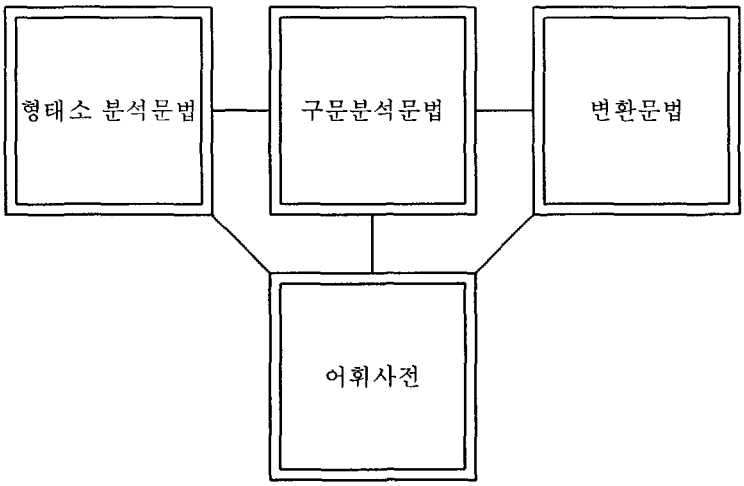
[표 3] 자동통역 관련 국내외 연구 동향

수행기관	국가	특징
ATR-ITL	일본	국제회의 참가 신청을 위한 회화 음성 자동 통역 예제 기반 번역 방식을 이용
NEC	일본	연주회 예약 통역 시스템 INTERTALKER 개발 500단어 처리
CMU	미국	대화번역(JANUS 프로젝트), 독일어의 통합언어 모델(Vermobil TP2) 연구 Interactive System Lab에서 자동통역 연구 수행
SRI	미국	화자 독립의 연속 음성 인식 기술 DECIPHER 보유 대화 기계번역, 화자 인식, 음성 데이터베이스 등의 STAR 프로젝트 수행
MIT	미국	VOYAGER 시스템(영어, 일본어, 이태리어 대상) 개발 Lincoln Labs에서 자동통역 연구 수행
AT&T	미국	영어와 스페인어간의 VEST 번역 시스템 개발
LIMSI	프랑스	대용량의 다국어 연속 음성인식, 언어 종류 인식 연구
DFKI	독일	자연 발화된 대화의 번역 시스템 Vermobil 개발 독일어와 일본어간의 자동통역 시스템 개발
L&H	벨기에	화자독립 음성 인식기 VoiceXpress 개발 자연언어에 가까운 음성 생성기 Real Speak 개발 기존의 번역기에 이들을 통합 장착하여 자동통역 시스템 개발
British Telecom	영국	호텔 예약에 관한 회화번역 (영어, 불어 대상)
KT	한국	호텔 예약과 교환 서비스에 관한 영역에서 한일간의 낭독체 음성언어 번역 시스템 개발
ETRI	한국	멀티미디어 환경하에서의 음성 번역 통신 기술 개발

3. L&H 시스템의 개괄

어휘-기능 문법에 기반을 둔 L&H 한국어-영어 양방향 기계 번역 시스템은 크게 세 가지 작업으로 구분해 볼 수 있다. 첫째는 입력문 분석 규칙의 구성과 변환 문법의 구성으로 한국어나 영어의 분석 입력문 중에서 해결할 수 있는 문장의 종류를 결정하게 된다. 둘째는 형태소 규칙의 구성인데 영어의 경우는 술어나 명사의 활용이 비교적 간단하지만 한국어의 경우 형태소 규칙의 구성 및 처리는 분석 규칙의 구성 중에 중요한 비중을 차지하고 있다. 셋째는 사전의 구성

으로 크게 단일어 사전과 이중 언어 사전으로 나누어 볼 수 있다. 단일어 사전은 입력 어휘 하나 하나에 대해서 형태소 분석 및 구문 분석 규칙이 필요로 하는 정보를 담고 있다. 그리고, 이중 언어 사전은 출발어와 목표어의 변환에 관한 정보를 담고 있다.



[그림 4] L&H 기계 번역 시스템의 언어영역 구성도

문법규칙의 기술은 어휘-기능 문법의 형식과 표기 방식을 모태로 한 LECS라는 비교적 언어학자에게 친숙한 고급어 (HIGH-LEVEL)를 사용한다. LECS는 일종의 인터프리터 방식의 프로그래밍 언어라고 이해하면 되겠다. 즉, 규칙이 LECS로 쓰여지면 컴파일 하는 과정을 거치게 되고, 이때 LECS의 형식에 맞지 않은 표기라든지 이미 선언되지 않은 기호(SYMBOL)의 사용을 막아서 규칙이 무질서하게 양산되는 것을 막는다. 그리고, 사전은 핵심어 사전, 전문용어 사전 및 사용자 사전의 세 개의 파일로 구분되어서 유지된다. 사전의 입력은 개발 과정에서는 전문 어휘론자가 작업을 하기 때문에 효율적인 편집기를 사용해서 하지만, 일반 번역가와 같은 사용자들을 위해서는 사전편집기를 이용해서 쉽게 어휘의 추가가 이루어질 수 있도록 구성되어 있다.

L&H 양방향 기계 번역 시스템의 흐름을 살펴보면 크게 분석 (PARSING) 단계, 변환(TRANSFER) 단계 및 생성(GENERATION) 단계로 구분해 볼 수 있다. 출발어의 입력문이 들어오면 제일 먼저 시스템은 초기화 (INITIALIZATION)를 통해서 파서가 요구하는 모든 기호들을 읽어 들이고, TOP-DOWN TABLE이나 CACHE 영역에 있는 최적화 테이블을 읽는다. 그리고 문제 해결(DEBUG)에 필요한 품사 인식 파일을 읽어 들인다. 초기화가 되고 나면 분절(SEGMENTATION) 과정을 거치게 된다. 분절과정에서는 문장의 시작과 끝을 알리는 여백이나 탭, 문장 부호 등을 인식하여 문장 단위로 입력문을 끊는다. 분절 과정을 거쳐서 나오게 된 결과는 단어별로 하나씩 사전을 찾는 과정을 거친다. 사전을 찾는 순서는 먼저 사용자 사전을 찾고, 다음에 사용자가 지정한 전문 용어 사전을 찾고 마지막으로 핵심 사전을 찾는다.

여기서 사전에 발견되지 않는 어휘가 나왔을 때 일단 KWAN 이라는 엔트리 속에 넣어서 다음과 같이 세 가지 방법으로 해결한다. 첫째, 관용어구의 한 부분인지를 검사한다. 둘째, 형태소 규칙을 적용해야 할 것인지 적용 형태소 규칙들이 나와있는 KWAN 엔트리를 본다. 셋째, 사전에 나와 있지 않는 경우에 문법적인 위치를 추정하여 해당하는 문법 범주로 투사시킨다.

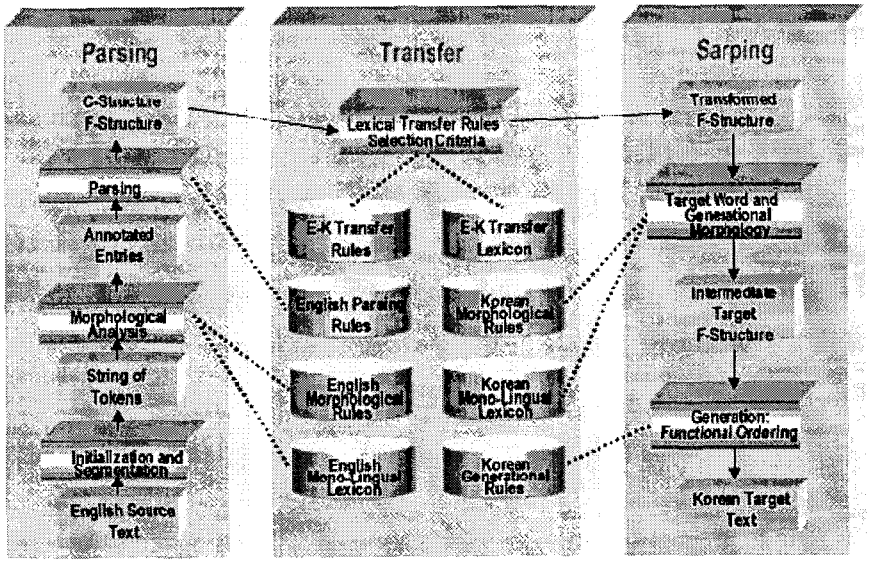
두 번째의 경우는 KWAN 엔트리 속에 있는 형태소 규칙의 입력문이 되어서 형태소 분석 과정을 거치게 된다. 사전에서 찾은 어휘정보는 곧 바로 챠트에 입력된다. 그리고 이때 몇 개의 단어가 관용어를 형성하면 이들을 찾아서 묶는다.

구문 분석 단계에서는 먼저 상향 규칙 (BOTTOM-UP RULES)들이 챠트에 올라온 사전

정보에 따라서 실행된다. 그리고, 실행 가능한 모든 상향 규칙들이 실행되고 나면 조건을 만족시키는 하향 규칙 (TOP-DOWN RULES)들이 적용된다. 물론 규칙이 성공적으로 적용되면 그때마다 새로운 문법 범주가 챠트에 만들어 진다. 이런 과정을 거쳐서 마지막 문법 범주가 만들어지고 나면 그 결과로 어휘-기능 문법의 구성 성분 구조와 기능 구조가 구성 되고, 이어서 기능 구조가 다시 변환규칙의 입력문이 된다. 변환규칙으로 넘어가기 전에, 파서는 수형조건 (TREE CONDITIONS; TCND)을 적용해서 분석을 통해서 얻어진 기능 구조에 합성한다.

변환 단계에서는 우선 출발어의 기능 구조를 수정하여 목표어의 기능 구조로 바꾸는 데, 이때에 데이터 베이스에 있는 변환 규칙과 변환 사전에 따라서 처리한다. 변환 규칙의 적용은 1단계에서부터 6단계까지 진행된다. 그리고, 이때 정확한 번역을 선택하기 위해서 의미 선호도, 사전 영역 선별, 사전의 종류별 찾기 순서와 더불어서 하나의 출발어 단어가 여러 개의 목표어 단어로 번역될 수 있을 때에 이들 각 단어들의 구문적 조건을 지정해 놓은 SMOKEY 등에 있는 정보를 이용한다. 그리고, 번역어 속에 DAG으로 직접 합성되어 들어갈 수 있는 SPARKEY를 사용할 수도 있다. 간단히 말하면, 입력문의 영어 단어를 영한 변환 사전을 이용하여 한국어 정보로 바꾸고 기능 구조상에서 처리되어져야 될 구조적인 차이를 해결한다.

생성 단계에서는 한국어 사전의 정보를 읽어 들이고 형태소 생성 규칙인 SMRULES에 나열된 필요한 형태소 규칙을 거친 다음 GRAFT의 값이 되어 있는 기능 구조상의 문장 기능을 배열하는 배열 규칙에 따라서 최종적인 목표어 문장이 생성되게 된다. 이때, 만약 GRAFT의 값이 주어져 있지 않다면 CONTROL-LABELS에 있는 목표어인 한국어의 기본 어순 규칙에 따라서 배열한다.

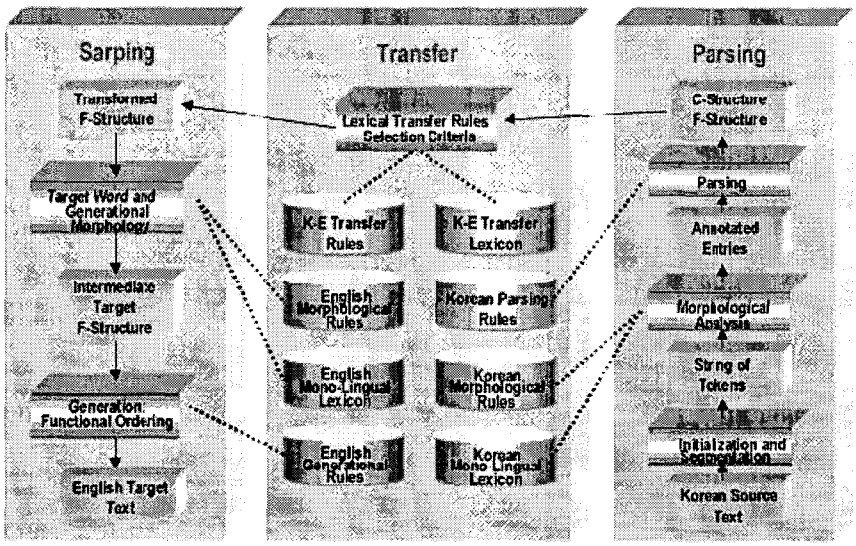


[그림 5] 한국어→영어 기계번역 과정

여기서 한 가지 주목할 것은 L&H 시스템의 생성 단계에서 전통적인 변환 번역 방식에서 해오던 구문 생성 부문을 생략하게 된 것이다. 형태소 생성이 끝난 다음에는 구문 생성을 대신하는 기능 성분 배열 규칙을 적용하여 이들 성분간의 선형 구조를 생성해 내는 것이다. 다시 말해, 목표어의 구성 성분간의 수직적 구조의 생성은 생략한다는 것이다. 이렇게 하므로써 번역에 걸리는 시간의 약 15%를 줄일 수 있게 되었다. 시스템의 구성에 따라서 번역 과정에서 생성부가 차지하는 시간이 여러 가지로 다르기 때문에 일반적화 시키기는 힘들지 만, L&H 시스템의 경우는 단위 문장의 번역에서 생성에 걸리는 시간이 전체의 약 30% 정도였으나 구문 생성을 생략할 수 있게 되면서 생성부의 처리 속도는 두 배정도 증가되었다.

반대로 한국어에서 영어로 가는 기계 번역 과정도 흐름도 유사하지만 언어의 특성상 몇

가지 차이점이 있다. 분절 과정의 경우에 영어와 달리 한국어는 한 글자 속에 두 개의 형태소가 붙어있는 경우가 드물지 않게 나타나기 때문에 한 음절 단위로 된 글자를 알파벳 낱자 하나씩 풀어서 형태소 규칙의 적용을 받을 수 있도록 해야 한다. 분절이 끝나면 형태소 분석 과정을 거치는 데 한국어의 경우는 영어와는 비교가 안될 정도로 많은 형태소 규칙을 포함하고 있어서 몇 개의 종적 단계를 두어서 흔히 나타나는 규칙에서 좀 드물게 나타나는 순서대로 분석한다. 형태소의 분석이 끝나고 나면 구문 분석 과정을 거치게 되고 구문 분석이 끝나면 한국어의 구성 성분 구조와 기능 구조로 나뉘어져 나온다. 다시 기능 구조는 변환 과정에서 한영 사전을 이용하여 어휘교체 및 기능 구조 차이가 해결되게 된다. 그리고, 변환된 기능 구조는 생성 과정으로 넘어가서 형태소 생성 과정과 기능 성분 배열 규칙을 통해서 최종적인 영어 문장이 도출되게 된다.



[그림 6] 영어→한국어 기계번역 과정

4. 어휘-기능 문법에 근거한 기계 번역 과정

어휘-기능 문법과 L&H 기계 번역 시스템이 공유하는 개념 중에 중요한 것이 바로 언어 자료에 근거한 확장성이다. 문법이나 번역 시스템의 확장성은 일반적으로 동일 언어에 대한 수직적 확장성과 다른 언어에 대한 적용성을 고려한 수평적 확장성으로 나누어 볼 수 있다 (Pentheroudakis (1990)). 동일 언어에서 수직적으로 간단한 구조의 언어 자료에서부터 복잡한 언어 자료에 이르기까지 많은 문법적 또는 비문법적 언어 재료를 제대로 설명할 수 있는 문법이나 분석할 수 있는 기계 번역 시스템이 되어야 한다는 것이다. 그리고, 여러 가지 다른 언어들의 분석에 공히 적용될 수 있는 문법 이론이 되어야 하고, 번역 시스템의 틀을 바꾸지 않고 분석 규칙 데이터 베이스만 바꾸어 갈아 끼면 다른 언어의 번역이 가능한 확장성이 있어야 한다. 이를 위한 최선의 방법은 언어학에서 널리 언어의 분석 기술에 사용되어서 수직적/수평적 확장성을 검정 받은 문법 기반을 채택하도록 해야 한다. 그렇지 않고 잉여적으로 매번 시스템을 구성할 때마다 새로운 임의적인 문법을 만들어 쓴다면 이는 아무래도 확장성이 결여된 시스템일 수밖에 없어서 후에 시스템을 고쳐서 개선하기가 무척 까다로운 경우가 많다.

어휘-기능 문법을 L&H 기계 번역 시스템의 언어학적 기반으로 택한 이유는 위에서 언급한 확장성 이외에도 동일한 규칙의 세트를 가지고 분석과 생성을 모두 할 수 있다는 장점 때문이다. 이와 더불어, 위에서도 간단히 언급했지만 구성 성분 구조와 기능 구조의 분리는

정보 처리 과정에 중점을 둔 이론으로서 구성 성분 규칙에 문장 기능을 주석으로 붙여서 분석 과정에서 언어 기능간의 자질 통합이 이루어질 수 있도록 했다. 그리고, 무엇보다도 중요한 것은 이미 기존의 자연언어 처리 시스템이 어휘-기능 문법을 쓰고 있으면서 검증을 받았다는 데 있다. 예를 들면, 일본어-독일어 프로젝트인 Stuttgart 대학의 SEMSYN이라 든지 UMIST (University of Manchester Institute of Science and Technology)의 영어-일본어 프로젝트, CMU(Carnegie-Mellon University)의 지식 기반 기계 번역(Knowledge-based Machine Translation) 시스템 등을 들 수 있다.

4.1 입력문의 분석

입력문이 들어오면 먼저 초기화 과정, 분절 과정, 사전 정보 읽기, 형태소 규칙 적용 등을 통해서 구문분석 단계로 들어온다. 구문 분석 단계에서는 아래와 같은 규칙 정보를 이용하 여 구성 성분 구조와 기능 구조로 분리해서 분석한다. 먼저 아래의 (1)과 같은 영어 문장이 영-한 번역기의 입력문으로 들어왔다고 가정하자.

(1) That girl sings everyday.

(1)과 같은 문장이 들어와서 초기화 과정, 분절 과정을 거치고 형태소 규칙의 적용이 끝나서 구문 분석 단계로 들어와서 아래의 (6)과 같은 규칙들의 적용을 받게 된다. 일단 사전의 정보를 챠트에 읽어들인 상황에서 다음과 같은 어휘정보들이 챠트에 올라가게 된다.

(2) that $\begin{bmatrix} \text{DET} & \\ \text{DEFINITE} & + \\ \text{NUMBER} & \text{SG} \end{bmatrix}$

(3) girl $\begin{bmatrix} \text{NOUN} & \\ \text{FORM} & \text{“girl”} \\ \text{NUMBER} & \text{SG} \end{bmatrix}$

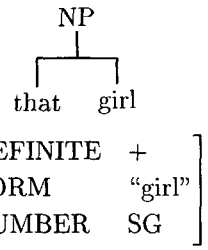
(4) *sings* $\begin{bmatrix} \text{VERB} & \\ \text{FORM} & \text{“sing”} \\ \text{PRED} & \langle \text{SUBJ} \rangle \\ \text{SUBJ} & [\text{NUMBER SG}] \end{bmatrix}$

(5) everyday $\begin{bmatrix} \text{ADV} & \\ \text{FORM} & \text{“everyday”} \end{bmatrix}$

(6) $\begin{array}{lll} \text{S} \rightarrow & \text{NP} & \text{VP} \\ & \langle\, !=\hat{}\text{SUBJ} \rangle & \\ \text{NP} \rightarrow & \text{DET} \quad \text{NOUN} & \\ \text{VP} \rightarrow & \text{VERB} & (\text{ADV}) \\ & & \langle\, !=+\hat{}\text{ADJUNCTS} \rangle \end{array}$

위와 같은 규칙을 입력문 (2)-(5)에 나와 있는 언어 정보에 적용하면 “that”과 “girl”이 NP규칙의 적용을 받아서 아래 (7)과 같은 구성 성분 구조 및 기능 구조가 생성 된다.

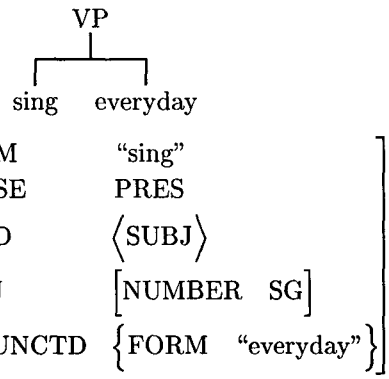
(7) that girl



(7)에서 어휘부의 품사 정보는 기능 구조에서는 중요하지 않기 때문에 위의 구구조 NP의 기능구조에 나타나지 않는다.

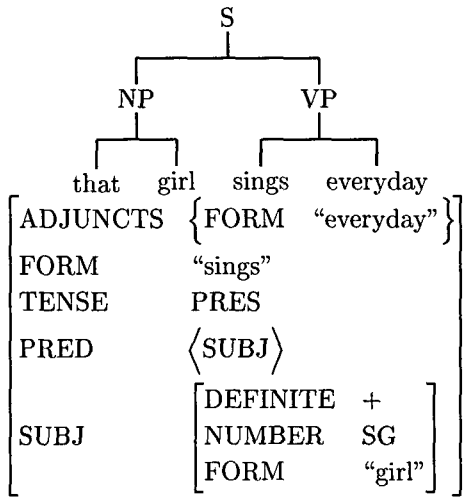
그리고 이어서 "sings"와 "everyday"에 VP 규칙이 적용되어서 아래와 같은 부분적인 기능 구조가 생성된다. 여기서 ADJUNCTS는 나타날 수 있는 개수가 통사적으로 영향을 미치지 않으므로 { }의 집합 속에 묶는다.

(8) sing everyday



이어서 문장 형성 규칙이 이미 만들어진 NP와 VP에 적용되어서 아래와 같은 구성 성분 구조와 기능 구조를 형성하게 된다.

(9) That girl sings everyday.



앞으로 변환과정을 통해서 자세히 보게 되겠지만 앞에서 언급했듯이 기능 구조는 구성 성분 구조에 비해서 비교적 개별 언어의 구조에 관계없이 독립적이다. 그리고, 두 개의 다른 구성 성분 구조라고 하더라도 동일한 기능 구조를 가질 수 있다. 예를 들어서 The girl sings everyday 와 Everyday the girl sings 는 동일한 기능 구조를 갖게 되고 한국어로 번역할 때는 동일한 문장으로 번역되게 된다.

4.2 분석문의 변환

(9)에 나와있는 기능 구조를 분석문으로 받아서 거기에 나타난 FORM의 값에 해당하는 단어들의 영한 사전의 변환 규칙을 적용받는다. 그리고 해당 영한 사전에 있는 변환 규칙에 따라서 관용어구인지 아닌지를 확인한 다음 단어 번역시에 사전의 엔트리에 있는 정보를 읽어 들인다. 참고로 영한 사전의 예는 아래 (10)에 나타나 있다.

```
(10) ek_sing ::                                          ` id 3001292x
          [
            \STK-T-V
            WORT { [ FORM "노래하"
                     TECH# GENERAL #
                   ] }
          ]
```

(10)에서 \STK-T-V는 아래 (11)에서 보여준 대로 동사의 동사 범주의 변환 규칙이다. 그리고, WORT는 IDIOM에 반하는 것으로 해당 어휘가 IDIOM의 한 부분이 아닌 경우, 즉 독립된 단어로 번역이 이루어져야 하는 경우를 나타낸다. 앞에서 언급했듯이 변환부의 규칙은 6단계로 문법 범주별로 구성되어 있고 동사의 경우를 예로 들어서 변환규칙을 살펴보면 아래와 같다.

```
(11) eX_STK-T-V ::                  ` id   1000DF8x   UPDrid   3006924x
          [
           FIRE [ 1 ? ( IDIOM ( ^, FORM ) `TRANSFER IDIOM인지를 체크
                        ? ⟨ ^ INSUX ⟩ = PLUS
                        : ⟨ ^ INSUX ⟩ = MINUS
                  )
                2 ? ( ⟨ ^ IDSUX ⟩ = MINUS `IDIOM이 아니면 FORM의 번역
                      ? TRAN ( ^, FORM )
                  )
                3 ? ( CHKLABEL ( ^, XCOMP ) `XCOMP를 논항으로 가지면
                      `필요한 한국어 형태소 자질 부여
                      ? REPVAL ( ⟨ ^ XCOMP ⟩, kS_VKON, PLUS )
                  )
                  && `SCOMP를 논항으로 취하면 필요한 형태소 자질 부여
                  ( CHKLABEL ( ^, SCOMP )
                      ? REPVAL ( ⟨ ^ SCOMP ⟩ , kS_SCOMPL , PLUS )
                        &&
                        REPVAL ( ⟨ ^ SCOMP ⟩ , SPACT, DECL )
                  )
                4 ? ( ⟨ ^ VOICE ⟩ == PASSIVE `수동문인 경우 능동문으로 바꿈
                      ? FIDO ( eX_DTD-EK-PASSIVE ⟩ ACTIVE )
                  )
                5 ? ( ⟨ ^ SUBJ FORM ⟩ == AIR `생성부에서 적용된 어순규칙 부여
                      ? ⟨ ^ SUBJ GRAFT ⟩ ~ =kR_SARP-NP
                      : ( ⟨ ^ OBJ1 FORM ⟩ == AIR
```

```
                                        ? ⟨ ^ OBJ1 GRAFT ⟩ ~ =kR_SARP-NP
                            )
                    )
                        &&
                        ⟨ ^ GRAFT ⟩ ~ = kR_SARP-SENT
                6 ? FIDO ( eX_STD-EK-V-DEFAULT )
                                        `DEFAULT 자질들의 부여
            ]
    ]
```

자세한 과정은 지면 관계상 생략하고, 변환부를 통과한 입력 구조의 결과를 살펴보면 다음과 같이 된다.

(12)
$$\begin{bmatrix} \text{ADJUNCTS} & \{\text{FORM} \quad \text{“매일”}\} \\ \text{FORM} & \text{“노래하”} \\ \text{TENSE} & \text{PRES} \\ \text{SPACT} & \text{DECL} \\ \text{GRAFT} & \text{kR\_SARP-SENT} \\ \text{PRED} & \langle \text{SUBJ} \rangle \\ \text{SUBJ} & \begin{bmatrix} \text{SPFORM} & \text{“그”} \\ \text{NUMBER} & \text{SG} \\ \text{CASE} & \text{NOM} \\ \text{FORM} & \text{“소녀”} \\ \text{GRAFT} & \text{kR\_SARP-NP} \end{bmatrix} \end{bmatrix}$$

여기서 볼 수 있듯이 변환부를 거치면서 목표어의 생성에 필요한 정보들을 형태소 결합 정보에서부터 어순배열 정보에 이르기까지 사전에서 읽어 들이거나 규칙에 있는 통합 정보를 이용하여 완성하게 된다. 이외에도 한 개의 출발어 단어가 목표어에서는 여러 개로 번역될 수 있는 경우 아래와 같이 SPARKEY나 SMOKEY를 사용한다.

(13) SPARKEY 사용의 예

```
ek_with ::                                  ` id    3002F33x
    [
      \STD-EK-PREP
      WORT { [ FORM “함께” `in case of 와 함께
               SPARKEY [ kS_PCASE kS_CNJ ]
               TECH# GENERAL #
               \ STD-EK-SELECT-IF-HUMAN
             ],
             [ FORM “” `~로 와 같은 도구격의 사용
               SPARKEY [ kS_PCASE kS_NST `INSTRUMENT CASE
                   CASEMINUS
                   ]
               TECH# GENERAL#
             ]
             }
    ]
```

(14) SMOKEY 사용의 예

```
ek_bank ::                                  ` id    30008F5x
        [
          \STK-T-N
          WORT { [ TECH #GENERAL#
                     FORM { "둑" } `river bank와 같이 합성어로 쓰인 경우
                     SMOKEY $ eX_SMOKEY-SUBSUME $
                     FS [ FORM "bank"
                          NOMAD [ form "강" ]
                          ]
                   ],
                     [ TECH# GENERAL #
                     FORM { "은행" }
                   ]
                   }
        ]
```

4.3 변환문의 생성

(12)와 같은 기능 구조가 변환부를 거치면서 완성되어서 생성과정으로 넘어오게 되면 먼저 목표어 사전에 있는 아래와 같은 해당 단어들 (15)-(17)을 읽어 들이게 된다.

(15) kW_소녀 :: ` id 20053BBx

```
        [
          \STK-N
          \STK-HEWMN-P
          \STK-CONCR-P
          \STK-COUNT-P
          \STK-INFL-N-V-ENDING `모음으로 끝난 명사의 형태소 정보
          FS     [
                   FORM "소녀"
                   KLAW "명"
                 ]
        ]
```

(16) kW_노래하 :: ` id 200549Bx

```
        [
          \STK-V
          \STK-COMPL-INTR
          \STK-COMPL-ACC
          \STK-INFL-V-IRR-HA-ENDING `하-변칙 형태소 정보
          FS     [
                   FORM "노래하"
                 ]
        ]
```

(17) kW_매일 :: ` id 2004877x

```
        {
        0 = [
                   \STK-INFL-N-L-ENDING `ㄹ로 끝난 명사의 형태소 정보
                   \STK-HEWMN-M
```

```
                \STK-N
                \STK-CONCR-M
                \STK-COUNT-M
                \
                FS  [
                      FORM "매일"
                      USE # TIME#
                    ]
            ]
    1 = [
                \STK-ADV
                \STK-ADV-MOD-V
                FS  [
                      FORM "매일"
                    ]
            ]
    }
```

이들 사전에 있는 형태소 규칙에 관한 정보를 이용하여 한국어 형태소 생성 규칙의 적용을 받는다. 위에서 예로 든 "노래하"와 "소녀"가 거쳐가는 형태소 규칙은 아래와 같다.

(18) 동사 형태소 생성 규칙의 예

```
kR_SARP-IND-VERB ::                    ` id 1002094x UPDrid 3006A46x
    [:
        VSTEM. / ? ⟨ ^ TENSE ⟩ == PRES
                   &&
                   ⟨ ^ FORMALITY ⟩ == MINUS
                   &&
                   ⟨ ^ SPACT ⟩ == DECLAR
        mVLAST-CHK
        mIND
        mSFM
    :]
    →
    SUCCEED ()
```

(19) 명사 형태소 생성 규칙의 예

```
kR_SARP-NPCASE ::                      ` id 1002084x UPDrid 3006924x
    [:
        NSTEM. / ? ~ ( ⟨ ^ CASE ⟩ = MINUS )
        ( mPLR )
        mCASE
    :]
    →
    SUCCEED ( )
        &&
```

FIDO (kX_KLAW-FORM) ` 필요한 경우에 한국어의 개사를 생성하는 규칙

이들 규칙의 적용을 받아서 나온 기능구조는 아래의 (20)과 같다.

(20) 생성 형태소 규칙 적용 후의 기능 구조

$$\begin{bmatrix} \text{ADJUNCTS} & \{\text{FORM} \quad \text{“매일”}\} \\ \text{FORM} & \text{“노래하”} \\ \text{SARPFORM} & \text{“노래한다”} \\ \text{TENSE} & \text{PRES} \\ \text{SPACT} & \text{DECL} \\ \text{GRAFT} & \text{kR\_SARP-SENT} \\ \text{PRED} & \langle \text{SUBJ} \rangle \\ \text{SUBJ} & \begin{bmatrix} \text{SPFORM} & \text{“그”} \\ \text{NUMBER} & \text{SG} \\ \text{CASE} & \text{NOM} \\ \text{FORM} & \text{“소녀”} \\ \text{SARPFORM} & \text{“소녀가”} \\ \text{GRAFT} & \text{kR\_SARP-NP} \end{bmatrix} \end{bmatrix}$$

생성 형태소 규칙의 적용 결과인 (20)에 GRAFT의 값으로 되어 있는 아래와 같은 어순 배열 규칙이 적용된다. 이들 어순배열 규칙은 문법 범주는 상관없이 기능 구조상에 해당 문법 기능을 가진 논항들은 규칙에 나와 있는 순서대로 배열된다. (21)에서 주어-직접 목적어-간접목적어-수식어-동사-문장부호의 순서로 배열된다.

(21) 문장 어순 배열 규칙

```
kR_SARP-SENT ::                          ` id 100208Cx UPDrid 3006924x
    [
        SARPKAT : SUBJ
        SARPKAT : OBJ1
        SARPKAT : OBJ2
        SARPKAT : ADJUNCTS
        SARPKAT : ASRPFORM
        SARPKAT : PUNCK
    ]
    →
    SUCCEED ( )
```

아래의 명사구의 어순 배열 규칙도 명사 자신은 SARPFORM이라고 표시된 자리에 나타나고, 나머지 논항이나 수식어들은 아래에 나와있는 차례대로 수식어-관사-소유격-합성명사의 앞부분-명사의 순서대로 나타난다.

(22) 명사구 어순 배열 규칙

```
kR_SARP-NP ::                                  ` id 100208Ax UPDrid 3006924x
    [
        SARPKAT : ADJUNCTS  ? ~ ( ⟨ ! POSTNOM ⟩ == PLUS )
        SARPKAT : SPFORM    ? ~ ( ⟨ ! POSTNOM ⟩ == "그" )
        SARPKAT : POSS
        SARPKAT : PUNCK
        SARPKAT : NOMAD
        SARPKAT : SARPFORM
    ]
    →
    SUCCEED ( )
```

(20)의 기능 구조가 (21)과 (22)의 문장 어순배열 규칙 및 명사구 어순배열 규칙의 적용을 받아서 나온 결과는 아래와 같다.

(23) 그 소녀가 매일 노래한다.

이상과 같이 구문 분석 단계를 통해서 나온 기능 구조를 중심으로 변환부를 거쳐서 생성부에서 생성 형태소를 적용하고 이어서 어순 배열 규칙을 적용하여 최종적으로 목표어의 문장이 번역되어 나타나는 과정을 자세히 살펴보았다.

5. 맺는말

초기의 영어-러시아어 중심의 직접 번역 방식의 기계번역의 성과를 비판한 ALPAC 보고서가 1966년에 나온 후로 (Somers and Whitelock (1995) 기계번역은 한동안 동면에 들어갔다. 그러다가 70년대 말에 이르러 언어학의 발달로 언어들 간에 보다 복잡한 문장의 구조적인 해결이 가능하게 되면서 변환 번역 방식의 기계 번역이 시작되어서 지금까지도 채택되어 쓰이고 있다. 물론, 그 사이에 구조 지향적 번역에서 어휘의 중요성을 인식하고 어휘쪽에 보다 많은 정보를 기록해서 번역상의 어려움을 해결하려는 방향의 전환도 있었다. 이와 더불어 최근에는 통계적 방법에 의한 번역도 어휘의 공기 확률에 따른 애매성의 해결 방법과 interlinear text를 이용한 적정 번역도 시도되고 있다. 이와같은 여러 가지 새로운 시도들은 이 글에서 논의된 변환 번역 방식의 L&H 한-영 양방향 번역 시스템에서도 필요한 대로 채택해서 쓸 수 있는 방법들이다. 즉 구조 중심의 변환 번역 방식에 풍부한 어휘사전 및 통계적인 방법과 여러 가지 담화 영역 정보들을 처리할 수 있는 모듈들이 맞물려 돌아간다면 훨씬 더 질높은 번역이 가능할 것이다.

앞으로 기계번역은 인터넷의 급속한 성장과 더불어 요약 번역을 할 수 있는 능력이 강조되고, 입력문에서 흔히 일어나는 수행 과오 (PERFORMANCE ERRORS)를 처리할 수 있는 시스템이 요구될 것으로 보인다. 그리고, 끊임없는 기계번역의 숙제인 번역의 질을 고양시키기 위해서는 결국 기계번역은 근본적으로 언어 지식 처리 기술이라는 생각이 자리 잡아야한다. 다시말해, 기계 번역의 문제는 대부분 언어학적인 문제들로 언어학자들이 고민하고 처리해야 되는 부분들이다. 전문가의 지식 기반이 컴퓨터에 잘 결합이 되고, 그것을 기반으로 해서 지식 기반 데이터 베이스가 실제로 나타나는 언어 자료들과 더불어 구성되어 있어야 한다. 그리고, 담화단위나 텍스트 단위의 연구가 더욱 열심히 되어서 문장의 구조를 설명하듯이 담화의 구조가 설명이 되고 텍스트를 구성에 대한 이해가 심화되면 기계번역의 수준은 한층 높아질 것이다.

이러한 기계번역 시스템의 개발로 인하여 미국, 유럽, 일본을 비롯한 해외 정보를 누구나 쉽고 빠르게 접할 수 있게 되고, 이는 곧 바로 개인은 물론 국가 경쟁력 향상에 크게 도움이 될 것이다. 아울러 언어의 장벽 때문에 주춤하고 있는 인터넷 이용을 확산할 수 있으며 PC 판매량 증가, PC 번들링 소프트웨어로 PC 자체가 번역기가 되므로 지금까지 PC를 이용하지 않

았던 사용자들이 대거 PC의 적극적인 사용자들이 되어서 국가의 정보화 기반이 두터워지고 더욱 강화될 것이다.

참고문헌

한국 과학기술연구원 부설 시스템 공학 연구소. 1998. '98 국가 정보화 백서. 한글정보처리.

최승권. 1999. 자동번역 기술 동향: 언어학적 관점에서. 한국 언어정보학회 소식, 20.

Pentheroudakis, Joseph. 1990. Complex bidirectional transfer in the ecs korean-english system. *Proceedings of SICONLP '90.*

Somers, Harold. and Peter Whitelock. 1995. *Linguistic and Computational Techniques in Machine Translation System Design.* University College London Press, London.

접수일자: 1999년 5월 29일
게재결정: 1999년 6월 28일