

독-한 명사구 기계번역시스템의 구축

이민행 · 최승권 · 최경은
연세대 · SERI · 연세대

Lee, Minhaeng, Sung-Kwon Choi, and Kyung-Eun Choi. 1998. **The Construction of a German-Korean Machine Translation System for Nominal Phrases.** *Language and Information* 2.1, 79-105. This paper aims to describe a German-Korean machine translation system for nominal phrases. Besides, we have two subgoals. First, we are going to reveal linguistic differences between two languages and propose a language-informational method to overcome the differences. The method is based on an integrated model of translation knowledge, efficient information structure, and concordance selection. Then, we will show the statistical results about translation experiment and its evaluation as an evidence for the adequacy of our linguistic method and translation system itself. (Yonsei University, SERI, Yonsei University)

1. 머리말

컴퓨터가 자동으로 한 언어의 텍스트를 다른 언어의 텍스트로 번역하기 시작한 지 거의 50여 년이 지났다. 국내에서의 기계번역 시스템은 영어와 관련되어 상품화되거나 혹은 신기술에 대한 연구가

*이 논문은 1996년 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음.

지속적으로 연구되고 있는 반면 (시스템공학연구소 1992, Kim *et al.* 1993, 시스템공학연구소 1996), 독일어와 관련해서는 언어학적으로 이론적인 토대가 형성되고 (Lee 1992, 이민행 1994), 독-한 기계번역에 대한 프로토타입이 마련되었지만 (Choi 1995, 최승권 1996), 본격적인 독한 기계번역시스템에 대한 연구는 거의 없었다.

이러한 배경하에 이 연구에서는 번역단위를 (관계절을 포함하지 않는) 명사구로 한정하는 독-한 명사구 기계번역 구축을 시도하였다. 독일어 명사구를 한국어로 자동번역하는 데 있어서 두 언어간의 언어적 차이로 발생할 수 있는 언어학적 구조와 정보의 차이를 체계적으로 정리하며, 구조와 정보의 차이를 자동번역에서 해결하기 위한 효율적인 언어학적 번역지식 모델과 방법론을 제안하고 이를 번역실험을 통해 검증하고자 한다.

본 독-한 명사구 기계번역 시스템연구는 영어, 독일어, 불어, 중국어를 대상으로 프로토타입 수준의 다국어 기계번역을 시도한 CAT2 다국어 기계번역 시스템 (Sharp 1988, Choi 1995)을 독-한 명사구 기계번역 시스템으로 이용했다.

논문의 구성은 다음과 같다. 2장에서는 정형화된 독일어 명사구 이외에 실제 독일인에 의해 사용되는 독일어 명사구의 언어학적 구조를 살펴본다. 이를 위하여 독일어 명사구 코퍼스를 수집하고 전문번역가에 의한 한국어로의 번역을 거친 후 이 두 언어 코퍼스를 바탕으로 독일어와 한국어간의 명사구 차이를 대조 언어학적 관점에서 유형별로 정리.기술한다. 3장에서는 다국어 기계번역을 목표로 한 독-한 명사구 기계번역 시스템의 전체 번역 구성도를 소개한다. 4장에서는 본 연구에서 사용된 언어정보 기반 번역지식 즉 번역규칙 모델과 정보구조를 소개하며 5장에서는 2장에서 정리.기술된 독-한 명사구간의 대조 언어학적 차이를 해결하기 위한 다양한 정보언어학적 해결방안을 제시한다. 그리고 6장에서는 앞서 제안된 언어학적 해결방안들의 타당성을 검증 및 평가하기 위하여 번역실험한 결과를 통계치로써 제시하고자 한다. 마지막 7장 맺음말에서는 지금까지의 논의를 종합하고, 남은 문제를 살펴본다.

2. 독일어 명사구 코퍼스와 독-한 명사구의 언어학적 차이

2.1. 독일어 명사구 코퍼스

코퍼스란 컴퓨터가 읽을 수 있는 형태(machine-readable form)로 저장된 자연언어자료와 이들 자료에 대한 부속 정보들의 묶음이라고 할 수 있다. 언어학 연구에 코퍼스를 이용하였을 때의 장점은 실세계의 다양한 언어현상에 대한 규칙 기반 접근 방법의 제한적인 설명력을 보완하고 확인할 수 있으며 또한 코퍼스로부터 통계 정보도 이용할 수 있다는 것이다.

이 연구에서는 다양한 독일어 명사구 언어현상을 포함하는 균형화된 독일어 명사구 코퍼스를 만들기 위해 독일어 DUDEN사전의 용례와 독일의 잡지와 신문기사로부터 독일어 명사구를 수집한 후 독일어 전공 번역전문가가 각 독일어 명사구에 대한 한국어 대역어를 달고 독일어 명사구의 성분들에 품사를 기입하였다. 이 연구를 통해 수집된 독일어 명사구의 코퍼스 양은 다음의 [표 1]과 같다 :

수집한 자료	명사구의 수
DUDEN 독일어 사전	1400개
독일잡지 DER SPIEGEL 37호	680개
독일신문 DIE ZEIT 37호	250개

[표 1] 수집된 독일어 명사구 코퍼스

위의 표에서 보듯이 수집된 독일어 명사구는 총 2330개이고 이들 명사구의 평균 단어 수는 4개이다.

2.2. 독-한 명사구의 언어학적 차이

수집된 독일어 명사구 코퍼스와 그것의 한국어 대역어 코퍼스를 비교분석함으로써 발견된 독일어 명사구와 한국어 대역어간의 언어학적인 차이는 다음과 같이 몇 가지로 정리될 수 있다.

▷ 어순의 차이

코퍼스로부터 독일어 명사구에 대한 한국어 대역어를 비교분석해 보면 그들의 기본적인 어순차이는 다음과 같다.

독일어	한국어
수사-명사	수사-명사
지시사-명사	지시사-명사
소유격 명사구-명사 / 명사-소유격 명사구	소유격 명사구-명사
형용사-명사	형용사-명사
명사-관계절	관계절-명사

[표 2] 독일어 명사구와 한국어 대역어의 기본어순 차이

이상의 코퍼스에 기반한 독일어와 한국어의 명사구 어순은 Hawkins (1983)에서 논의된 바, 언어유형론에 따른 두 언어간의 명사구 기본어순과 다르지 않다는 것을 다시 확인할 수 있다.

▷ 범주적 차이

범주적 차이란 독일어 명사구와 한국어 대역어 사이에 품사적인 전이현상이 일어난 것을 의미한다. 다음의 코퍼스의 예들이 이 사실을 보여준다.

- (1) a. hohes Wachstum 고도 성장
 Adj N N N
- b. keine Videoueberwachung 비디오감시를 하지 않음
 Det N PP FV-Neg-NSuff
- c. die Gewerkschaft der Polizei 경찰노조
 Det N Det N N N

(1a)는 형용사-명사로 이루어진 독일어 명사구가 한국어 대역어에서 조합적 번역이 이루어지기는 했지만 형용사-명사의 명사구로 만들어지지 않고 명사-명사의 합성명사로 만들어진 예이다. (1b)는

관사-명사로 이루어진 독일어 명사구가 한국어 대역어에서는 조사구(PP)-기능동사(FV)-부정어(Neg)-명사화접미사(NSuff)로 만들어진 예이다. (1c)에서는 관사-명사-관사-명사의 독일어 명사구가 한국어 대역어에서는 명사-명사로 만들어진 예를 나타낸다.

▷ 어휘적 변이

어휘적 변이는 독일어의 어휘가 조합적으로 번역되지 않고 함께 사용된 단어들에 따라 다르게 번역되는 것을 의미한다. 다음의 코퍼스가 어휘적 변이를 나타내는 예이다.

(2) a. Die guten Auswirkungen

auf die wirtschaftliche Dynamik

경제의 동력에 미치는 좋은 영향

b. Die Auswirkung des Teigs

가루 반죽 하기

c. Ein schwerer Beutel

묵직한 지갑

d. Ein besonders schwerer Verstoß

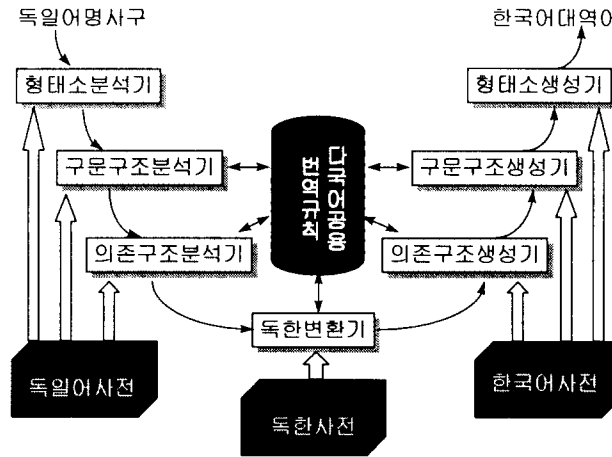
특히 심한 위반

(2a)와 (2b)에서 독일어의 Auswirkung(en)은 원래 '마무리, 성취, 도달'의 의미이나 함께 사용된 다른 단어들과의 연관성에 의해 다른 의미로 변이되는 것을 보인다. (2c)와 (2d)에서 독일어의 schwer는 원래 '무거운'의 뜻이나 함께 사용된 어휘에 따라 (2d)에서는 변이된 의미로 쓰이고 있다.

3. 전체 번역시스템 구성도

이 연구가 기반으로 삼고 있는 독-한 명사구 기계번역은 기계번역 방법론중에서 변환기반 기계번역 방법에 속하며 언어학 정보를 기반으로 하는 언어정보지향 기계번역 방법에 속한다. 일반적인 변환기반 기계번역 시스템들과 마찬가지로, 본 연구를 통해 구축한 독-한 명사구 기계번역 시스템도 독일어 명사구의 분석부분과 독-

한 변환부문 그리고 한국어 생성부문에 나뉘어 진다. 그러나, 기존의 양국어(bilingual) 혹은 다국어(multilingual) 변환기반 기계번역 방법과는 달리, 언어 유형론에 이론적인 토대를 두고 유사 언어 유형군에 속하는 언어들은 공통의 번역지식을 공유한다는 다국어 공용 번역지식을 사용한다는 점에 본 시스템의 특징이 있다. 다국어 공용 번역규칙에 대해서는 다음절에서 자세히 기술한다. 독-한 명사구 기계번역의 전체 번역시스템 구성도는 다음과 같다.



[그림 1] 독-한 명사구 기계번역시스템 구성도

또한, 각 모듈의 입출력 자료와 기능은 각각 다음과 같다.

- ▷ 형태소 분석기 : 독일어 명사구 문자열을 입력으로 받아 빈칸을 경계로 하여 단어 문자열들을 만들고 단어 문자열을 독일어 사전과 비교하여 독일어 사전으로부터 단어 문자열의 기본어휘와 사전정보를 출력하는 모듈.
- ▷ 구문구조 분석기 : 단어의 기본어휘와 그것의 사전정보를 입력으로 받아 독일어 명사구의 올바른 구문구조를 출력하는 모듈. 하나의 독일어 명사구 문자열에 대해 한 개 이상 만들어

지는 통사적 중의성을 해소하는 제약작업도 이루어진다.

- ▷ 의존구조 분석기 : 구문구조 분석기로부터 출력된 독일어 명사구의 구문구조를 입력으로 받아 전치사, 관사와 같은 기능어를 삭제하고 삭제된 기능어의 정보를 기본어휘에 전달하고 전체적으로 어휘의미 지향적 의존구조를 출력하는 모듈.
- ▷ 독-한 변환기 : 독일어 명사구의 의존구조를 입력으로 받아 해당되는 한국어 의존구조로 출력하는 모듈.
- ▷ 의존구조 생성기 : 한국어 의존구조를 입력으로 받아 한국어의 어순에 맞는 어휘의미 지향적 의존구조를 생성하는 모듈.
- ▷ 구문구조 생성기 : 한국어 어순에 맞는 의존구조를 입력으로 받아 한국어 구문구조에 맞는 기능어와 구구조를 생성하는 모듈.
- ▷ 형태소 생성기 : 한국어 구문구조를 입력으로 받아 규칙 및 불규칙 활용이 반영되는 한국어 문자열을 만드는 모듈.
- ▷ 다국어 공용 번역규칙 : 다국어에 공통으로 적용될 수 있는 번역규칙이 저장되어 있는 번역지식 베이스. 언어보편적인 번역규칙이라기 보다는 언어유형론에 입각한 번역규칙의 개념이며 언어유형간 유사도에 따라 유사 언어유형에 속하는 언어들은 번역규칙을 공유하는 정도가 높다.

4. 번역지식 모델

이 장에서는 독-한 명사구 기계번역 시스템에서 사용한 번역 지식, 즉 번역규칙 모델과 사전정보구조 모델의 형식과 내용을 살펴본다.

4.1. 번역 규칙 모델

본 독-한 명사구 기계번역을 위한 번역규칙은 문법규칙과 변형규칙으로 구성된다. 문법규칙은 각 모듈 단계에서 개별언어들의 언어

현상을 설명하고 기술하기 위해 사용되는 규칙을 말하며 변형규칙은 모듈 단계와 모듈 단계를 연결짓는 규칙을 말한다. 이들 번역규칙 모델은 현재 발전하고 있는 문법형식(grammar formalism)의 틀 내에서 “문법기술의 편리성”과 “규칙의 재활용성”이란 두가지 목표에 초점을 두고 만들어졌다. 첫째 목표인 문법기술의 편리성은 문법작성자가 각 언어에 대해 축적하고 있는 언어학 지식을 기계번역에서 편리하게 기술할 수 있도록 지원할 수 있는 번역규칙 모델이어야 한다는 요구이며, 두번째 목표인 규칙의 재활용성은 독-한 명사구 기계번역에서 적용되었던 번역규칙을 다른 언어와의 기계번역에도 재활용할 수 있어야 한다는 요구이다. 이러한 두가지 요구사항을 충족하는 번역규칙 모델은 다음과 같은 효과를 가지게 된다 (Choi *et al.* 1997).

▷ 번역지식 크기의 축소

번역규칙이 독일어와 한국어에 의해 단지 한번 호출(load)되어 공유될 수 있기 때문에 전체 번역시스템의 메모리 크기를 줄일 수 있다.

▷ 번역지식 관리의 편리성

번역규칙이 독일어와 한국어에서 한번의 수정으로 공통으로 적용될 수 있기 때문에 번역규칙의 작성 및 수정에 일관성 유지 및 관리가 편리.

▷ 번역지식의 재활용

기존의 번역규칙을 새로 첨가된 언어에서도 전체 혹은 부분적으로 재활용할 수 있다.

이상에서 설명된 번역규칙 모델이 실제 어떠한 모습을 가지는 지에 대해서 다음 절에서 논의한다.

4.1.1. 문법 규칙

언어간의 차이에도 불구하고 언어들간의 언어현상을 동일한 문법 규칙으로 기술할 수 있기 위해 독-한 명사구 기계번역에서는 문법 규칙 모델로 X-bar 통사 이론[Jackendoff, 1977]과 HPSG[Pollard and Sag, 1994]를 혼합한 문법 규칙 모델을 사용하였다. 이 문법 규칙 모델에서 X-bar 통사이론의 개념은 독일어와 한국어의 관찰 구조와 선형구조를 설명하는 데 기여를 하며 HPSG의 형식모델 개념은 X-bar통사론에 의해 구성된 구구조 내에서의 효율적인 언어학 정보 이동에 이용된다. 이 혼합된 문법 규칙 모델은 다양한 독일어 언어현상에 대응력을 높이기 위해 이분지 방법을 기초로 하였다. 그러나 접속어구는 이분지 방법을 택하는 대신에, 분석구조의 깊이를 고려하여 삼분법 규칙으로 기술하고 접속어구의 상위노드는 접속성분이 아닌 접속사가 중심어가 되도록 하였다. 독-한 명사구 기계번역을 위해 작성한 문법규칙 모델을 비형식화하여 나무구조로 그리면 전체적인 모습은 다음 쪽의 [그림 2]와 같다. [그림 2]에서 중심어의 위치에 따라서 규칙들은 중심어가 오른쪽에 오는 '오른쪽 중심어 구조'와 중심어가 왼쪽에 오는 '왼쪽 중심어 구조'로 구분되어 있으며 중심어의 기능에 따라 술어와 그것의 논항간의 관계를 기술하는 '술어논항 처리규칙', 수식어와 중심어간의 관계를 기술하는 '수식어 처리규칙', 기능어와 비기능어간의 관계를 기술하는 '기능어 처리규칙', 접속어와 비접속어들을 기술하는 '접속어구 처리규칙'으로 나뉘어져 있다.

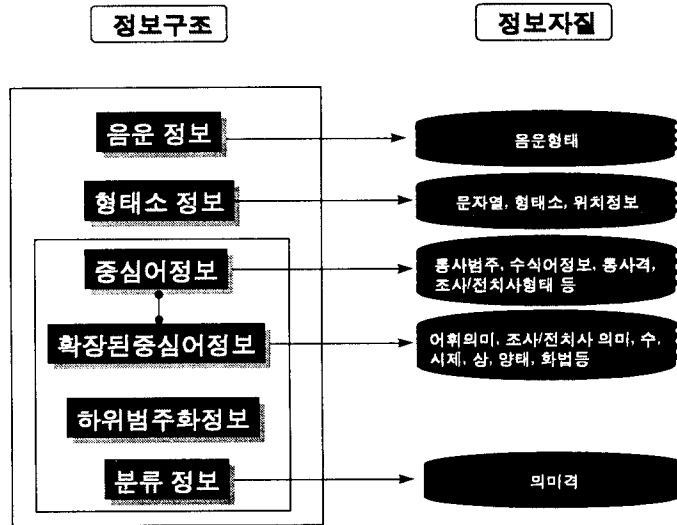
의존구조 => 의존구조 (변환)	1. 어휘의미정보의 전달 2. 어휘 및 구조 변환 3. 어순 조정
의존구조 => 구문구조 (생성)	1. 기능어 정보로부터 기능어 구조 생성 2. 기능동사 정보로부터 기능동사 생성 3. 어휘의미정보로부터 통사정보 생성 4. 의존구조로부터 구구조 생성

[표 3] 단계별 변형규칙과 언어학적 기능

4.2. 정보 구조 모델

기계번역에서 번역을 위한 언어학 지식은 일종의 언어정보로 간주되어, 모두 “속성:값”의 형태를 가진 자질구조를 통해 표상된다. 하위층위의 언어정보들은 통합(Unification)이란 조작을 거쳐 상위층위의 언어정보를 생성된다. 이러한 정보들을 사전과 번역규칙이 필요로 한다.

이 연구에서 다루어지는 정보 구조는 언어학 지식에 따라 크게 3개 군으로 분류되었다. 즉 음운 정보, 형태소 정보, 통사-의미 정보이다. 이 3개군 중 통사-의미 정보군은 다시 4개의 세부군으로 나뉘는데 이들은 각각 중심어 정보, 확장된 중심어 정보, 하위범주화 정보, 분류정보이다. 중심어 정보군은 중심어의 통사정보를 담은 정보군이며 확장된 중심어 정보군은 기능어와 기본어휘가 공유할 수 있는 통사적 정보뿐만 아니라 어휘의미 정보를 담고 있는 정보군이다. 또 하위범주화 정보군은 중심어가 하위범주화할 수 있는 성분들의 수와 정보들을 기술하는 군이며 분류정보군은 의미격이 담겨 있는 정보군이다. 이상 정보구조에서 기술된 대정보군과 세부정보군은 다음쪽의 [그림 3]과 같이 표현될 수 있다.



[그림 3] 독한 기계번역 시스템의 정보구조

이들 정보구조와 정보자질들은 문법기술가에 의해 삭제, 첨가, 수정이 가능하다.

■ 확장된 중심어(extended head) 정보

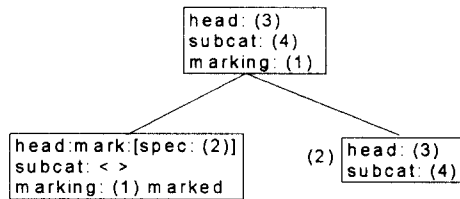
통사-의미 정보군에 속하는 확장된 중심어 정보는 기능어와 기본 어휘가 공유할 수 있는 정보를 담고 있는 정보군으로서 중심어 정보군의 부분집합이다. 이 확장된 중심어 개념은 Grimshaw (1991)에서 새롭게 제안된 개념으로, 이 개념을 본 연구에서는 범주 뿐만 아니라 모든 자질에까지 확장 적용하였다. 이러한 확장 적용에 따르면, 전치사구가 전치사와 명사구로 이루어지고 명사구가 관사와 명사로 이루어져 있을 때 전치사와 관사와 명사는 확장된 중심어 정보군에 있는 모든 정보자질들을 공유하게 되며 명사구와 전치사구도 또한 확장된 중심어 정보를 공유하게 된다.

확장된 중심어를 이용하여 번역지식을 기술하고 기계번역을 실행

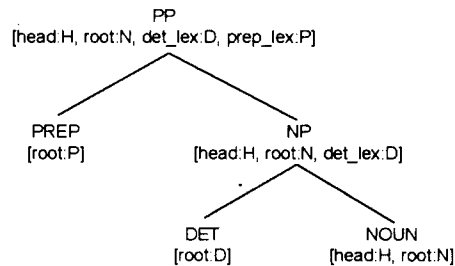
하는데 있어서의 장점은 기능어 처리 및 번역을 더욱 효율적으로 할 수 있다는 점에서 찾을 수 있다. 왜냐하면 기능어가 가지고 있는 통사 정보와 의미정보를 기능어와 관련된 기본어휘가 항상 가지고 있어서 기능어를 삭제하더라도 그것이 번역에 기여하는 정보를 손실하지 않기 때문이다. 또한 확장된 중심어가 중심어 정보의 부분집합이기 때문에 정보를 이동할 때에도 중심어 정보를 이동함으로써 확장된 중심어가 함께 이동되기 때문에 따로 확장된 중심어를 위한 이동절차를 만들 필요가 없다는 잇점이 있다.

기능어 처리와 관련된 확장된 중심어 개념의 효율성은 타 문법 이론이나 문법형식과 다음과 같이 비교된다.

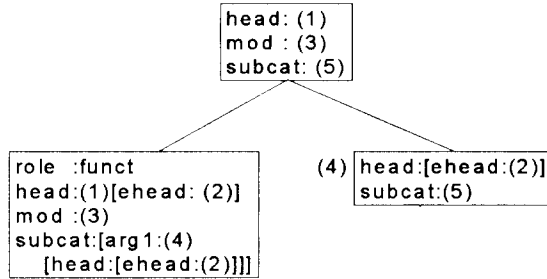
(3) a. HPSG에서의 기능어 처리 (Pollard and Sag 1994)



b. 확장된 CFG에서의 기능어 처리 (한국과학기술원 1992)



c. 확장된 HEAD에서의 기능어 처리



(3a)는 HPSG에서의 기능어 처리에 관한 것이다. HPSG에서는 기능어 정보가 marking이란 자질에 저장되어 상위노드로 이동하게 되며 필요시 제약조건으로 사용하게 된다. 하지만 이러한 방법에 의하면 기능어가 1개 이상 나타날 때 그들간의 관계를 규정하고 제어할 방법이 없다는 것이 단점이다. (3b)는 확장된 문맥자유문법에서의 기능어 처리이다. 정보를 통합하는 완전한 기능이 없기 때문에 기능어 정보가 상위노드로 이동하기 위하여 매번 기술되어야 한다는 불편이 있다. (3c)는 확장된 중심어 개념에 의한 기능어 처리이다. 기능어가 중심어가 되기 때문에 기능어의 특수 정보도 함께 상위로 이동하게 된다. 이상의 관찰에 의하여 확장된 중심어 개념은 기능어가 여러 개 나타날 지라도 그들의 정보가 확장된 중심어 자질에 모두 쌓이기 때문에 기능어들간의 관계를 설명할 수 있으며 중심어 정보와 더불어 함께 이동되기 때문에 정보이동도 편리하다는 장점을 가진다.

5. 독-한 명사구 기계번역

5.1. 명사구 어순의 기계번역

앞서 2.2절에서 언급된 독일어와 한국어의 명사구 기본어순의 차이는 변환 방식에 따라 크게 두가지로 구분된다. 구조적 변환없이 조합적으로 변환할 수 있는 방식과 구조적 변환이 필요한 비조합적

변환방식이다. 2.2절에서 소개된 명사구 기본어순을 조합적 변환방식과 비조합적 변환방식으로 나눈 결과는 다음의 [표 4]와 같다:

변환 방식	독일어	한국어
조합적	수사-명사	수사-명사
	지시사-명사	지시사-명사
	형용사-명사	형용사-명사
	소유격 명사구-명사	소유격 명사구-명사
비조합적	명사-소유격 명사구	소유격 명사구-명사
	명사-관계절	관계절-명사

[표 4] 독일어와 한국어의 명사구 기본어순 차이의 변환방식 구분

조합적 변환방식은 어휘적 변환만이 이루어지므로 특별한 변환과정을 요구하지 않기 때문에, 다국어 공용 번역규칙에 의해 이루어지는 반면 비조합적 변환방식은 언어간 구조적인 차이를 반영하기 때문에 독일어와 한국어간의 특수변환 규칙의 적용을 받는다. 이상의 변환방식은 다음과 같은 번역규칙을 통해 처리된다.

- (4) a. 조합적 변환을 위한 다국어 공용 번역규칙
 $\{ \}.[+] \Rightarrow \{ \}.[+]$.
- b. 비조합적 변환을 위한 특수번역규칙
 $\{ \text{head:(1)} \}.[\{ \text{head:(1)} \}, \{ \text{role:mod} \}] \Rightarrow$
 $\{ \text{head:(2)} \}.[\{ \text{role:mod} \}, \{ \text{head:(2)} \}]$.

(4a)의 '+'는 하위노드들의 자질구조를 나타내며 동일한 구조가 변환없이 목표언어의 구조로 전달된다는 의미이다. (4b)는 독일어의 수식어 구조는 한국어에서 중심어의 앞에 온다는 사실을 나타낸다. 위의 두 번역규칙은 적용순서로 보아 특수번역규칙 이후에 공용번역규칙이 적용된다. 때문에 특수번역규칙에 적용되지 않는 구조는 공용번역규칙의 적용을 받는다.

5.2. 파생어의 기계번역

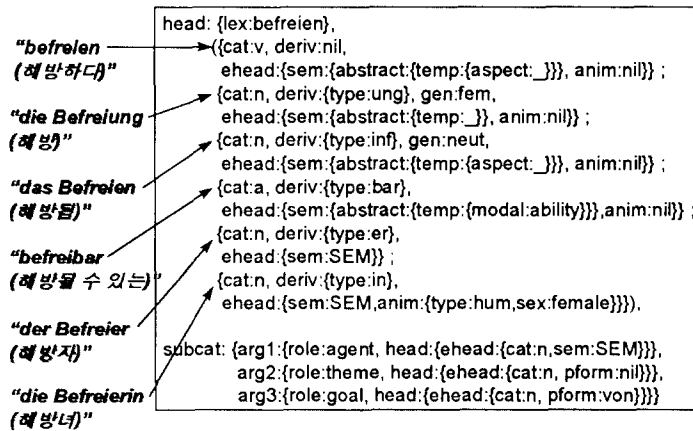
파생어는 어휘의 수를 증가시키는 중요한 언어현상중의 하나이며 파생어가 기계번역에서 중요한 이유는 품사전이 현상과 밀접한 관련이 있기 때문이다. 파생어 처리와 관련해 파생어 전체를 사전에 등록할 것인지 아니면 기본어휘만을 사전에 등록하고 파생어에 대한 정보는 형태소 분석시 얻을 것인지에 대한 논의가 여전히 자연어처리의 주요이슈이다. 이 연구에서는 파생어 처리와 관련해 기본어휘만을 사전에 등록하고 형태소 분석시 파생어에 대한 정보를 얻는다는 입장을 취하며, 이 입장을 뒷받침하는 경험적인 자료로 다음과 같은 파생어 번역 예를 들 수 있다.

(5) a. Befreien		해방됨
N		N-FV-NSuff
b. keine Videoueberwachung		비디오감시를 하지 않음
Det	N PP	FV-Neg-NSuff

(5a)는 독일어의 파생명사(N)가 한국어에서 명사(N)-기능동사(FV)-명사화접미사(NSuff)로 번역되는 경우이며 (5b)는 독일어의 관사(Det)-파생명사(N)가 한국어에서 조사구(PP)-기능동사(FV)-부정어(Neg)-명사화접미사(NSuff)로 번역되는 경우이다. 위의 예들은 파생어 전체가 단어로써 사전에 등록될 경우 대역어에서 파생어에 삽입된 다른 품사의 어휘를 생성하기 위해 파생어에 대한 분석이 생성시에 이루어져야 한다는 사실을 보여주며, 파생어에 의한 품사전이현상을 설명하기 위해 모든 파생명사에 대해 변환과정에서 품사전이현상에 대한 설명이 이루어져야 한다는 점을 명확히 한다. 이러한 설명은 파생어 전체를 사전에 등록할 지라도 어디에선가 그 파생어에 대한 형태소 분석이 이루어져야 한다는 것을 의미하는 것이다.

이러한 맥락에서 이 연구는 파생어의 경우, 기본어휘만 사전에

기입하고 파생어에 대한 정보는 사전 정보로써 기입하여 형태소 분석시 올바른 파생어 형태에 대한 기본어휘와 관련 언어정보를 얻을 수 있도록 하였다. 이러한 방안은 기본어휘만이 사전에 등재되기 때문에 전체적으로 사전의 크기를 줄이는 효과를 가진다. 다음의 예는 본 논문에서 파생어를 사전에 등록한 독일어 사전의 예를 보여준다.



[그림 4] 기본어휘 befreien(해방하다)의 파생어 사전정보

위의 [그림 4]에서 기본어휘인 befreien은 파생접미사 deriv의 값과 관련해 Befreiung, Befreien, befreibar, Befreier, Befreierin등으로 파생될 수 있는데 파생어들 모두 기본어휘인 befreien의 어휘값과 하위범주화 정보를 공유하며 각 파생어들의 고유정보는 각각의 중심어 정보에 기입된다는 것을 알 수 있다.

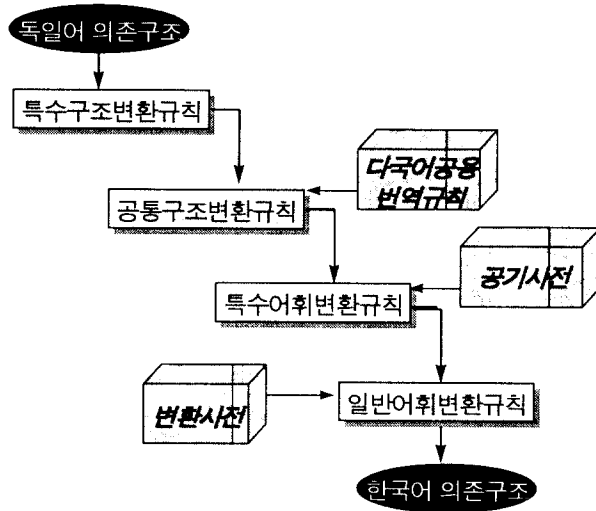
본 연구에서의 파생어의 기계번역은 파생어의 기본어휘와 해당 파생 정보에 의해 이루어지는 데 예를 들어 (5b)의 'keine Videoueberwachung'이 한국어로 변환되는 규칙과 과정을 보이면 다음과 같다.

- (6) a. 의미전달 규칙
 $\{\text{head:}\{\text{sem:SEM}\}\}.[*] \Rightarrow \{\text{head:}\{\text{sem:SEM}\}\}.[*]$.
- b. 부정어 전달 규칙
 $\{\text{head:}\{\text{neg:NEG}\}\}.[*] \Rightarrow \{\text{head:}\{\text{neg:NEG}\}\}.[*]$.
- c. 변환사전
 $\{\text{lex:ueberwachen, deriv:ung}\}.[] \Rightarrow \{\text{lex:감시, deriv:nil}\}.[]$.
- d. 변환과정

독일어 형태 소 분석	독일어 의존 구조	한국어 의존 구조	한국어 형태소 생성
kein(Det, neg:yes) / video(N) / ueberwachen (N, deriv:ung,se m:{abstract:{ temp:_}})	ueberwachen(N, deriv:ung,sem: {abstract:{tem p:_}}, neg:yes) / video(N)	감시(N, deriv:nil,sem: {abstract:{temp :_}}, neg:yes,vsup: 하다) / 비디 오(N)	비디오(N)/ 감시 (N)/하(FV)/지 않(Neg)/음(NSu ff)

5.3. 다단계 대역어 선택방법

독일어 명사구에 대한 한국어 대역어 선택시 품사적 전이현상이나 어휘적 변이 현상등을 해결하기 위해 이 연구에서는 다단계 대역어 선택방식을 추구하였다. 다단계 대역어 선택방식이란 대역어 선택 방법을 여러계층으로 나누어 올바른 대역어 선택을 하도록 유도하는 것이다. 이와 같은 선택방법을 그림으로 나타내면 다음과 같다.



[그림 5] 다단계 대역어 선택방법

위의 [그림 5]에서 특수구조변환규칙은 구조-대-구조로 변환시키는 비조합적 변환규칙을 말하며 공통구조변환규칙은 구조-대-구조로 변환시키는 조합적 변환방식을 의미한다. 그리고 특수어휘변환규칙은 단어-대-구조나 구조-대-단어와 같은 어휘와 관련된 비조합적 변환규칙들로서 일반적으로 공기되는 속어나 기능동사구문과 관련이 있다. 그리고 마지막으로 변환사전에 의한 일반어휘변환규칙은 단어-대-단어 변환이 이루어지는 곳이다. 각 규칙들의 예를 살펴보면 다음과 같다:

- (7) a. 특수구조변환규칙:
 $\{ \}.[a b:(\text{head}:\{\text{theme:yes}\}) c] \Rightarrow \{ \}.[b, a, c].$
- b. 공통구조변환규칙: $\{ \}.[+] \Leftrightarrow \{ \}.[+].$
- c. 특수어휘변환규칙:
 $\{ \}.[\{ \text{head}:\{\text{ehead}:\{\text{concept:ass}\}\} \}, \{ \text{lex:hoch} \}] \Rightarrow$
 $\{ \}.[\{ \text{head}:\{\text{ehead}:\{\text{concept:ass}\}\} \}, \{ \text{lex:고도,cat:n} \}].$

d. 일반어휘변환규칙: {lex:wachstum} <=> {lex:성장}

(7a)의 규칙은 테마-레마에 관한 규칙으로 화제가 되는 성분은 한국어 의존구조에서 가장 앞으로 이동한다는 것을 의미한다. (7b)인 공통구조변환규칙은 '+'로 표시된 구조를 변화없이 그대로 한국어 의존구조로 복사한다는 것을 의미한다. (7c)는 독일어의 'hoch(높다)'가 Wachstum(성장)과 같은 사회·경제 현상을 나타내는 개념 단어들을 수식하게 되면 hoch는 '고도'라는 한국어로 번역된다는 것을 말한다. (7d)는 일반적인 어휘변환을 나타내는 규칙이다.

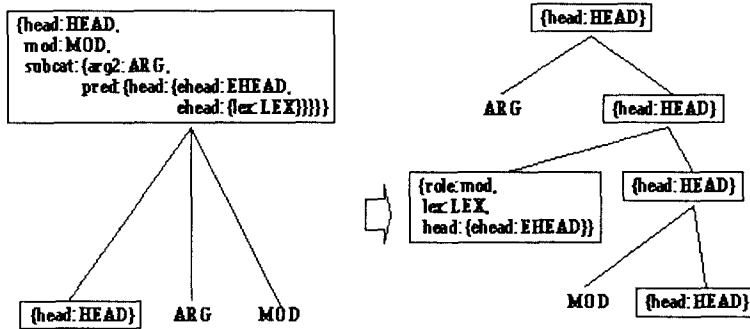
5.4. 기능동사구문의 처리

어휘적 변이와 관련해 술어명사에 숨겨졌던 기능동사가 한국어 대역어에서 나타나는 언어현상을 앞절에서 논의한 바 있다. 그에 대한 예는 다음과 같다.

- (8) Die guten Auswirkungen auf
die wirtschaftliche Dynamik
경제의 동력에 미치는 좋은 영향

(8)에서 나타난 바와 같은 언어들간의 어휘적 변이현상을 설명하기 위하여, 최승권 (1996)에서는 기계번역에서 하위범주화구조에 기능동사를 위한 자질을 만드는 새로운 방안을 제시했다.

숨겨진 기능동사가 대역어에서 나타나는 현상은 다음과 같은 변환규칙을 통하여 설명된다.



[그림 6] 기능동사구문 처리

[그림 6]의 기능동사구문을 위한 변환규칙은 독일어의 술어명사가 기능동사를 가지고 있을 경우, 한국어에서는 논항(ARG)-기능동사(EHEAD)-(수식어(MOD))-술어명사(HEAD)의 순으로 번역된다는 의미를 담고 있다.

6. 시스템의 검증을 위한 번역실험

6.1. 번역지식의 통계치

독-한 명사구 기계번역에서 작성된 번역규칙과 사전의 크기는 다음과 같다.

	독일어		변환	한국어	
	구문구조 조분석	의존구조 분석		의존구조 생성	구문구조 생성
다국어공용 번역규칙	9	19	21	19	9
언어별 특수규칙	9	15	3	8	55

[표 5] 번역규칙의 통계치

	독일어 사전	독-한 변환사전	한국어 사전
사전의 크기	25,000 단어	2000 단어	2000 단어

[표 6] 사전 목록의 통계치

[표 6]에서 번역지식을 크기를 줄이기 위하여 본 논문에서 제안한 번역규칙 모델인 다국어 공용번역규칙은 독일어와 한국어의 구문구조 분석과 의존구조 분석에 동일하게 적용된 것을 나타낸다. 이와는 대조적으로 언어별 특수규칙은 독일어나 한국어에만 적용되는 특수한 규칙의 수를 나타낸다.

다른 사전과 비교하여, 독일어 사전의 규모가 두드러지게 큰 것은 이미 오랜동안 독-영-불 기계번역을 위해 독일에서 구축된 독일어 해석 사전을 사용하기 때문이다.

6.2. 번역평가 및 번역결과

번역결과의 평가를 위해, 본 연구에서는 다음의 [표 7]과 같은 번역평가기준을 설정하여 사용하였다.

이해도	기준
5	명사구 전체의 의미가 명확하고, 명사구를 구성하는 개별표현의 의미도 명확하다.
4	명사구 전체의 의미가 명확하나, 명사구를 구성하는 개별표현의 의미는 불명확한 곳이 약간 있다.
3	명사구 전체의 의미는 파악이 가능하나, 명사구를 구성하는 개별표현의 의미는 불명확한 곳이 더러 있다.

2	명사구 전체의 의미는 몇번 읽어 보아야 알 수 있으며, 명사구를 구성하는 개별표현의 의미는 불명확한 곳이 상당히 있다.
1	여러 번 읽어도 이해하기 어렵다. 그러나 뉘센스는 아니다.
0	전혀 이해할 수 없다.

[표 7] 번역 평가 기준

수집된 독일어 명사구 코퍼스를, 기계번역시스템에 의해 자동번역한 결과를, 위에 제시된 평가기준을 기초로 기계번역에 대한 사전지식이 없는 독일어 전공 석사과정 학생 3명으로 하여금 평가하게 했다. 그 결과 번역이해도 3이상을 번역성공으로 가정하였을 때 전체 번역 성공률은 83%에 달하였다.

다음의 도표는 독일어 원문에 대한 전문 번역가의 번역과 독-한 명사구 기계번역시스템의 번역결과 예이다.

독일어 원문	전문 번역가 번역	독-한 명사구 기계번역시스템 번역
fuehrende Industrielaender	주도적 선진국	주도하는 산업국가들
schrittweise Oeffnung	단계적 개방	단계적 열림 단계적 개방
andauernder Exportboom	지속적인 수출붐	지속하는 수출 붐
hohes Wachstum	고도 성장	높은 성장
fremde Arbeitskraft	외국인 노동력	외국의 노동력 생소한 노동력
auslaendische Arbeiter	외국인 노동자	외국적 노동자

starke Nachfrage	많은 수요	강한 수요
steigende Produktivitaet	상승하는 생산력	상승하는 생산력
eine staatlich beschlossene Sache	국가적으로 결정된 사항	국가적 결정된 물건
der vereinbarte offizielle Start	합의된 공식 출발	합의된 공식적 출발

[표 8] 기계번역시스템의 번역결과

위의 비교표에서 독-한 명사구 기계번역이 번역한 결과에서 두 개의 번역결과가 나온 이유는 번역과정중에 분석결과나 변환결과가 한 개 이상의 결과가 나온 경우, 그들을 생성한 결과이다.

7. 맺음말

기계번역 분야에서의 그동안 국내외의 연구성과를 검토해 보면, 독일어-한국어 언어쌍과 관련하여, 언어학적으로 이론적인 토대가 마련되거나, 한독 기계번역에 대한 프로토타입이 마련된 수준에 머무르고 있음을 확인하게 된다.

이러한 국내외의 연구현황을 배경으로 하여, 이 연구를 통해, 독-한 명사구의 대조언어학적인 연구를 시도하고 전산언어학적인 특성을 기술하고자 했다. 이 연구에서 그 기반으로 삼고 있는 기계번역 시스템은 영어, 독일어, 불어간의 다국어 기계번역을 프로토타입 수준으로 만들었던 다국어 기계번역 시스템 CAT2이다. 이 논문에서는 다국어 기계번역 시스템 CAT2 를 토대로 하여, 구축된 독-한 명사구 기계번역의 설계구조를 기술했다. 이러한 과정에서 두 언어간의 언어정보의 공통점과 차이점을 비교,검토하는데 초점을 맞추었다. 또한 이러한 구체적인 언어학적 분석에서 출발하여 효율적인 번역지식 모델 구축을 중점적으로 기술했다.

보다 구체적으로, 이 논문에서는 기계번역 시스템 CAT2의 틀 내

에서의 독-한 명사구 기계번역의 전체 구성도가 논의되었으며, 독일어와 한국어를 동일한 번역지식을 가지고 분석 및 생성할 수 있는 번역지식 모델이 제안되었다. 더 나아가 독-한 명사구의 구조적, 품사적, 어휘적 차이를 정리.설명하고 자동번역하기 위한 언어학적 처리과정을 소개했으며, 독-한 명사구 기계번역에서 사용된 번역지식의 통계치와 번역평가기준과 번역실험 결과들을 제시했다.

이 연구에서 대상으로 한 원천언어 독일어의 명사구는 동격절과 관계절이 포함되지 않은 표현들로 한정했는데, 이 구조들을 포함하는 포괄적인 연구는 후속작업으로 남겨두고자 한다.

참고문헌

- 시스템공학연구소. 1992. 기계번역을 위한 언어모델링 및 골격시스템 개발(3), 시스템공학연구소.
- 시스템공학연구소. 1996. 영한 한영 텍스트 자동 번역 기술 개발, 시스템공학연구소.
- 이민행. 1994. “국어와 독일어의 대조통사론과 기계번역,” 독일문학 35권 1호 통권 52집, 480-507.
- 최승권. 1995. “한국어와 여러언어 통합기반 기계번역—기능동사구문을 중심으로—,” 어학연구 제31권 제1호, 139-64.
- 최승권. 1996. “한국어-독일어 자동번역,” 독일문학 37권 1호.
- 한국과학기술원. 1992. 영한기계번역시스템(III): 문법개발지원환경 및 해석문법개발.
- Choi, S. K. 1995. *Unifikationsbasierte Maschinelle Uebersetzung mit Koreanisch als Quellsprache*, Dissertation, Saarland, Germany.
- Choi, S. K., T. Kim, and D. Park. 1997. “Common and Constraint Grammar in Transfer-based Multilingual Machine Translation,” in *Proceedings of NLPRS 97*, Phuket,

Thailand.

- Grimshaw, J. 1991. *Extended projection*. Brandeis University manuscript.
- Hawkins, J. A. 1983. *Word Order Universals*. New York: Academic Press.
- Jackendoff, R. S. 1977. *X-bar Syntax: A Study of Phrase Structure*. Cambridge, Mass.: MIT Press.
- Kim, Y. T., H. G. Lee, J. H. Yang, B. R. Seo, N. R. Kim, E. K. Im and K. W. Kim. 1993. "Korean-English Machine Translation," in *Natural Language Processing Status Report*, Seoul Nat'l Univ/IBM Korea, Seoul.
- Lee, M. H. 1992. *Kontrastive Syntax und Maschinelle Sprachanalyse im Rahmen einer Unifikationsgrammatik: Untersuchungen zum Deutschen und Koreanischen*. Frankfurt a. M.: Peter Lang.
- Pollard C. and I. Sag. 1994. *Head-Driven Phrase Structure Grammar*. CSLI, Stanford University.
- Sharp, R. 1988. "CAT2-Implementing: A Formalism for Multi-Lingual MT," in *Proceedings of the 2nd International Conference on Theoretical & Methodological Issues in Machine Translation of Natural Language*, Carnegie Mellon University, Pittsburgh, PA.

이민행, 최경은

서울시 서대문구 신촌동 134

연세대학교 문과대 독어독문학과

120-749

E-mail: {leemh,choikec}@bubble.yonsei.ac.kr

최승권

대전시 유성구 어은동 1

시스템공학연구소 기계번역실

305-333

E-mail: skchoi@seri.re.kr

접수일자: 1998. 4. 15.

게재결정: 1998. 6. 29.