

# 한국어 부사격 조사와 독일어 전치사간의 기계번역문제에 대하여

홍문표 (한국전자통신연구원)

## I. 서론

본 논문에서는 한국어를 독일어로 컴퓨터를 이용하여 자동번역할 때 발생하는 한국어 부사격 조사의 처리문제를 다루고자 한다. ‘에서’, ‘으로’ 등과 같은 한국어의 부사격 조사는 그 의미의 다양성으로 인해 독일어와 같은 인도유럽어로의 번역에서 많은 문제점들을 일으킨다. 그 문제점들로 들 수 있는 것들은 첫째, 부사격 조사가 자체적으로 가지고 있는 의미적 모호성이며 (1), 둘째, 의미가 결정된 이후에도 발생하는 역어 선택의 문제이다 (2).

- (1) ㄱ. 철수가 열쇠로 문을 열었다 (도구)

Cheolswu öffnete die Tür mit einem Schlüssel

- ㄴ. 철수가 서울로 갔다 (방향)

Cheolswu ging nach Seoul

- ㄷ. 철수가 자전거로 출근했다 (교통수단)

Cheolswu fährt mit dem Fahrrad zur Arbeit

- ㄹ. 철수가 영어로 말한다 (언어)

Cheolswu spricht auf Englisch

- ㄱ. 철수가 암으로 죽었다 (원인)

Cheolswu starb an Krebs

(2) ㄱ. 그 학생이 방에서 잔다 (위치)

Der Student schläft im Zimmer

ㄴ. 그 학생이 침대에서 잔다 (위치)

Der Student schläft auf/in dem Bett

ㄷ. 그 학생이 책상에서 잔다 (위치)

Der Student schläft am Tisch

출발언어 Ausgangssprache의 형태-, 통사구조 분석을 거쳐 이를 해당하는 목표언어 Zielsprache의 구조로 변환한 다음 번역문장을 생성해내는 변환기반 기계번역방법론 Transferbasierte Maschinelle Übersetzung에서 사전 Lexikon의 역할과 비중은 점점 더 커지고 있다. 변환기반 번역방법론의 개발초기에는 통사구조 분석과 목표언어로의 변환 등을 위한 지식 Wissen으로 규칙 Regel이 사용되었다. 그러나 규칙의 수가 늘어나면 늘어날수록 규칙관리가 어려워지고 많은 규칙들이 상호 모순되는 정보 등을 담고 있는 문제점들로 인해 순수규칙기반 기계번역 방법론은 점차 그 이론적 타당성을 잃어가고 있다. 예전에는 번역을 위한 지식을 규칙이라는 틀을 통해 인코딩했다고 한다면, 최근에는 그 지식들을 사전에 담아 처리하고 있다. 이에 따라 HPSG (Head-driven Phrase Structure Grammar), LFG (Lexical Functional Grammar) 등과 같은 전산언어학 이론에서의 추세와 마찬가지로 기계번역 분야에서도 이론적 관심이 점차 규칙으로부터 사전으로부터 넘어가고 있는 추세이다. 이러한 맥락 하에서 본 논문에서는 한-독 기계번역을 위한 한국어 부사격 조사의 사전구조와 이에 기반한 독일어 전치사의 생성 방안에 대해 논의하고자 한다.

본 논문의 2장에서는 본 논의의 이론적 틀을 이루는 CAT2 형식들 Formalismus에 대해 간단히 소개한다. 3장에서는 한국어를 출발언어로 하는 기계번역 시스템의 설계에 있어, 부사격 조사 처리를 위한 기존 연구 및 문제점들을 다룬다. 4장은 3장에서 소개되는 부사격 조사의 전치사로의 번역을 위한 방법론들의 문제점을 해결할 수 있는 3단계 방법론을 제시한다. 3단계 방법론은 첫째, 부사격 조사의 의미 모호성 해소, 둘째, 유사중간언어표상 Quasi-Interlingua Repräsentation의 생성, 셋째, 생성 어휘부 Generatives Lexikon 이론에 기반한 전

치사 선택의 3단계로 구성되어 있다. 4.1장에서는 부사격 조사의 의미 모호성 해소를 위한 방법론을 제시한다. 이 방법론은 부사격 조사가 취하는 명사의 의미자질에 기반한 방법론이다. 4.2장은 부사격 조사의 전치사로의 번역을 위해서는 유사 중간언어 표현과 같은 중간단계를 거쳐야 함이 필요하다는 것을 밝히고, 유사 중간언어 표현으로 Streiter (1996)에서 제시된 Relation들을 도입한다. 4.3장에서는 생성 어휘부 이론을 간략히 소개하고 특질 구조 Qualia-Struktur가 어떻게 독일어에서 올바른 전치사의 선택을 가능하게 하는지에 대해 설명한다. 끝으로 5장 결론에서는 남은 연구 과제들을 소개한다.

## II. 한-독 기계번역을 위한 형식틀: CAT2

일반적으로 언어학 분야에서 형식틀 Formalismus이라 함은 자연언어의 문법을 기술하기 위한 형식언어를 일컫는다. CAT2 형식틀에서 단어, 구, 문장 및 언어 규칙 등과 같은 언어학적 단위들은 다른 많이 알려진 형식문법에서와 같이 ‘자질-값 구조 Merkmal-Wert Struktur’에 의해 기술된다. CAT2 형식틀에서는 컴퓨터상의 기술편의를 위해 자질-값 구조는 { } 괄호를 사용하여 표현된다.

(3) “Haus” : {phon=haus, cat=noun, gen=neut}

위의 자질-값 구조는 ‘Haus’의 정보를 담은 구조이다. 단어의 구조뿐만 아니라 구(Phrase) 구조도 { } 괄호와 [ ] 괄호를 사용하여 표기된다.

(4) S -> NP, VP : {cat=s}.[ {cat=np}, {cat=vp} ].

CAT2 형식틀 상에서 수형도 Baumstruktur의 어미매듭 Mutterknoten는 { } 괄호안에, 자매매듭 Tochterknoten는 [ ]안에 표기된다. 자매매듭의 각 개별 노드들은 { } 괄호안에 표기된다.

구구조문법 Phrasenstruktur-Grammatik의 PS규칙 Phrasenstruktur-Regel은 CAT2 형식틀 상에서 b-규칙 b-Regel에 의해 표현된다. (4)의 규칙은 S-> NP, VP 규칙

을 CAT2 표기 Notation에 따라 기술한 것이다. b-규칙이 언어구조를 만들어내는 역할을 하는 반면, f-규칙 f-Regel은 만들어진 언어구조의 적합성 Wohlgeformtheit을 검사하는 역할을 한다. 예를 들어, (4)와 같은 b-규칙을 통해 문장의 구조가 결정되었으면, 다음과 같은 f-규칙이 적용되어 주어와 술어간의 수일치 Kongruenz에 관한 제약조건을 적용할 수 있다.

(5) f\_kongruenz-constraint={cat=s}.[{cat=np}>>{agr=AGR}, {cat=vp}>>{agr=AGR}].

(5)에서 >> 연산자 Operator는 >> 왼쪽의 선행부가 만족될 경우 >> 오른쪽의 결과부도 반드시 만족해야 함을 나타낸다. 따라서 위의 규칙에 따르면 첫 번째 자매매듭의 카테고리가 np이고 (cat=np), 두 번째 자매매듭의 카테고리가 vp일 경우, AGR이라는 동일한 변수의 사용을 통해 그것들의 일치 자질(agr)에 대한 값은 일치하거나 통합 Unifikation이 가능해야 함을 알 수 있다.<sup>1</sup> 이와 같이 CAT2 형식들에서는 통합연산이 규칙적용을 위해 핵심적인 도구로서 사용된다.

b-규칙과, f-규칙을 통해 만들어진 구조는 t-규칙 t-Regel을 통해 다른 언어로 변환된다. CAT2에서 한-독 변환사전은 다음과 같은 t-규칙을 통해 기술된다.

(6) t\_haus={lex=집}.[] <=> {lex=haus}.[].  
t\_eyse={lex=에서}.[] <=> {lex=(in;an;zu;bei)}.[].

(6)의 t\_eyse 규칙에서 볼 수 있는 바와 같이, 하나의 단어에 대해 여러 개의 대역어가 대응될 수 있을 때 사전에는 ; 연산자를 사용하여 여러 개의 후보 대역어를 기술해준다. 이 중 어떤 대역어를 선택하는가의 문제는 대역어 선택의 문제이므로, 사전에는 단지 가능한 모든 대역어에 대한 정보만을 기술하게 된다.

CAT2는 형식들로서 독일어와 한국어 같은 자연언어의 문법을 기술할 수 있을 뿐만 아니라, 기계번역 시스템으로서의 기능도 할 수 있다. 즉, 문장의 입력, 분석, 변환, 생성, 출력 등을 위한 인터페이스를 제공하고 있다.

1 CAT2 형식들은 프롤로그(Prolog) 언어로 구현되어 있으므로, Prolog 프로그래밍언어와 마찬가지로 대문자로 시작하는 문자열을 변수로 사용한다.

### III. 기존연구

부사격 조사의 처리에 대한 연구는 한국어 처리 분야, 특히 기계번역 분야에 있어서 그 중요성에도 불구하고, 상대적으로 다른 테마에 비해 연구가 덜 이루어져왔다. 그 이유는 아마도 부사격 조사에 대한 전치사로의 번역이 전체 번역 문장을 이해하는데 큰 비중을 차지하지는 않기 때문으로 볼 수 있을 것이다. 그러나 고품질 번역을 지향하는 기계번역 시스템의 개발에서는 반드시 고려되어야 할 테마이다.

한국어 처리 분야에서 지금까지 이루어진 부사격 조사 처리에 대한 연구는 주로 부사격 조사의 의미 모호성 해소 분야에 집중되어 왔다. (박성배 et al. 2000) 그러나 부사격 조사의 의미 모호성 해소는 올바른 대역어 선택을 위하여 필수적이지만 반드시 정확한 대역어의 선택을 보장하지는 못한다. 즉, 예를 들어, ‘에서’의 의미가 ‘장소’의 의미로 결정되더라도 독일어로의 번역이 ‘in’일 수도 있으며, 경우에 따라서는 ‘an’, ‘zu’, ‘auf’ 등으로 번역된다.

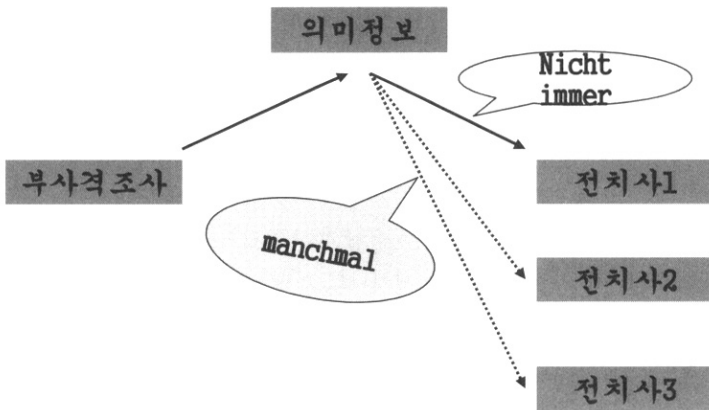


그림 1 : 의미모호성 해소후 전치사 선택의 문제

중간언어방식 Interlingua의 기계번역 시스템에서는 부사격 조사의 의미 처리를 위하여 개념 Konzept들과 개념들 간의 관계 Relation를 도입한다. 예를 들어 UNL 시스템<sup>2</sup>의 경우, 부사격 조사 ‘에서’가 나타내는 의미 표상을 위하여 plc(place)라는 관계를 사용한다. ‘방에서 자다’, ‘침대에서 자다’, ‘책상에서 자다’ 등의 의미를 나타내기 위해 각각 ‘plc(sleep(icl>event),room(icl>space))’, ‘plc(sleep(icl>event),bed(icl>furniture))’, ‘plc(sleep(icl>event),desk(icl>furniture))’ 와 같은 UNL 표기를 사용한다. 이러한 접근법에서의 문제점은 중간언어 표상에서 잠을 자는 주체 (lokalisiertes Objekt)와 잠을 자는 곳 (Referenzobjekt)의 공간적 관계가 구체적으로 나타나지 않는다는 점이다. 즉 plc라는 관계는 제1논항으로 취하는 사건 Ereignis과 제2논항간의 2차원적인 위치정보만을 나타낼 뿐, 3차원적인 공간정보는 나타나지 않는다. 이 정보는 언어쌍에 따라 전치사의 생성에 충분할 수도 있으나, 위의 예 등에서 본 바와 같이 한국어와 독일어간의 번역에서는 불충분하다. 이러한 경우에 올바른 전치사의 생성을 위해서는 plc와 같은 관계를 더욱 세분화하거나, 특정 명사들 간의 경우 올바른 전치사가 생성되게 하는 언어개별적인 모듈이 필요하다. (홍문표 1999)

일반적인 변환방식의 기계번역 Transferbasierte MÜ에서는 부사격 조사의 의미를 계산한 후 그 의미에 해당하는 전치사를 선택하는 방법이 주로 사용된다. 이 방법론의 문제점은 어떤 부사격 조사의 의미에 대해 반드시 하나의 전치사만이 대응되지는 않는다는 점이다. 또 다른 문제점은 일반적으로 대역어 선택의 단계에서 부사격 조사가 하위범주화하는 명사의 의미자질이 사용된다는 점이다. 그러나 일반적으로 기계번역 시스템에서 사용되는 [human], [furniture], [building] 등과 같은 일차원적인 의미자질만이 사용될 때, (2)의 예문들에서 볼 수 있는 ‘책상’과 ‘침대’의 의미자질을 상이한 전치사의 선택을 위하여 어떻게 차별화 할 수 있는가하는 문제가 대두된다.(Palmer et al. 1999)

동사패턴 기반 방법론에서도 유사한 문제가 발생한다. 이 방법론에서는 개별 부사격 조사가 패턴 내에서 어떠한 의미를 지니는 명사와 같이 사용되느냐에 따라 대역어가 결정되며, 변환기반 방법론에서와 마찬가지로 명사에 대해 단순 의미코드가 사용되므로 변환방식에서 발생하는 문제와 동일한 한계에 부딪힌다. 이

2 UNL은 Universal Networking Language의 약자임

러한 일차원적 의미기술의 문제를 해결하기 위해 어휘레벨에서의 모호성 해소 및 대역어 선택방법이 시도되지만, 무수한 패턴이 필요로 되는 데이터 부족현상의 문제가 있다. (Kim et al. 2002)

4장에서는 이러한 문제점들을 해결할 수 있는 3단계 방법론이 소개된다.

#### IV. 부사격 조사의 번역을 위한 3단계 방법론

한국어에서 독일어 혹은 영어와 같은 유럽어로의 기계번역에서 부사격 조사는 많은 경우에 다음과 같은 일반적인 변환 규칙에 의해 처리된다.

- (7) {lex=으로} <=> {lex=(nach;zu)}
- (8) {lex=에서} <=> {lex=(in;an;zu;bei)}

한국어나 일본어의 조사 및 독일어 등의 전치사와 같은 기능어 Funktionswort 들은 일반적으로 위의 변환 규칙 등과 같은 단순한 변환 규칙에 의하여 번역이 이루어져왔다. 이러한 직접 변환 방식은 한국어와 다른 유럽어에서 뿐만 아니라 영어와 독일어 같은 유럽어 사이에서도 사용되었다. 그러나 이러한 직접 변환 방식의 문제점은 Luckhardt (1987)에서 제시된 다음의 영어와 독일어 예문에서 명백히 드러난다.

- (9) ㄱ. He sent us a message from the USA  
     ↳ Er schickte uns eine Nachricht aus den USA
- (10) ㄱ. He sent us a message from the moon  
     ↳ Er schickte uns eine Nachricht vom Mond

위의 예에서 볼 수 있듯이 영어의 전치사 ‘from’은 독일어에서 ‘aus’ 혹은 ‘von’ 등으로 번역된다. 마찬가지로 한국어의 부사격 조사도 (2)의 예문들에서 볼 수 있는 바와 같이 조사가 취하는 명사에 따라 각각 다른 전치사로 번역된다.

부사격 조사와 전치사가 일대다 eins-zu-viele의 관계를 이루기 때문에 본 논문

에서는 부사격 조사가 독일어의 전치사로 직접 대응되는 방식이 아니라, 의미 모호성 해소의 과정을 거쳐 유사중간언어 표현 Quasi-Interlingua Repräsentation으로 분석이 된 후, 올바른 전치사 선택의 과정을 거쳐서 최종적으로 하나의 전치사가 선택되는 3단계 방법론을 제시한다.

본 논문에서 제시하는 3단계 방법론은 부사격 조사의 번역에만 국한되는 것은 아니다. 이미 홍문표 (2004)에서 보인 바와 같이 일반적인 명사의 대역어 선택을 위해서도 의미 모호성을 해소하고, 계산된 의미에 해당하는 후보 대역어군에서 문맥에 가장 알맞은 대역어를 선택하는 방법론이 적용된다. 문맥에 가장 적합한 대역어를 선택하는 방법으로는 코퍼스에서 통계정보를 추출하여 가장 확률값이 높은 대역어를 선택하는 경험적인 방법론 empirische Methode과 본 논문에서 제시하는 사전 지식을 사용하는 합리적인 방법론 rationale Methode 두가지가 있으며, 한-독 기계번역에서는 두가지를 모두 사용하는 복합적인 방법론이 응용될 수 있다.

### 1. 부사격 조사의 의미 모호성 해소

‘방에서’, ‘책상에서’, ‘집에서’ 등과 같은 ‘명사+부사격 조사’의 구조에서 부사격 조사는 핵심어 Kopf로서 명사를 보족어 Komplement로 취하는 것으로 간주된다. 그 이유는 첫째, 부사격 조사는 격조사와는 달리 ‘장소’, ‘도구’, ‘방향’ 등과 같은 의미 정보를 지니기 때문이다. 둘째로는 부사격 조사가 자신의 보족어로 취하는 명사를 의미적으로 제약한다는 점을 들 수 있다.

(11) 그 학생이 학교로 간다

(?) 그 학생이 꿈으로 갔다

‘방향’의 의미를 지니는 부사격 조사 ‘으로’가 의미적으로 특정한 속성을 지니는 명사들과만 결합할 수 있다는 정보는 ‘으로’의 사전에 다음과 같이 하위범주들에서 명사들에 대한 의미적 제약을 통해 표기될 수 있다.



(12)

```
{lex=으로, head={cat=advp}, subcat={a={head={cat=(n;d), sem={{BUILDING};
{REGION};{CITY};{INSTITUTION}}}}}}
```

‘으로’의 사전 정보에 따르면 이 부사격 조사는 ‘building’, ‘region’, ‘city’, ‘institution’ 등의 의미 타입을 가진 명사들만을 하위 범주화한다. 위 사전 엔트리에서 BUILDING, REGION, CITY, INSTITUTION 등은 일종의 의미적 매크로 Makro로서 다음과 같은 정보들을 내포하고 있다.

(13)

```
define (BUILDING, 'ag=nil, th~=(move), gl~=(perc;info;act;emot;comm.;exc;asp),
ccr={t=building, sex=nil})
```

예를 들어 BUILDING 매크로는 구체물 (concrete)이면서 의미적 타입은 ‘building’이며 어떤 동사의 행위자 Agens로는 사용될 수 없고 (ag=nil), 대상역으로 사용될 때는 이동동사의 경우를 제외하면 가능하고 (th~=move), 방향역으로는 지각, 정보, 행위, 감정, 통신, 교환, 양상의 의미를 지니는 동사들과는 사용될 수 없음을 의미한다 (gl~=(perc;info; act;emot;comm.; exc;asp)).<sup>3</sup>

위에서 제시한 ‘으로’와 보족어 명사들간의 결합은 다음과 같은 통사규칙에 의하여 제약된다.

(14)

```
b_func_subcat= {role=ROLE, head=HEAD, subcat=nil},{head={cat=CAT,
sem=SEM}}, {role=ROLE, head=HEAD, subcat={a={head={cat=CAT,
sem=SEM}}}}}
```

보족어로 취하는 명사에 대한 의미 제약은 조사의 하위범주화를 통해 이루어

3 명사의 의미자질을 이를 취하는 동사의 프로세스 타입에 따라 분류하는 방법론에 대해서는 Streiter (1998), 홍문표 (2004)를 참조하기 바란다.

지고 또 부사격조사의 모호성 해소도 이러한 방식으로 이루어진다. 즉, ‘으로’가 ‘방향’을 나타내는지 혹은 ‘도구’의 의미로 쓰이는지의 여부는 보족어로 취하는 명사들의 의미자질에 따라 결정된다.

## 2. 유사 중간언어 표상 (Quasi-Interlingua Repräsentation)

우리는 이미 영어와 독일어간에서 전치사의 번역시 어휘레벨에서 일대일 대응 규칙을 통한 변환방법은 많은 문제가 있음을 예문을 통하여 보았다. 이점은 바로 부사격 조사와 전치사간의 번역에도 똑같이 적용될 수 있을 것이다. 어휘 레벨에서의 직접 변환 방식을 통해 전치사간의 변환 문제를 해결하려는 시도는 그 밖에 더 심각한 문제를 간과하는 오류를 범할 수 있다. 이 점은 Buschbeck & Nübel (1995)에서 다음과 같이 지적되었다:

“이러한 변환규칙(직접변환을 위한 규칙)을 사용함으로써 다른 현상을 다루는 데에 필수적인 공간 정보를 잃게 될 수 있다. 예를 들어 수식하는 전치사구의 의미 분석에 대한 정보는 동사나 명사의 의미모호성 해소를 위하여 응용될 수 있다. 또한 그 정보는 시제와 양상의 처리에 있어서도 영향을 미친다. 그러나 전통적 의미의 변환규칙은 출발언어와 목표언어 단어간의 번역관계에 대한 어떠한 설명도 하지 못하며, 단지 두 단어를 연결시켜줄 뿐이다...”<sup>4</sup>

Streiter (1996)에서는 영어와 독일어의 전치사 번역을 위해 12가지 관계를 설정하고 이를 중간언어 표현으로 사용하고 있다. 본 논문에서도 한국어의 부사격 조사가 나타내는 의미를 기술하기 위해 다음의 12가지 관계 Relation를 도입하기로 한다. 이 관계에 해당하는 한국어의 부사격 조사와 영어와 독일어의 전치사의

4 “...Bei der Verwendung derartiger Transferbedingungen geht die Information über die eigentliche denotierte Relation verloren, die aber für die Behandlung anderer Phänomene unbedingt notwendig ist. Die Information über die Lesart von Modifikator PPn kann beispielsweise für die lexikalische Desambiguierung von Verben oder Nomen herangezogen werden, ebenso nimmt sie Einfluss auf Verarbeitung von Tempus und Aspekt. Transferregeln im traditionellen Sinne liefern keine Erklärung für die Übersetzungsrelationen zwischen quell- und zielsprachlichen Lexemen, sie setzen lediglich zueinander in Beziehung....”

대응관계를 표1에 소개한다.

이에 따라 모든 부사격 조사는 사전에서 ‘rel’ 자질에 대해 기술된다. ‘rel’ 자질이 첨가된 방향을 나타내는 ‘으로’와 장소를 나타내는 ‘에서’의 수정된 사전구조는 다음과 같다.

(15)  
 {lex=으로, rel='DIR', head={cat=advp}, subcat={a={head={cat=(n;d), sem={({BUILDING}; {REGION};{CITY});{INSTITUTION}})}}}

(16)  
 {lex=에서, rel='LOC', head={cat=advp}, subcat={a={head={cat=(n;d), sem={({BUILDING}; {REGION};{CITY});{INSTITUTION}})}}}

관계	한국어	독일어
'SPACE'	으로부터, 까지	ab, an, auf
'FUNCTION'	으로, 로서	als
'DIR'	으로, 에	an, in, zu, zwischen, über
'LOC'	에, 에서	an, auf, in, zu, zwischen, über
'TEMP'	에, 까지	um, bis
'TRANSPORT'	으로	mit
'CAUSE'	으로, 때문에	aufgrund, aus, bei
'USE'	으로	mit
'QUAL'	으로	auf
'EQUAL'	처럼	wie
'POSS'	의	von
'SET'	없이, 와	ohne, mit

표 1 : 한-독 기계번역을 위한 유사중간언어표현

이러한 유사 중간언어 표현의 단계를 거치므로 부사격 조사가 독일어에서 하나의 전치사로 직접 번역되는 것이 아니라 독일어에서도 이 표상에 해당하는 전치사들의 집합으로 사상 Abbildung된다.

예를 들어, ‘집에서’의 경우 핵심어 ‘에서’가 유사 중간언어 표상으로 ‘LOC’값을 지니므로 목표 언어에서 이 표상에 해당하는 독일어 전치사 ‘an’, ‘auf’, ‘in’, ‘zu’ 등이 대역어 후보로 선택된다.

부사격 조사의 변환을 위해 ‘rel’ 자질을 도입했으므로 이에 따라 (14)의 규칙도 다음과 같이 수정된다. 즉, ‘집에서’, ‘학교로’ 등과 같은 부사격 조사구의 경우, 부사격 조사가 가지고 있는 ‘rel’-값이 어미매듭에서도 그대로 승계되어야 한다.

(17)

```
b_funct_subcat= {role=ROLE, rel=REL, head=HEAD, subcat=nil,}{head={cat=CAT, sem=SEM}}, {role=ROLE, rel=REL, head=HEAD, subcat={a={head={cat=CAT, sem=SEM}}}}
```

위 규칙에 의해 ‘집에서’, ‘집으로’ 등의 부사격 조사구가 각각 ‘LOC’, ‘DIR’ 등의 rel 값을 갖게 되고, 이 값이 독일어 변환구조로 전이된다.

### 3. 생성 어휘부 이론에 기반한 부사격 조사의 대역어 선정

4.2장에서 소개된 유사 중간언어 표상 Quasi-Interlingua Repräsentation에 의해 분석단계가 끝난 부사격 조사에 대해 대역어에서 하나 이상의 대역어 후보가 선택된다. 예를 들어 ‘집에서’의 경우, 부사격 조사 ‘에서’가 rel값으로 ‘LOC’을 갖기 때문에 이에 대한 대역어 후보로서 ‘an’, ‘auf’, ‘in’, ‘bei’ 등이 선택된다. 따라서 독일어에서 올바른 전치사의 선택을 위하여 유사 중간언어 표상만으로는 부족하며 좀 더 풍부한 정보가 필요함을 알 수 있다. 이에 대한 논의를 위하여 (2)의 예문을 다시 소개한다.

(2) ㄱ. 그 학생이 방에서 잔다 (위치)

Der Student schläft im Zimmer

ㄴ. 그 학생이 침대에서 잔다 (위치)

Der Student schläft auf/in dem Bett

ㄷ. 그 학생이 책상에서 잔다 (위치)

Der Student schläft am Tisch

위의 예문에서 알 수 있는 것은 ‘위치’의 의미를 내포하는 부사격 조사 ‘에서’가 쓰임에 따라 그 의미가 모호하다는 점이다. (2.ㄱ)에서 ‘에서’로 나타내지는 주어 ‘학생’과 장소 ‘방’의 관계는 ‘학생’이 ‘방’이라는 3차원 공간 안 (in)에 위치해 있다는 것이고, (2.ㄴ)에서 ‘학생’은 ‘침대’의 위 (auf)에 위치하면서 ‘침대’와의 물리적인 접촉을 갖고 있으며, (2.ㄷ)에서 ‘학생’은 ‘책상’의 근처 (an) 혹은 위에 위치한다는 것이다.

이러한 의미적 모호성을 해결하기 위해서는 다음의 두가지 질문에 대한 대답을 찾아야 할 것이다.

- i) (2.ㄱ~ㄷ)에서 나타나는 의미적 모호성을 해소할 수 있는 한국어 모국어 화자의 언어 능력은 무엇인가?
- ii) 이러한 언어 능력이 기계 번역에 적용되기 위하여 어떻게 형식화 되어야 하는가?

‘방에서’, ‘침대에서’, ‘책상에서’의 경우 독일어에서 전치사의 선택에 영향을 미치는 것은 부사격 조사가 아니라 이에 의해 선택되는 명사임을 알 수 있다. 따라서 위에서 제기된 첫번째 질문에 대한 가능한 대답은 명사와 부사격 조사간의 연결관계 Kollokation 정보를 이용한 해결 방법일 것이다. 이 방식은 간단히 말하면, 부사격 조사 ‘에서’는 명사 ‘방’과 결합하면 ‘in’으로 번역되고, ‘침대’의 경우는 ‘auf’으로 번역되고, ‘책상’의 경우는 ‘an’으로 번역된다는 규칙을 상정하는 것이다. 그러나 이러한 접근법은 언어학적 일반성도 지니지 못하며 지식 획득의 문제도 크기 때문에 좀 더 일반적이고 언어학적으로 타당한 방법론이 필요하다.

다른 한 가지 방법은 부사격 조사에 의해 취해지는 명사들의 의미 자질에 의존하는 방법이다. 즉, 조사 ‘에서’가 ‘공간’의 의미자질을 나타내는 ‘방’과 같은 명사와 결합하면 ‘in’으로 번역되고, ‘가구’의 자질을 가지는 ‘침대’와 결합하면 ‘on’으로 변환되게 하는 것이다. 그러나 이러한 단순 의미자질에 의존하는 방식은 많은 문제점을 내포하고 있다. 가장 큰 문제점은 일반적인 명사들의 의미분류

는 'is-a' 관계에 의존하기 때문에 많은 경우에 변별력이 떨어진다는 점이다. 위의 예에서 '침대'와 '책상'을 의미적으로 구분하는 개념적 차이는 이러한 기술방식에서 어떻게 형식화 될 수 있는가가 큰 문제점으로 대두된다. 물론 올바른 역어 선택을 위하여 '책상'과 '침대'를 구별해주는 어떤 하나의 자질을 임의로 도입할 수 있을 것이다. 그러나 이러한 접근방법은 언어학적 일반성이 결여되어 있다고 할 수 있다.

이러한 문제점을 해결하기 위해 이 논문에서는 이미 많은 언어들에 대해 적용되었고 역어 선택의 문제만이 아닌 많은 다른 언어현상들의 설명에 있어서 그 타당성을 검증받은 Pustejovsky (1995)의 생성 어휘부 이론 Generatives Lexikon을 도입한다. 생성 어휘부 이론의 가장 큰 장점은 자연언어의 창조적인 용법과 규칙적인 다의 현상을 다계층 어휘목록과 몇 개의 생성 메커니즘으로 자연스럽게 설명할 수 있다는 데에 있다. 생성 어휘부 이론에서 어휘의 의미는 사건구조 Ereignisstruktur, 특질구조 Qualia-Struktur 그리고 논항구조 Argumentstruktur 등 3개의 레벨에서 기술된다. 이 중 부사격 조사의 대역어 선택을 위해 주목하는 구조는 명사의 특질구조 Qualia-Struktur이다.

#### (18) 특질구조 Qualia-Struktur

```
{constitutive ...
  telic ...
  agentive ...
  formal ... }
```

특질구조에는 일반적으로 사람이 어떠한 사물에 대해 설명하고자 할 때 사용하는 네 가지의 양상들이 반영된다. 즉, 첫째, 무엇으로 구성되어 있는가 (constitutive), 둘째, 누가 만들었는가 (agentive), 셋째, 어떤 카테고리에 속하나 (formal), 넷째, 용도가 무엇인가 (telic)의 네 가지 양상들이 그것이다.

이 중 특히 부사격 조사의 번역에서 사용될 수 있는 자질은 기능역 telic이라고 할 수 있다. 명사들의 기능역에 대해 Pustejovsky (1995)는 직접 기능역 direct telic과 목적 기능역 purpose telic을 구별한다. 예를 들어, '맥주'라는 단어는 '마

시다'라는 동사가 지시하는 행위의 직접적인 대상이 되고, '망치'라는 명사는 '때리다'라는 동사가 지시하는 행위의 직접적인 대상이 되는 것이 아니라 때리는 행위의 도구가 된다. 이러한 이론적 토대 위에 '방', '침대', '책상'의 부분적인 의미구조를 살펴보면 다음과 같다.<sup>5</sup>

(19)

```
{lex=방, argstr={arg=x:space}, qualia={formal=x, telic=(sleep(e,x,y); read(e,x,y);
...))}}
```

(20)

```
{lex=침대, argstr={arg=x:furniture}, qualia={formal=x, telic=(sleep(e,x,y); read(e,x,y);
...))}}
```

(21)

```
{lex=책상, argstr={arg=x:furniture}, qualia={formal=x, telic=(sleep(e,x,y); read(e,x,y);
...))}}
```

특질 구조가 첨가된 '방', '침대', '책상'의 의미구조는 이 단어들이 지시하는 개체, 즉 x가 어떤 행위, 예를 들어 'schlafen'을 위한 도구가 됨을 나타내지만, 이러한 행위의 주체 y와 x가 어떠한 공간적 관계를 맺는지는 나타내지 못한다. 본 논문에서는 기능역 telic 구조에서 주체와 대상간의 공간적 관계를 나타내기 위해 'reln' 이란 자질을 도입한다. 이 'reln' 자질은 목적 기능역 purpose telic을 갖는 모든 명사들의 경우에 기술된다. 'reln' 자질에 대한 값으로는 앞 장에서 도입한 유사 중간언어 표상인 'rel'의 값이 'LOC'을 제외한 'FUNCTION'이나 'TRANSPORT' 등이면 'rel'의 값과 동일하다. 그러나 'rel'의 값이 'LOC'이면 'reln'의 값은 'top', 'near' 혹은 'empty'를 갖는다.

5 (19)-(21)에 보인 사전구조는 실제 CAT2에서 사용하는 사전구조가 아니라 논지전개의 편의상 단순하게 표현한 사전구조임을 밝힌다

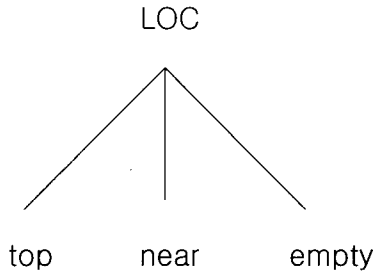


그림 2 : 'LOC'의 Spezifikation

'top'값은 y가 x의 위에 위치할 때, 'near' 값은 y가 x의 근처에 위치할 때, 'empty'값은 y가 x의 안에 위치할 때 reln의 값으로 사용된다. 이러한 정보는 다음과 같은 방식으로 CAT2 사전 구조에 반영될 수 있다.

(22)

```
{lex=방, qualia={telic={reln=empty, arg1=_, arg2=(study;read;sleep)}}}
```

(23)

```
{lex=침대, qualia={telic={reln=top, arg1=_, arg2=(study;read;sleep)}}}
```

(24)

```
{lex=책상, qualia={telic={reln=near, arg1=_, arg2=(study;read;sleep)}}}
```

'공간'이라는 의미를 갖지만 주체와 대상간의 구체적인 공간적 관계에 대해서는 모호한 '에서'와 같은 부사격 조사에 대해 이제 하위 범주화하는 명사의 기능역 (telic)의 reln정보를 사용하여 정확한 대역어를 선택할 수 있다. 즉, '에서'가 reln 값이 'empty'인 명사를 하위 범주화 하면 독일어에서는 'in', 영어에서도 'in'으로 번역되고, reln 값이 'near'인 명사를 하위 범주화하는 경우에는 독일어에서는 'an', 영어에서는 'at'으로 번역될 수 있고, 'top'을 하위 범주화 할 때는 독일어에서는 'auf', 영어에서는 'on'로 번역된다. 이를 통합기반 기계번역 시스템 CAT2의 형식들을 사용하여 표현하면 다음과 같다.



(25)

$t\_reln\_top = \{rel='LOC', \quad lex=에서, \quad head=\{cat=advp\},$   
 $subcat=\{a=\{head=\{sem=\{qualia=\{telic=\{reln=top\}\}\}\}\}\} \Leftrightarrow \{rel='LOC', \quad lex=auf,$   
 $head=\{cat=p\}\}$   
 $t\_reln\_near = \{rel='LOC', \quad lex=에서, \quad head=\{cat=advp\},$   
 $subcat=\{a=\{head=\{sem=\{qualia=\{telic=\{reln=near\}\}\}\}\}\} \Leftrightarrow \{rel='LOC', \quad lex=an,$   
 $head=\{cat=p\}\}$   
 $t\_reln\_empty = \{rel='LOC', \quad lex=에서, \quad head=\{cat=advp\},$   
 $subcat=\{a=\{head=\{sem=\{qualia=\{telic=\{reln=empty\}\}\}\}\}\} \Leftrightarrow \{rel='LOC', \quad lex=in,$   
 $head=\{cat=p\}\}$

(22)와 같이 reln 값이 포함된 사전구조와 (25)의 변환규칙에 따라 ‘침대에서’는 ‘auf dem Bett’, ‘책상에서’는 ‘an dem Tisch’, ‘방에서’는 ‘in dem Zimmer’로 각각 변환된다.

## V. 결론

본 논문은 한국어의 부사격 조사를 다국어 기반 기계번역 시스템에서 독일어로 번역할 때 생기는 대역어 선택의 문제에 대해 다루었다. 대역어 선택의 문제는 부사격 조사의 의미적 모호성과 의미적 모호성이 해소된 이후에도 발생하는 대역어 선택 단계에서의 언어적 자의성에 기반한다. 본 논문에서는 부사격 조사의 의미적 모호성 해소를 위하여 명사의 의미자질에 기반한 방식을 제안하였으며, 대역어 선택의 단계에서는 Pustejovsky (1995)의 생성 어휘부 Generatives Lexikon 이론을 도입하여 문제를 해결하고자 했다. 특히 ‘에서’와 같이 공간의 의미는 나타내지만 주체와 대상간의 구체적인 공간적 관계에 대해 모호한 *unterspezifiziert* 부사격 조사의 처리를 위하여 생성 어휘부 이론의 특질 구조 *Qualiastruktur*를 도입할 수 있음을 보였다. 그러나 기존의 특질 구조로는 역시 완전한 처리에 문제가 있음을 밝히고 기능역 *telic*에 ‘reln’이란 자질을 도입하여 명사가 내포하는 주체와 대상간의 공간적 관계를 사전에서 제시하는 방법을 통해 문제를 해결할 수 있음을 밝혔다. 또한 이러한 이론적 해결방안을 CAT2 형식들

내에서 구현하므로써 전산적으로도 처리 가능성을 보였다.

앞으로의 연구과제로는 대규모 엔트리에 대한 풍부한 사전 정보의 반자동 획득 방안을 들 수 있다. 단순한 프로토타입 Prototyp 기계번역 시스템의 개발이 목적이 아니라, 상용화를 목적으로 하는 기계번역 시스템의 개발을 목적으로 할 경우, 최소한 20만 단어 이상의 엔트리를 갖는 사전의 구축이 필수적이다. 이 중 대략 10만 단어 이상이 명사라고 할 때, 이들 명사들에 대한 특질구조의 반자동 획득 방안은 본 논문에서 제시하는 해결책의 실제 개발 단계에서의 적용 가능여부를 가름하는 가장 중요한 요소라고 할 수 있다. 10만개 이상의 명사들에 대해 특질구조를 수동으로 부착하는 것은 개발에 있어 많은 비용과 시간을 필요로 할 것이 분명할 것이기 때문이다.<sup>6</sup> 따라서 명사들의 기능역 telic 정보가 본 논문에서 제시된 해결책에서 핵심적인 역할을 한다고 할 수 있는데, 이에 대한 반자동 지식 획득 여부가 매우 중요한 추후 연구과제라고 할 수 있다.

## 참고문헌

- 박성배, 장병탁, 김영택 (2000). 의미 부착이 없는 데이터로부터의 학습을 통한 의미 중의성 해소, 한국정보과학회 봄 학술 발표 논문집 (B) 27권 1호
- 홍문표 (1999). UNL-프로젝트에 대하여 - 독/영 기계 번역 시스템에 기반한 독일어 UNL-분석기의 구현을 중심으로, 독일문학 71집, 40권 3호
- 홍문표 (2004). 한-독 기계 번역의 대역어 선택 문제에 대하여, 독일문학 91집, 45권 3호
- Buschbeck-Wolf, B. & R. Nübel (1995): Die Behandlung übersetzungsmäßiger Präpositionen im Transfer des VERBMOBIL-Demonstrators, Verbmonil Report 87
- Kim, C.H., M. Hong, Y. Huang, Y. Kim, S. Yang, Y. Seo, & S. Choi (2002): Korean-Chinese Machine Translation based on Verb Patterns, in the Proceedings

6 명사의 특질구조를 반자동으로 획득하는 방안은 아직 본격적으로 연구되고 있지는 않지만, 기술적으로 충분히 현실성이 있다. 특히 정보통신부의 지원하에 2005년도에 진행된 “표준형 한국어 언어/음성DB” 사업의 결과물 중 50만개 규모의 용언-부사격조사구 언어 대역패턴을 이용할 경우 상당히 정확률이 높은 특질구조를 반자동으로 추출할 수 있을 것으로 기대된다.

- of the Conference of the Association for Machine Translation in the Americas (AMTA 2002)
- Hong, M.P. & O. Streiter (1999): Overcoming the Language Barriers in the Web: The UNL Approach, in Tagungsband der 11. Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung (GLDV'99), Frankfurt am Main
- Hong, M.P. (2001): Linguistische Probleme in der Maschinellen Übersetzung Koreanisch-Deutsch, Ph.D Dissertation, Universität des Saarlandes
- Luckhardt, H-D. (1987): Der Transfer in der maschinellen Sprachübersetzung, Sprache und Information, Band 18, Niemeyer, Tübingen
- Pustejovsky, J. (1995): The Generative Lexicon, The MIT Press, Cambridge/London
- Streiter, O. (1996): Linguistic Modeling for Multilingual Machine Translation, Shaker Verlag
- Streiter, O. (1998): A Semantic Description Language for Multilingual NLP, in Tuscan Word Center Institut für Deutsche Sprache Workshop on Multilingual Lexical Semantics

## Zusammenfassung

### Die Behandlung der adverbialen Postpositionen in der Maschinellen Übersetzung Koreanisch-Deutsch

Munpyo Hong (ETRI)

Die vorliegende Arbeit beschäftigt sich mit der maschinellen Übersetzung der adverbialen Postpositionen in die deutschen Präpositionen. Die adverbialen Postpositionen entsprechen in der Regel den Präpositionen im Deutschen. Das Hauptproblem bei der Übersetzung der adverbialen Postpositionen liegt darin, dass die adverbialen Postpositionen Ambiguitäten aufweisen. Zur Disambiguierung der adverbialen Postposition wird die Semantic Description Language (SDL) von Streiter

(1998) herangezogen. Es soll gezeigt werden, dass die SDL zwar in der Lage ist, die grobe Disambiguierung der adverbialen Postposition durchzuführen, aber bei den feinen Bedeutungsunterschieden der adverbialen Postpositionen an ihre Grenze stößt. Aus diesem Grunde wird die SDL dahingehend erweitert, dass die sogenannte Qualia-Struktur aus der 'Generatives Lexikon'-Theorie von Pustejovsky (1995) bei der Beschreibung der adverbialen Postpositionen mit einbezogen wird. Nach der Disambiguierung wird eine Quasi-Interlingua Repräsentation für die zugrundeliegende Bedeutung erzeugt. Jede Präposition im Deutschen ist im Lexikon an eine Quasi-Interlingua Repräsentation gekoppelt, so dass die Generierung der deutschen Präposition einfach eine Auswahl der passenden Präposition darstellt. Die Qualia-Struktur konnte zur Übersetzung der adverbialen Postpositionen in die Präpositionen angewandt werden.

주제어: 한-독 기계번역, 부사격조사, 생성어휘부, 대역어선택

Schlüsselbegriffe: Koreanisch-Deutsch MÜ, adverbiale Postposition, Generatives Lexikon, Auswahl der Zielsprache

필자 이메일 주소: munpyo@etri.re.kr

접수일: 2005. 12. 31 / 심사일: 2006. 1. 24 / 심사완료일: 2006. 2. 21