독일언어문학 제30집(2005 · 12). 39-60

한-독 기계번역상의 연어처리 문제에 대해

홍 문 표(ETRI)

1. 서 론

수십년동안 프레게(Frege)의 합성성의 원리(Kompositionalitätsprinzip)는 우리가 언어의 의미와 의미합성의 매커니즘을 이해하는데 가장 핵심적인 근간으로 여겨져 왔다. 이와 유사하게 기계번역(Maschinelle Übersetzung, MÜ)에서도 다음과 같은 '번역합성성의 원리 (Kompositionalitätsprinzip der Übersetzung)'는 번역 프로세스를 이해하는데 중요한 매커니즘으로 여겨진다.

> "Two expressions are translation-equivalent if they can be built up from parts which are translation-equivalent by means of translation-equivalent rules" Landsbergen(1998)

현존하는 대부분의 규칙기반 기계번역(regelbasierte MÜ) 방법론[1]에서는 위와 같은 번역 합성성의 원리에 따라 어떠한 언어표현의 번역을 위해 그 언어표현을 구성하는 단위 언어표현의 번역을 선행해야 하는 bottom-up 방식의 번역방식을 따른다. 즉, 문장의 번역을 위해서는 문장을 이루는 명사구와 동사구의 번역이 선행되어야 하며, 동사구의 번역을 위해서는 그 동사구를 이루는 동사와 명사의 번역이 선행되어야 하는 것이다. 이렇게 번역된 단위들은 그 단위들을 결합하는 통사규칙에 대응하는 번역규칙에 의해 번역된다.

모든 언어표현이 이와 같이 합성성의 원리에 따라 기계적으로 처리될 수만 있다면 이미 언어간의 장벽은 상당수 해소되었을 것이다. 그러나 우리의 언어는 그 언어가 사용되는 사회, 자연환경, 문화, 역사, 화자 등과 같은 요인에 의해 이와 같은 번역 합성성의 원리를 허용하지 않는 많은 표현들을 내포하고 있다. 우리가 흔히 연어(Kollokation)라고 부르는 언어표현이 이의 대표적인 예라고 볼 수 있다.

---

1) 여기서 언급하는 규칙기반 기계번역 방법론이란 통계정보와 대용량 코퍼스와 같은 경험적 지식에 의존하는 통계기반 기계번역 방법론 (Statistik-basierte MÜ), 예제기반 기계번역 방법론 (Fallbasierte MÜ) 등에 상반되는 개념으로서, 협의의 규칙기반 기계번역 방법론, 지식기반 방법론, 중간언어방식 기계번역 방법론 등이 여기에 속한다고 볼 수 있다.

본 논문에서는 한독 기계번역에서 나타나는 연어의 문제를 다루고자 한다. 일반적으로 기계번역을 포함한 자연언어처리 분야의 연어 관련 연구는 크게 두 가지 방향으로 생각할 수 있다.

1) 연어의 개념을 어떻게 규정할 것이며, 어디까지를 연어로 볼 것이며, 연어 데이터를 어떻게 추출할 것인가의 문제[2)]
2) 연어 관계를 주어진 이론적 틀 내에서 어떻게 파악, 기술할 것인가에 관한 문제

본 논문의 연구 방향은 두 번째 문제, 즉, 기계번역을 위해 고안된 형식문법 틀 내에서 어떻게 한국어의 연어 관계를 기술할 것이며, 이렇게 기술된 연어표현을 독일어로 변환할 것인지에 관한 것이다.

한독 기계번역에서 연어표현들은 번역합성성의 원리를 따르지 않으므로, 연어표현 전체가 하나의 번역단위로서 다루어져야 한다. 연어표현이 하나의 번역단위로 다루어지지 않으면, 연어표현을 이루는 개별 단어들이 각각 번역되어 목표언어(Zielsprache), 즉, 한독 기계번역의 경우 독일어에서는 독일어의 자연스러운 표현이 아닌 어색하거나 잘못된 번역을 생성하게 된다.

본 논문에서는 한독 기계번역에서의 한국어 연어표현 분석과 독일어로의 변환 등을 위해 Streiter(1996)에서 소개된 기능동사(Funktionsverbgefüge) 개념을 도입한다. 또한 이러한 방법론을 사용할 경우 단순히 한국어와 독일어간의 번역만이 아닌, 한국어, 독일어, 영어, 일본어 등과 같이 다국어를 동시에 번역해야하는 상황에서 어떠한 문제가 발생하는지도 살펴보게 될 것이다. 본 논문에서 주장하는 방법론을 구현하기 위한 이론적인 틀로써는 CAT2 형식틀(Formalismus)이 사용되었다. 한독 기계번역상의 연어처리문제를 직접 다루기에 앞서 CAT2 형식

---

2) Lewandowski(1985)에 따르면 연어의 개념은 "die Möglichkeit und die Wahrscheinlichkeit des gemeinsamen Vorkommens lexikalischer Einheiten in einem Syntagma aufgrund syntagmatischer Beziehungen zwischen lexikalischen Einheiten"와 같이 정의되었다. 본 논문에서 정의하는 연어의 개념은 쭈어-동사, 목적어-동사와 같이 특정한 문법적 관계에 있으면서, 대용량 코퍼스를 분석하였을 때 다수 출현하는 어휘-어휘의 관계로 이해될 수 있다. 물론 '다수출현'이라는 모호한 개념이 개념 정의에 사용되는 문제점이 있지만, 특정 크기의 코퍼스를 분석할 시, 다른 어휘-어휘쌍보다 두드러지게 많이 발견되는 특정 어휘-어휘쌍을 말한다고 볼 수 있다. 여기에서 '목이 마르다', '배가 고프다' 등과 같은 숙어적표현과 '그네를 타다' 등과 같이 숙어적 표현은 아니지만 다른 어휘쌍에 비해 상대적으로 자주 등장하는 표현도 연어의 범위에 속한다고 볼 수 있다.

틀이 2장에서 소개된다.

2. CAT2

일반적으로 언어학 분야에서 형식틀(Formalismus)이라 함은 자연언어의 문법을 기술하기 위한 형식언어 혹은 수학적인 언어를 일컫는다. CAT2 형식틀에서 단어, 구, 문장 및 언어규칙 등은 다른 많이 알려진 형식문법에서와 같이 '자질-값 구조(Merkmal-Wert Struktur)'에 의해 기술된다. 컴퓨터상의 기술편의를 위해 자질-값 구조는 { } 괄호를 사용하여 표현된다.

(1) "Buch" : {phon=buch, cat=noun, gen=neut}

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

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

CAT2 형식틀 상에서 수형도(Baumstruktur)의 어미노드(Mutterknoten)은 { } 괄호안에, 자매노드(Tochterknoten)은 [ ]안에 표기된다. 자매노드의 각 개별 노드들은 { } 괄호안에 표기된다.

구구조문법(Phrasenstruktur-Grammatik)의 PS규칙(Phrasenstruktur-Regel)은 CAT2 형식틀 상에서 b-규칙 (b-Regel)에 의해 표현된다. (2)의 규칙은 S-> NP, VP 규칙을 CAT2 표기(Notation)를 따라 기술된 것이다. b-규칙이 언어구조를 만들어내는 역할을 하는 반면, f-규칙(f-Regel)은 만들어진 언어구조의 적합성(Wohlgeformtheit)을 검사하는 역할을 한다. 예를 들어, (2)와 같은 b-규칙을 통해 문장의 구조가 결정되었으면, 다음과 같은 f-규칙이 적용되어 주어와 술어간의 수일치에 관한 제약조건을 적용할 수 있다

(3) f_kongruenz-constraint={cat=s}.[{cat=np}>>{agr=AGR}, {cat=vp}>>{agr=AGR}].

(3)에서 >> 연산자(Operator)는 >> 왼쪽의 선행부가 만족될 경우 >> 오른쪽의 결과부도 반드시 만족해야 함을 나타낸다. 따라서 위의 규칙에 따르면 첫번째 자매노드의 카테고리가 np이고 (cat=np), 두 번째 자매노드의 카테고리가

vp일 경우, AGR이라는 동일한 변수의 사용을 통해 그것들의 일치 자질(agr)에 대한 값은 일치하거나 통합(Unifikation)이 가능해야 함을 알 수 있다.[3] 이와 같이 CAT2 형식틀에서는 통합연산이 규칙적용을 위해 핵심적인 도구로서 사용됨을 알 수 있다.

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

(4) t_haus={lex=집}.[]. <=> {lex=haus}.[].
t_produzieren={lex=만들다}.[] <=> {lex=(produzieren;herstellen;zubereiten)}.[].

(4)의 t_produzieren 규칙에서 볼 수 있는 바와 같이, 하나의 단어에 대해 여러 개의 대역어가 대응될 수 있을 때 사전에는 ; 연산자를 사용하여 여러개의 후보 대역어를 기술해준다. 이 중 어떤 대역어를 선택하는 가의 문제는 대역어 선택의 문제이므로, 사전에는 단지 가능한 모든 대역어에 대한 정보만을 기술하게 된다.[4]

3. 한독 번역상의 연어(Kollokation)

컴퓨터를 통해 한국어를 독일어로 자동 번역하는 과정에서 연어로 취급될 수 있는 것은 크게 숙어적 표현 (idiomatischer Ausdruck)과 소위 konflationelle Divergenz[5]를 나타내는 표현들이다.

한독 기계번역에서 다음과 같은 표현들은 한국어 표현의 개별 단어들이 독일어 표현의 개별단어들과 별다른 연관관계가 없으므로 숙어적인 표현이라고 볼 수 있다.

(5) a. 그 남자가 **배가 고프다**
a'. Der Mann **hat Hunger**
b. 그 아이가 **목이 마르다**
b'. Das Kind **hat Durst**

---

3) CAT2 형식틀은 Prolog 프로그래밍언어와 마찬가지로 대문자로 시작하는 문자열을 변수로 사용한다.

4) 한독 기계번역의 대역어 선택 문제에 대해서는 홍문표(2004) 참조.

5) 저자는 Dorr(1993)에 의해 소개된 Konflationelle Divergenz (conflational divergence) 용어의 한국어 번역에 있어 신중을 기하고자 본 논문에서는 원어 그대로 사용하고자 한다. 아직 이 용어에 대한 통용되거나 합의된 번역이 존재하지 않으므로, 본 용어가 나타내는 개념이 분명한 이상, 적절하지 못한 역어의 선택으로 인한 논쟁을 피하고자 한다.

한국어의 '배가 고프다', '목이 마르다' 등과 같은 표현은 코퍼스 상에서 찾아보더라도 '배'는 '고프다'라는 술어와 동시에 출현하는 빈도가 다른 어휘에 비해 상대적으로 높고, 마찬가지로 '목'이라는 단어는 '마르다'와 동시에 출현하는 빈도가 높다.[6] 이와 같은 측면에서 뿐만 아니라, 이 한국어 표현들은 독일어 번역, 즉, 'Hunger haben', 'Durst haben' 등과 단어레벨에서 서로 연관성이 없으므로, 숙어적인 표현으로 볼 수 있을 것이다.

Dorr(1993)에 의해 소개된 konflationelle Divergenz 개념은 한 언어의 어떤 단어가 다른 언어에서는 여러개의 단어로 번역되거나, 혹은 그 반대가 되는 경우를 일컫는다. 이 현상은 Hutchins&Somers(1992), Santos(1993), Trujillo(1999) 등의 연구에서는 '어휘구멍(lexical holes)'이라고도 불리워진 개념이다. 이 언어현상은 언어라는 창을 통해 세상을 바라보는 각 언어구성원들의 시각이 역사적, 문화적, 사회적 이유 등으로 인해 약간씩 다를 수 있기 때문에 나타나는 것으로 여겨진다. 다음의 그림은 konflationelle Divergenz 현상을 설명한다.

6) 이러한 정보는 보통 MI (Mutual Information)라고도 불리며, 연어를 추출하기 위한 정보로서 많이 사용된다.

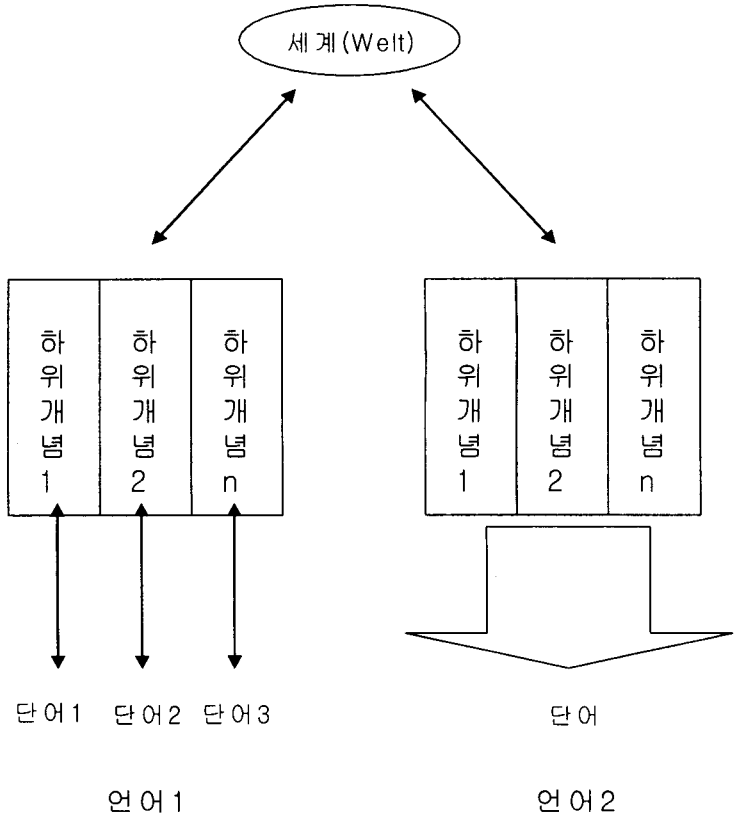

그림 1 : Konflationelle Divergenz 개념

위 그림에 따르면 언어1에서 개별 단어들로 표현되는 하나의 큰 개념이 언어2에서는 하나의 단어로 표현되고 있다. 이러한 언어현상은 한국어와 독일어간에서 매우 자주 볼 수 있다. 특히 독일어의 경우 접두사(Präfix)등의 발달로 인해, 한국어에서는 하나의 내용어(Inhaltswort)로 표현되는 개념이 독일어에서는 종종 접두사등의 사용으로 표현되는 경우가 많으므로 자주 접하게 되는 현상이라고 볼 수 있다. 예를 들어, 독일어 단어 'versalzen'의 경우 이 단어가 지시하는 개념의 하위 개념(Subkonzept)들이 한국어에서는 개별적으로 어휘화 되는 현상을 다음의 그림이 나타낸다.

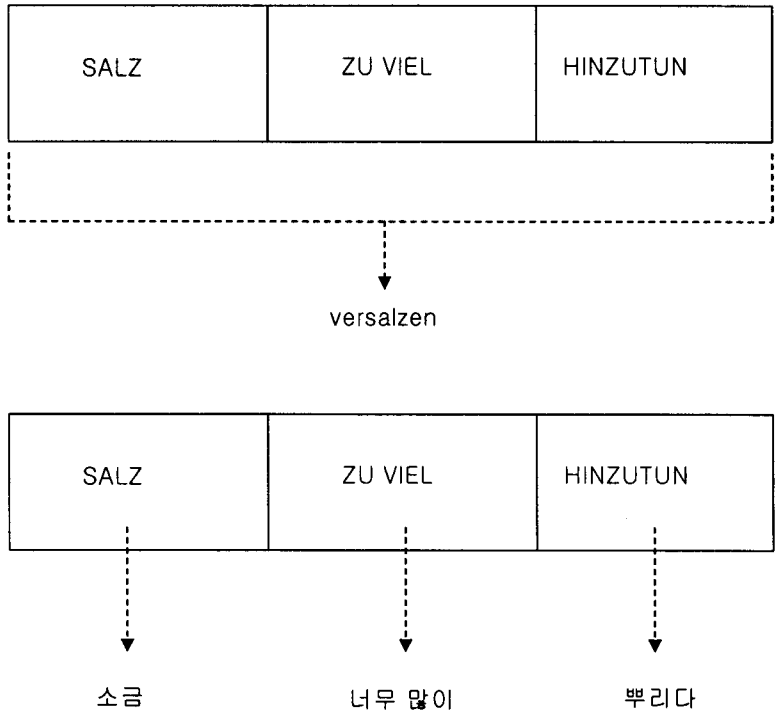

그림 2 : Konflationelle Divergenz의 예: 'versalzen'

'versalzen'의 예 이외에도 한독 번역상에서 Konflationelle Divergenz를 많이 찾아볼 수 있다. 예를 들어 접두사 'ab-'이 붙는 단어들의 경우 상당수가 한국어로 번역시에 혹은 한국어에서 독일어로의 번역시에 Konflationelle Divergenz 현상을 보인다.

독일어	한국어
abfällen	쓰레기로 나오다
abarbeiten	일해서 갚다
abärgern	오랫동안 몹시 화내다
abbacken	구워서 완성하다
abbeißen	물어 뜯다
abbestellen	주문을 취소하다
abbezahlen	할부금을 지불하다
abbiegen	옆으로 꺾이다
abblasen	불어서 없애다

위의 예에서 보는 바와 같이 한독기계번역에 있어서 한국어 표현들을 하나의 단위로 다루어야만 독일어를 생성하는데 있어서 비문법적이거나 어색한 스타일의 독일어 문장을 만들어내는 오류를 방지할 수 있다.

4. 숙어적 표현의 처리

(5)의 예에서 본 것과 같은 숙어적 표현을 다루기 위해 본 논문에서는 독일어의 기능동사구문(Funktionsverbgefüge)의 처리와 유사한 방법론을 제안한다.

(5) a. 그 남자가 **배가 고프다**
a'. Der Mann **hat Hunger**
b. 그 아이가 **목이 마르다**
b'. Das Kind **hat Durst**

독일어에서 'Entwicklung nehmen', 'Anerkennung finden', 'Initiative ergreifen' 등과 같은 기능동사구문은 자체적으로 논항구조를 지니는 술어적 성격의 명사와 시제와 양태 정보 이외에는 별다른 의미정보를 지니지 않는 동사의 결합으로 이루어진다. 한국어 숙어표현에서의 '배', '목'과 같은 명사는 언어학의 관점으로 볼 때 술어적 성격의 명사라고 보기는 어렵다. 그러나 위의 문장들에서 '고프다', '마르다' 등과 같은 동사는 시제와 양태 정보이외에는 별다른 의미를 지니지 않고, 또한 독일어에서도 정확하게 일치하는 대역어가 없다고 볼 수 있다. 따라서 위와 같은 문장들의 분석과 번역에 기존의 기능동사구문(Funktionsverbgefüge)을 확대 적용한 확대기능동사구문(Erweiterte Funktionsverbgefüge)의 적용을 제안한다. Streiter(1996)는 CAT2 형식틀 상에서 기능동사의 처리를 위해 Grimshaw&Mester(1988)의 논항전이(Argument Transfer) 개념을 도입한 바 있다. 이 방법론의 핵심은 술어명사는 자신의 논항구조를 가지고 기능동사는 시제와 양태 정보만을 가지며, 논항전이 조작을 통해 기능동사가 술어명사의 논항구조를 전이 받고 다른 문법 요소들을 하위범주한다는 것이다. 이와 같은 방법론을 한국어의 숙어적 표현에도 적용할 수 있을 것으로 본다. 즉, 한국어 명사 '목'은 'theme' 의미역을 가지며 주격을 취하는 명사를 하위범주화 한다. 동시에 이것은 또 '마르다'라는 동사를 하위범주화 한다. '마르다' 동사는 주격 술어명사를 하위범주화 하는데, 이 동사는 논항전이 연산을 통해 술어명사의 논항구조를 공유하게 된다. 다음 자질-값 구조는 '목'의 사전정보를 나타낸다.

{lex=목, head={cat=n, ehead={pred=yes, sem=EVENT}}, subcat={arg1={role=theme, head={ehead={cat=n, case=nom, sem=animal}}}, vsup={head={ehead={cat=v, vsup=yes, lex=마르다}}}}}[7]

위 사전정보를 보면, '목'은 논항구조(subcat)를 가짐을 알 수 있다. 이 논항구조의 첫 번째 논항(arg1)은 'theme'의미역을 가지며, 주격(case=nom)을 취한다. 또한 기능동사(vsup)를 하위범주화 하는데, 이 기능동사는 반드시 '만들다'이어야 함이 명시되어 있다(lex=만들다). 이제 '마르다' 동사의 사전 정보를 보기로 한다.

{lex=마르다, head={cat=v, ehead={vsup=yes, sem={aspect=dur}&SEM}}, subcat={arg1=ARG1, arg2=ARG2, arg3=ARG3, arg4={head={cat=n, ehead={sem=SEM, pred=yes, case=nom}}, subcat={arg1=ARG1, arg2=ARG2, arg3=ARG3}}}

'마르다'의 사전정보를 보면, 이 동사는 양태 정보이외에 아무런 의미정보를 가지고 있지만, 하위범주화 하는 술어성 명사의 의미정보와 동일한 변수를 가짐으로써 (SEM) 술어성 명사의 의미를 같이 갖게 된다. 또한 위 사전정보에서 가장 주목할 점은 이 동사의 논항구조가 비어있다는 점이다. 그러나 역시 이 경우에도, 하위범주화 하는 술어성 명사(arg4)의 논항구조내의 논항들과 동일한 변수(ARG1, ARG2, ARG3)를 가짐으로써 술어성 명사의 논항구조를 공유하게 된다. '마르다' 동사가 취하는 술어성 명사 '목'의 격을 주격(case=nom)으로 제약함으로써, 다음과 같은 비문을 생성하지 않게 된다.

(6) * 그 아이가 목을 마르다
Das Kind hat Durst

'목'과 '마르다'의 사전정보를 보면 두 단어가 서로 하위범주화하고 있는 것을 알 수 있다. 이러한 특성을 반영하고 논항전이(argument transfer)를 가능하게 하는 두 단어의 결합규칙은 다음의 b-규칙과 같다.

b_fv={head={cat=vp, ehead={sem=SEM, voice=VOICE}}, subcat=SUBCAT}.[

7) 이 자질-값 구조에 등장하는 'ehead', 확대핵심어(erweiterter Kopf) 개념은 명사구(NP)의 DP 투영현상 등을 설명하기 위해 도입된 개념이나, 본 논문에서는 일반적인 핵심어(Kopf)로 이해하면 무리가 없을 것으로 생각된다.

{head={cat=n, ehead={sem=SEM}}&HEAD, subcat=SUBCAT, subcat={vsup={head={ehead={cat=v, lex=LEX, vsup=yes}}}}, {lex=LEX, head={ehead={cat=v, voice=VOICE, vsup=yes}}, {subcat={arg4={head=HEAD}}}}].

위 b_fv 규칙은 술어성 명사와 기능동사를 결합하게 하는 기능을 한다. 기능동사(vsup=yes)는 술어성 명사를 하위범주화 하는데, 특히 어휘 정보를 직접 제약하므로써(lex=LEX), '마르다'의 경우 '목'이라는 명사만을 하위범주화 하게 된다. 또한 두 노드의 결합으로 만들어진 상위 노드의 경우, 하위범주 정보를 술어명사로부터 직접 전이받게 되어(subcat=SUBCAT), 동사구 자체가 나머지 논항들을 하위범주화 할 수 있게 된다. 즉, '목이 마르다'라는 동사구는 술어성명사 '목'의 논항구조를 그대로 전달받아, 주어 명사만을 논항구조에 남겨놓게 된다.[8] (그림 3 참조)

분석단계에서는 위와 같은 방식으로 '목이 마르다'가 하나의 단위로 인식되어 분석이 가능하다. 이와 같이 분석된 한국어 구조를 독일어로 변환(Transfer)하는 단계에서는 기능동사는 삭제되고 술어성 명사만 남게 된다. 기능동사를 삭제하더라도 기능동사가 가지고 있는 시제, 양태 등의 정보는 '마르다'의 사전정보를 보면 알 수 있듯이 sem자질에 대해서 '마르다'가 취하는 술어성 명사의 sem자질과 동일한 값을 공유하므로써(sem=SEM), 술어성 명사에 그 정보를 전달하게 된다.[9]

---

8) b_fv 규칙을 보면, 상위노드, 즉, vp가 술어성 명사의 논항구조를 그대로 가져옴을 알 수 있다. 그러나 이렇게 되면, 이미 술어성 명사가 기능동사를 하위범주화 하여 동사구 레벨에서는 하위범주틀내에서 기능동사는 삭제되어야 함에도 불구하고, 동사구가 또 다른 기능동사를 하위범주화하게 되는 결과를 초래한다. 따라서 HPSG 이론의 하위범주화 원칙(Subkategorisierungsprinzip, subcategorization principle)과 같은 장치가 필요하다. 이와 같은 장치는 CAT2 형식틀 상에서 f-규칙에 의해 구현이 가능하나, 본 논문에서는 이 규칙에 대해서는 언급하지 않기로 한다.

9) 그림 3에서 '그 아이가', '목이'와 같은 조사구 혹은 명사구의 구조에서 핵심어를 무엇으로 보느냐에 따라 PostP 혹은 NP로의 분석이 가능하다. 어떤 분석방법을 택하던지 본 논문에서 주장하는 방법론의 적용에는 문제가 없으므로, 본 논문에서는 조사구의 관점을 취하고자 한다.

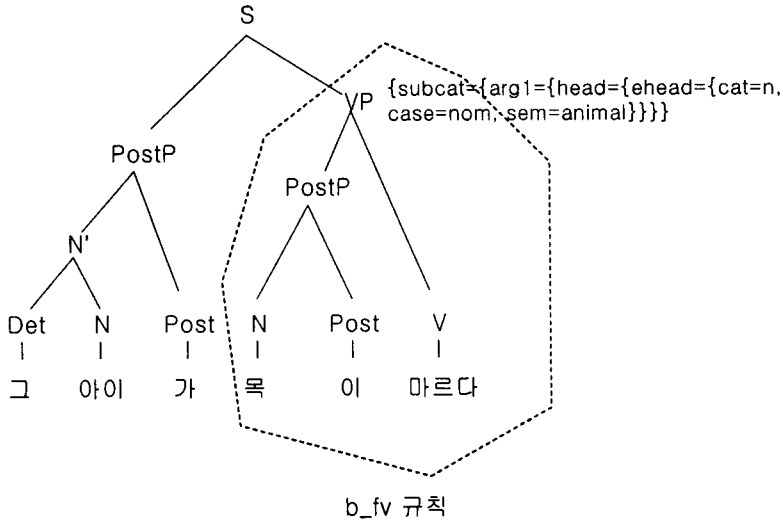

그림 3 : '그 아이가 목이 마르다'의 통사구조

그 결과 변환(Transfer) 단계에서는 한국어와 독일어간에 구조변화가 전혀 없이 변환을 수행할 수 있게 된다.

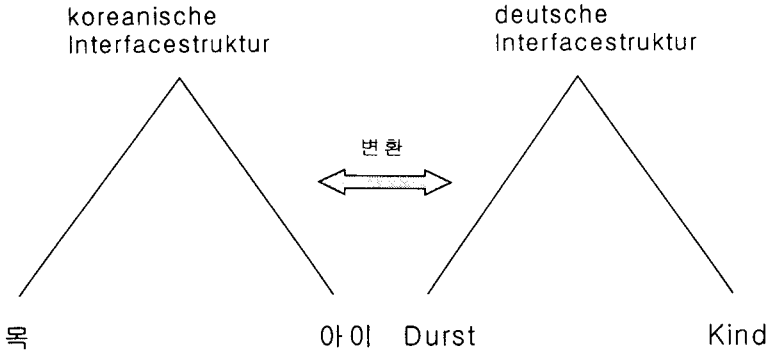

그림 4 : '그 아이가 목이 마르다 '의 한-독 변환구조

술어성 명사 '목'을 'Durst'로 변환시켜주는 규칙은 다음과 같다.

t_mok_durst={lex=목, head={ehead={pred=yes}}}.[] <=> {lex=durst}.[].

위 변환규칙에서 볼 수 있는 바와 같이 한국어의 '목'이 항상 'Durst'로 번역되는 것이 아니라 술어성 명사(pred=yes)일 경우에만, 'Durst'로 번역된다. 위와 같이 독일어 인터페이스구조(interfacestruktur)로 변환된 후에는, 독일어 어휘들에 들어있는 정보를 통하여 독일어 문장이 생성되게 된다. 독일어의 경우에도 한국어와 마찬가지로 'Durst'가 술어성명사로 사전에 등록이 된다. 이 때 한국어의 인터페이스 구조에서 술어성명사 '목'이 가지고 있던 시제와 양태 등에 관한 정보는 독일어 'Durst'로 그대로 전달된다.[10] 'Durst'는 한국어의 '목'과 마찬가지로 기능동사 'haben'을 하위범주화한다. 따라서 이번에는 한국어의 분석과 정반대의 절차로 이러한 정보로부터 기능동사 'haben'을 생성하게 된다.[11] 이렇게 생성된 'haben'은 'Durst'로부터 물려받은 시제, 양태 정보를 가지고 최종 생성 단계에서 정확한 최종 표층형태인 'hat'로 생성되게 된다.

확대기능동사구문은 '그 아이가 목이 마르다'와 같은 이중주어를 보이는 숙어적 표현에만 적용될 수 있는 것이 아니라 '자동차를 길을 들이다(ein Auto einfahren)'와 같은 일반적인 연어에 대해서도 마찬가지로 적용이 가능하다.

(7) 그 남자가 그 새 자동차를 **길을 들이었다**
Der Mann **fuhr** das Auto **ein**

위의 예에서 '들이다'에 일대일로 대응되는 독일어 대역어는 없다고 볼 수 있다. 따라서 위의 경우도 '목이 마르다'의 경우와 마찬가지로 '길'을 술어성 명사로 분석하고, '들이다'를 기능동사로 보는 방법이 적용가능하다.

---

10) '목'은 '마르다'로부터 시제와 양태에 관한 정보를 이미 전달받은 상태인데, 이것은 '마르다'의 사전구조에서 '마르다'가 하위범주화하는 '목'과 의미정보를 공유하는(sem=SEM) 것을 통해 이루어진다.

11) 한국어 연어표현이 독일어에서 기능동사구문으로 번역이 되지 않고, 일반적인 N-V 형태로 번역이 될 경우에는, 변환규칙의 추가가 필요하다. 즉, 특정 한국어 명사-동사 표현을 독일어로 번역할 때, 기능동사의 생성을 방지하는 추가 변환규칙을 통해, 기능동사구문의 생성이 방지되고, 입력된 동사의 디폴트(default) 대역어 생성을 통해 일반적인 N-V 구문을 생성할 수 있다.

'길'과 '들이다'의 사전 정보는 다음과 같다.

{lex=길, head={cat=n, ehead={pred=yes, sem=event}}, subcat={arg1= {role=agent, head={ehead={case=nom, sem=human)}}, arg2={role=theme, head={ehead={case=acc, sem=(vehicle;animal)}}}, vsup={head={ehead= {vsup=yes, cat=v, lex=들이다}}}}}

{lex=들이다, head={cat=v, ehead={sem={aspect=dur}&SEM}}, subcat= {arg1=ARG1, arg2=ARG2, arg3=ARG3, arg4={head={cat=n, ehead={cat=n, sem=SEM, pred=yes, case=acc}}, subcat={arg1=ARG1, arg2=ARG2, arg3=ARG3}}}}

술어성명사 '길'은 하나의 주어와 목적어를 하위범주화 한다. 이 두개의 논항들의 의미역(semantische Rolle)과 격(Kasus)은 바로 이 술어성명사 '길'로부터 부여받는다. 기능동사 '들이다'는 이 술어성 명사만을 하위범주화 한다. 그러나 b_fv 규칙에서 본 바와 같이 이 술어성명사를 하위범주화함으로써 술어성명사가 하위범주화하는 논항들에게도 논항전이(Argument Transfer) 조작을 통해 접근할 수 있게되는 것이다.

본 절에서는 한국어의 숙어적인 표현들의 분석과 독일어로의 변환, 독일어 생성을 위해 독일어의 기능동사구문 분석을 위한 방법론이 확대적용될 수 있음을 보였다.

5. Konflationelle Divergenz의 처리

Konflationelle Divergenz 현상은 한 언어에서는 한 단어로 표현되는 의미가 다른 언어에서는 여러개의 단어로 표현되거나 혹은 그 반대로 되는 현상을 일컫는다. 이것은 세상의 사태에 대한 언어마다의 상이한 개념화 결과로 나타나는 것으로 이해된다. 일반적으로 한국어와 다른 언어간의 번역에서 흔히 볼 수 있는 Konflationelle Divergenz 현상은 주로 '~를 하다' 동사를 사용하는 경동사구문에서 찾아볼 수 있다.[12] 그러나 한국어와 독일어간의 번역에서는 이것 이외에도 아래와 같이 많은 예에서 이 현상이 발견된다.

한국어	Deutsch
그네를 타다	schaukeln
시소를 타다	wippen

12) 한국어와 일본어의 경동사구문에 대한 기계번역 시스템 상에서의 처리는 각각 Choi(1995), Fujinami&Nanz(1998) 참조

말을 타다	reiten
버터를 바르다	buttern
소금을 뿌리다	salzen
영화로 만들다	verfilmen
호루라기를 불다	pfeifen
비가 오다	regnen
눈이 오다	schneien
천둥이 치다	donnern
번개가 치다	blitzen
얼굴이 빨개지다	erröten

한국어에서는 예를 들어, '그네'라는 대상과 '타다'라는 동작의 개념이 각각 어휘화되어 '그네를 타다'라고 표현되는 반면, 독일어에서는 'schaukeln'이라는 하나의 동사로 어휘화된다 (그림 5 참조). 마찬가지로, '시소를 타다'는 'wippen'이라는 하나의 어휘로 표현된다.

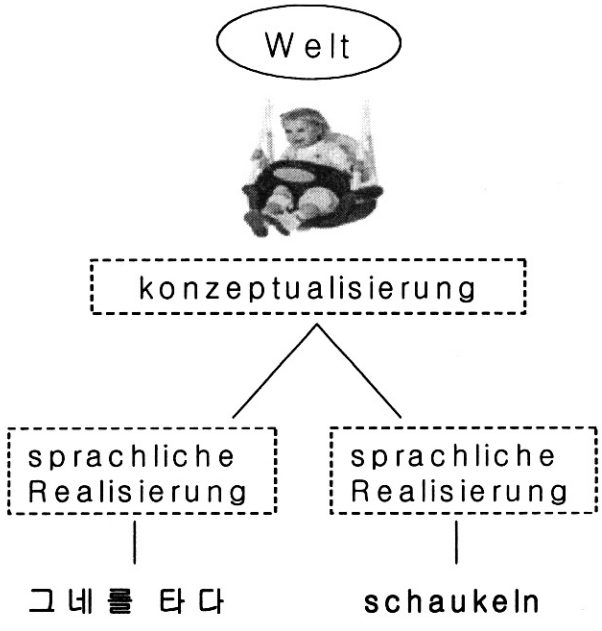

그림5:Konflationelle Divergenz의 예: 'schaukeln'

숙어적 표현과 마찬가지로 konflationelle Divergenz 관계에 있는 표현들도 한국어에서 독일어로의 번역시 단어 레벨에서의 합성적인 번역을 통해 구표현의 번역 결과

를 도출해 낼 수 없는 문제가 있다. 즉, '그네를 타다'의 번역을 '그네' 와 '타다'의 번역을 통해 'Schaukeln einsteigen' 등과 같이 번역을 하게 되면, 어색하거나 잘못된 번역을 초래하게 되는 문제가 생긴다. 따라서 이와 같은 표현들도 앞장의 숙어적 표현과 마찬가지로 하나의 단위로 분석을 하여 독일어로 번역을 해야한다. 이를 위해 앞장에서 소개된 확대기능동사구문(Erweiterte Funktionsverbgefüge) 매커니즘을 이 경우에도 사용할 수 있을 것이다. 이 방법론에 따르면, '그네를 타다'의 경우, '그네'가 술어성 명사로 분석되고, '타다'를 기능동사로 볼 수 있을 것이다. 술어성 명사 '그네'는 하나의 논항, 즉, 행위역(Agens)을 가지는 주격 명사를 하위범주화 하며, 또한 기능동사를 하위범주화 한다. '타다'의 경우 과연 다른 일반적인 기능동사(Funktionsverb)들과 마찬가지로 자체적인 의미를 가지지 않는 기능동사로 볼 수 있는가에 대한 논의가 있을 수 있을 것이다. 그러나 우리는 이러한 문제를 '타다' 동사에 대해 기능동사로서의 '타다'와 'EINSTEIGEN' 자체의 의미를 가지는 '타다'라는 두개의 엔트리를 가정하므로써 해결하고자 한다.[13] 이에 따라, '그네'와 '타다'의 사전정보를 구성해 보면 다음과 같다.

{lex=그네, head={cat=n, ehead={pred=yes, sem=EVENT}}, subcat={arg1={role= agent, head={ehead={cat=n, case=nom, sem=human}}}, vsup={head={ehead={cat=v, vsup=yes, lex=타다}}}}}

{lex=타다, head={cat=v, ehead={vsup=yes, sem={aspect=dur}&SEM}}, subcat={arg1=ARG1, arg2=ARG2, arg3=ARG3, arg4={head={cat=n, ehead={sem=SEM, pred=yes, case=nom}}, subcat={arg1=ARG1, arg2=ARG2, arg3=ARG3}}}

'그네를 타다'는 위 사전구조와 앞장에서 소개된 b_fv 규칙에 의해 다음과 같이 분석된다.[14]

---

13) 이와 같이 직관적으로 하나의 의미를 가진 것 같은 동사에 대해 두개 이상의 사전 엔트리를 구성하는 것에 대해서는 논란의 여지가 있을 수 있다. 그러나 자연언어처리(Sprachverarbeitung)분야와 일반언어학 분야에서 문제해결을 위한 접근 방법이 다를 수 있다는 점에서 '타다'와 같은 단어에 대해 두개 이상의 엔트리를 상정하는 것에 현재로서는 큰 문제는 없다고 본다.

14) '그네를 타다'와 같은 경우 HPSG이론에서의 관용어 분석방법과 같이 명사구가 동사의 형태 등만을 하위범주화 하는 방식의 분석도 가능하다. 이러한 분석방법도 CAT2 형식틀 내에서 가능하며, 본 논문에서 제안하는 방법론과의 차이는 계산상의 효율적인 측면이라고 볼 수 있다. 이 점을 지적해주신 익명의 논문심사자께 지면을 통해 진심으로 감사드립니다.

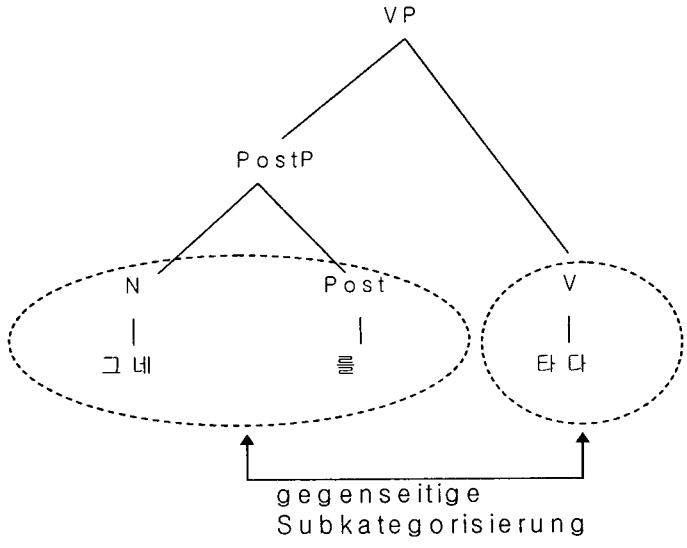

그림 6 : '그네를 타다'의 통사분석

'목이 마르다'의 경우와 마찬가지로 '그네를 타다'의 경우 술어성명사 '그네'만이 변환(Transfer)을 위한 인터페이스구조에 남게 된다. '그네'는 다음의 변환규칙에 의해 'schaukeln'으로 변환된다.

t_kuney_schaukeln={lex=그네, head={ehead={pred=yes}}}.[] <=> {lex=schaukeln, head={cat=v}}.[].[15]

독일어 동사 'schaukeln'으로 변환된 후 이것과 주어 명사를 선택하여 문장을 생성해내는 과정은 기술적으로 별다른 어려움이 없는 과정이므로 여기서는 기술을 생략하도록 한다.

지금까지 본 논문에서는 한국어와 독일어간의 번역시 발생하는 konflationelle Divergenz 현상에 대한 분석, 변환을 위한 방법론을 제시하였다. 독일어의 기능동사구문을 확대적용한 확대기능동사구문을 적용하면, 한국어의 분석과 독일어로의 변환

15) '그네를 샀다'에서처럼 '그네'가 술어성명사가 아닌 경우는 다음과 같은 변환규칙에 의해 일반적인 독일어 명사 'Schaukel'로 번역이 된다.

t_kuney_schaukel={lex=그네, head={ehead={pred=no}}}.[] <=> {lex=schaukel, head={cat=n}}.[].

이 성공적으로 가능함을 보였다.

그러나 konflationelle Divergenz 현상은 한국어와 독일어만을 번역해야하는 상황에서는 지금까지 제안한 확대기능동사구문 매커니즘을 이용해 해결할 수 있지만, 한국어와 독일어 이외에 다른 언어를 추가적으로 동시에 번역해야 하는 다국어 기계번역 환경에서는 문제가 발생할 수 있다.[16]

한국어와 독일어 이외에 일본어도 추가적으로 번역을 해야하는 상황을 '그 아이가 그네를 탄다' 예문을 통해 가정해보자.

(8) 한국어: 그 아이가 그네틀 탄다
독일어: Das Kind schaukelt
일본어: Sono kotomo-ka branko-ni norimatsu

위 예문에서 한국어와 독일어 간에는 konflationelle Divergenz 현상이 발견되어, 확대기능동사구문을 적용하여 이 문제를 해결할 수 있음을 보았다. 그러나 동일한 기계번역 엔진을 사용하여 한국어 문장을 일본어로 번역하고자 할 때, '그네틀 타다'를 확대기능동사구문을 사용하여 번역할 수 없다. 왜냐하면 한국어 '그네를 타다'는 일본어의 'branko-ni norimatsu'로 합성적으로 번역되기 때문이다. 즉, 한국어와 독일어만을 고려할 경우에는 '그네를 타다'를 하나의 단위로 분석하고 변환하였으나, 한국어와 독일어 이외에 일본어 등과 같은 제 3의 언어를 고려할 때는 지금까지 소개한 방법론이 그 한계에 부딪힌다.

이와 같은 한계는 애당초 변환(Transfer)기반 기계번역 방법론을 채택하였기 때문에 필연적으로 발생하는 것으로 여겨진다. 즉, 변환방식 기계번역 시스템은 출발언어(Quellsprache)와 목표언어(Zielsprache) 두개만을 고려하여 개발하기 때문에, 다국어 번역을 고려하면 문제가 발생할 수 밖에 없다.

이러한 문제를 해결하기 위해서는 변환기반 방법론 대신에 중간언어방식(Interlingua-Ansatz)을 고려할 수 있다. 중간언어방식의 방법론에서는 언어보편적인 표상을 만들어 출발언어의 의미를 이 언어보편적인 표상으로 매핑한 후, 이 언어보편적인 표상으로부터 거꾸로 목표언어의 문장을 생성해내는 기법을 사용한다. Dorr(1993)에서 소개된 UNITRAN 시스템과 Hong&Streiter(1999)과 홍문표(1999)에서 소개된 UNL 시스템이 이와 같은 중간언어방식을 사용하는 대표적인 시스템이다.

16) 실제 기계번역과 관련된 시장분석 등에 따르면, 한국어나 독일어와 같이 단 두개의 언어만을 번역할 수 있는 시스템이 아닌, 여러 개의 언어를 동시에 하나의 시스템에서 번역할 수 있는 다국어 번역시스템에 대한 요구가 급증하고 있다.

UNITRAN 방식에서는 문장의 의미를 Jackendoff(1990)의 개념구조(conceptual structure)를 기반으로 잘개 쪼개서 표현하고, 잘개 쪼개진 의미로부터 다른 언어의 문장을 다시 생성해내는 방법론을 사용한다. 반면, UNL 시스템에서는 영어라는 제3의 언어를 중간언어로 사용하여, 한국어-독일어-일본어 간의 번역의 경우, 각각의 언어를 영어를 사용한 중간언어표상으로 만든 후 그 표상으로부터 원하는 언어의 문장을 만들어 내는 방식을 사용한다.

그러나 이와 같은 중간언어방식은 그 이론적인 화려함 이외에 실제로 구현하기 매우 어려운 단점들을 가지고 있다. 즉, 완벽한 의미분석은 완벽한 형태소, 구문 분석을 전제라 하고 있는데, 형태소 분석과 구문 분석이 실세계에 등장하는 많은 문장들을 다루는 경우 완벽하게 되는 것은 거의 불가능한 일이기 때문이다. 그러나 이러한 문제점 이외에도 중간언어방식에도 변환방식에서와 마찬가지의 똑같은 문제점을 지니고 있다. '그네를 타다'의 예를 UNL 방식으로 번역을 한다고 가정을 해 보자. 한국어의 '그 아이가 그네를 탄다'는 다음과 같은 사전구조를 기반으로 (9)와 같이 중간언어로 번역이 될 것이다.

한국어 - Universal Word
아이 <=> child(icl>human)
그네 <=> swing(icl>instr)
타다 <=> get_in(icl>event)

(9) agt(get_in(icl>event)@entry@pres, child(icl>human)@def)
obj(get_in(icl>event)@entry@pres, swing(icl>instr)@def)[17]

(9)는 앞서 언급한 바와 같이 언어보편적인 중간언어표상이므로 이 구조로부터 독일어와 일본어의 번역문이 각각 다음의 사전구조에 기반하여 생성될 것이다.

독일어 - Universal Word
kind <=> child(icl>human)
schaukel <=> swing(icl>instr)
einsteigen <=> get_in(icl>event)

일본어 - Universal Word
kotomo <=> child(icl>human)

17) 이 표상은 'child'와 'get_in'이 행위역 관계에 있고, 'swing'과 'get_in'이 목적역 관계에 있음을 의미한다. 구체적인 UNL 형식틀에 대해서는 홍문표(1999)를 참조하기 바란다.

branko <=> swing(icl>instr)
noru <=> get_in(icl>event)

한국어가 입력문이고 독일어와 일본어를 출력해야 하는 경우, 독일어 생성 모듈은 obj(get_in(icl>event)@entry@pres, swing(icl>instr)@def)로부터 'Schaukel einsteigen'이 아니라 'schaukeln'을 생성해야 하며, 일본어 생성 모듈은 'branko-ni norimatsu'를 생성해야 한다는 정보를 가지고 있어야 한다. 즉, 변환방식에서와 마찬가지로 개별언어쌍 간의 특징을 반영한 정보를 가지고 있어야 하는 것이다. 이상에서와 같이 중간언어방식도 다국어 기계번역 환경에서는 konflationelle Divergenz를 다루는데 변환방식과 마찬가지의 취약점을 드러냄을 알 수 있다.

6. 맺는말

본 논문에서는 한국어와 독일어간의 기계번역상에 등장하는 연어의 처리문제에 대해 다루었다. 규칙기반 기계번역의 기본 아이디어는 번역합성성의 원리라고 할 수 있다. 즉, 어떤 언어표현을 번역하기 위해서는 그 표현을 구성하는 구성단위의 번역이 선행되고 그 구성단위를 통사적으로 결합하는 규칙과 대응되는 번역규칙에 의해 상위 단위의 언어표현이 번역된다. 그러나 연어의 경우 이와 같은 번역합성성의 원리에서 위반되기 때문에 분석과 변환에 많은 어려움을 초래한다. 본 논문에서는 이 문제의 해결을 위해 독일어 기능동사구문의 처리에서 아이디어를 얻어 이를 확장한 확대기능동사구문 매커니즘을 제안했다. 술어성 명사와 기능동사가 상호 하위범주하는 성격을 이용하여 연어적 표현을 분석단계에서 하나의 단위로 분석하고, 변환단계에서는 술어성 명사만을 남김으로서, 구조변환의 부담을 최소화하고 구조변환시의 오류를 피하였다. 이러한 방법론으로 우리는 숙어적 표현과 konflationelle Divergenz를 나타내는 표현들을 성공적으로 분석하고 독일어로 변환할 수 있음을 보였다.

그러나 한국어와 독일어 이외에 일본어 등과 같이 제3의 언어가 개입되는 다국어 번역의 환경에서는 확대기능동사구문이 그 적용의 한계를 드러냄을 보였다. 이러한 문제를 해결하기 위해 중간언어방식의 기계번역방법론도 응용할 수 있음을 보였으나, UNL 시스템의 예에서 본 것과 같이 중간언어방식에서도 개별언어간의 언어현상을 모두 고려해야 하는 변환방식에서의 약점과 똑같은 문제점이 노출되었다.

현재 프로토타입의 수준에만 머물러 있는 한-독 기계번역 시스템이 향후에 상용시스템으로 발전되기 위해서는 한국어와 독일어의 연어사전 및 두 언어 사이에 konflationelle Divergenz를 보이는 표현들에 대한 대규모 수집 및 리소스 구축이 절

대적으로 필요하며, 이에 대한 독어학도들의 많은 관심이 필요할 것으로 본다.[18]

참고문헌

홍문표 (1999). UNL-프로젝트에 대하여 - 독/영 기계 번역 시스템에 기반한 독일어 UNL-분석기의 구현을 중심으로, 독일문학70집, 40권 2호

홍문표 (2004). 한-독 기계 번역의 대역어 선택 문제에 대하여, 독일문학 91집, 45권 3호

Choi, S.K. (1995). Unifikationsbasierte Maschinelle Übersetzung mit Koreanisch als Quellsprache, IAI Working Papers 34, Saarbrücken

Dorr, B. (1993). A view from the lexicon, The MIT Press, Cambridge/London

Fujinami,T. & Nanz, C. (1998). The light verb construction in Japanese, Verbmobil Report 221, Universtität Stuttgart

Hong, M.P. & Streiter, O. (1999). Overcoming the Language Barriers in the Web: The UNL-Approach, in Tagungsband der 11. Jahrestagung der Gesellschaft für linguistische Datenverarbeitung (GLDV'99)

Hutchins, W.J. & H.L. Somers (1992). An Introduction to Machine Translation, The MIT Press, Cambridge/London

Grimshaw, J. & Mester, A. (1988). Light verb and theta-marking, Linguistic Inquiry 19

Landsbergen, J. (1998). Compositional Translation Revisited, in 10th European Summer School in Logic, Language and Information (ESSLI) Workshop "Machine Translation", Saarbrücken

Lewandowski, T. (1985). Linguistisches Wörterbuch, Quelle&Meyer, Heidelberg/Wiesbaden

Santos, D. (1993), Broad-Coverage Machine Translation, in Natural Language Processing: The PLNLP Approach, Boston/Dordrecht/London

Streiter, O. (1996). Linguistic Modeling for Multilingual Machine Translation, Shaker Verlag, Aachen

---

18) 본 졸고에 대해 정성껏 심사를 해주신 두 분의 익명의 심사위원님들께 진심으로 감사를 드리는 바이다. 두 분 심사위원들의 고견 덕분에 상당 부분을 내용적으로 개선할 수 있었다. 그럼에도 불구하고 본고에 있을 수 있는 오류는 전적으로 저자의 책임임을 밝히는 바이다.

Trujillo, A. (1999). Translation Engines: Techniques for Machine Translation, Springer Verlag, Berlin/Heidelberg/NewYork

Zusammenfassung

Über die Kollokation in der maschinellen Übersetzung Koreanisch-Deutsch

Hong, Munpyo (ETRI)

Die vorliegende Arbeit beschäftigt sich mit Kollokations-Phänomen in der maschinellen Übersetzung (MÜ) Koreanisch-Deutsch. Kollokationen sind vor allem deswegen schwierig zu behandeln, weil die kollokativen Ausdrücke dem Kompositionalitätsprinzip der Übersetzung nicht unterliegen. Daher sollen sie in der Übersetzung als eine Einheit behandelt werden. Dazu wurde sich des Funktionsverbgefüge-Mechanismus bedient. Die idiomatischen Ausdrücke im Koreanischen entsprechen im traditionellen Sinne nicht den Funktionsverben. Jedoch konnte gezeigt werden, dass sie mit dem Funktionsverbgefüge-Mechanismus behandelt werden können. Die lexikalisch-semantische Verschiedenheit, konflationelle Divergenz genannt, wurde anschließend vorgestellt. Die konflationelle Divergenz kommt vor, wenn zwei Sprachen über unterschiedliche Mittel verfügen, um ein Konzept sprachlich zu realisieren. Ähnlich wie bei den idiomatischen Ausdrücken konnte hierbei auch der Funktionsverbgefüge-Mechanismus angewendet werden. Allerdings wurde darauf hingewiesen, dass diese Strategie nur für die Übersetzung zwischen zwei Sprachen geeignet ist. Anhand des koreanischen, deutschen und japanischen Beispiels war zu sehen, dass diese Strategie bei der multilingualen MÜ die konflationelle Divergenz zwischen Sprachen nicht richtig erfassen kann. Als eine alternative Lösung wurde der UNL-Ansatz erörtert.

핵심어: 연어 Kollokation, 확대기능동사구문 Erweiterte Funktionsverbgefüge, 기계번역 Maschinelle Übersetzung, 규칙기반 기계번역 Regelbasierte MÜ, 중간언어기반 기계번역 Interlingua-basierte MÜ

필자 Email: munpyo@etri.re.kr
논문투고일: 2005. 9. 30. / 심사일: 2005. 11.8. / 심사완료일: 2005. 11. 30.