UNL-프로젝트에 대하여

- 독/영 기계 번역 시스템에 기반한 독일어 UNL-분석기의 구현을 중심으로

홍문표(자알란트 대학 IAI연구소)

I. 서론

I.1 UNL프로젝트

본 논문은 현재 독일을 비롯, 프랑스, 일본, 이탈리아, 스페인, 중국 등 세계 13개국의 연구소와 대학이 개발에 참여하고 있는 UNL프로젝트(Universal Networking Language-Project)를 간단히 소개하고, 현재 필자가 개발에 참여 중인 독일어 UNL-분석기의 구조 및 특징을 소개하는 것이 목적이다. 또한 이를 통하여 현재 독일 내에서 진행 중인 독일어 전산 처리 기술의 전개 방향을 간접적으로나마 소개하는 것도 본 논문의 작은 목적이다.

1990년대 초부터 서서히 일반인들의 생활 속에까지 보급된 인터넷(Internet)은 이제 많은 사람들에게는 그들의 사업 활동이나, 학문 연구 또는 취미 생활에 이르기까지 없어서는 안 될 존재가 되었다. 인터넷 초창기 때의 많은 전문가들의 예상과는 달리 현재에는 영어가 아닌 다른 언어로 작성된 웹 페이지의 비율이 점차 늘어가는 추세이다.[1] UNL프로젝트는 이러한 인터넷상의 언어 장벽을 뛰어 넘기 위해 1996년 말 일본의 IAS 연구소 (http://unl.ias.unu.edu)에 의해 발기되었고 2006년까지 진행될 예정이다. 이 프로젝트에서는 자연 언어가 표현할 수 있는 의미를 정보의 손실 없이 표상화(repräsentieren)할 수 있는 인공 언어 UNL (Universal Networking Language)을 정의하고 자연 언어를 자동으로 UNL 포맷으로 변형하고, 거꾸로 UNL로부터 일반 사용자와 컴퓨터간의

---

1) 이에 대한 통계 자료를 위해서는 Allied Business (1998) 참조.

인터페이스를 위하여 자연 언어로 된 문장을 생성해 내는 도구들을 개발하는 것이 그 목적이다.[2)]

독일에서는 잘란트 대학의 응용 정보 연구소(IAI)가 프로젝트의 주 파트너로서 독일어 UNL-분석기(UNL-Enconverter)를 개발하고 있으며, 마찬가지로 잘란트 대학의 독일 인공지능 연구 센터(DFKI)가 협력 파트너로서 독일어 UNL-생성기(UNL-Deconverter)를 개발하고 있다.

I.2 UNL-시나리오

일종의 형식 언어(formale Sprache)인 UNL은 자연 언어가 담을 수 있는 다양한 정보를 표현하기 위해 정의되었다. UNL은 자연 언어가 표현할 수 있는 의미를 표기한다는 점에서 메타언어(Metasprache)이자 의미적 개념의 마크업(Mark-up)언어이다.[3)]

일반 사용자의 입장에서 볼 때 웹페이지의 내용을 UNL을 사용하여 표기한다는 것은 매우 어렵고 많은 훈련을 요하는 일이다. 이러한 문제를 해결하기 위해 어떠한 자연 언어, 예를 들어 독일어로 작성된 웹페이지를 UNL로 자동으로 변환시켜 주는 모듈이 필요하다. 이러한 모듈을 UNL-분석기(UNL-Enconverter)라고 한다. 독일어 UNL-분석기는 입력된 독일어 문장을 분석하여 그 문장이 나타내는 명제적 의미를 UNL로 표현하는 역할을 담당한다.

UNL로 표기된 문서를 일반 사용자가 접근하기 위해서는 UNL문서를

---

2) 이미 기계 번역(maschinelle Übersetzung)분야에서는 각 언어 쌍들의 번역을 위한 개별 트랜스퍼 모듈을 상정하지 않는, 언어 중립적인 표상을 이용한 번역 방식인 중간 언어 번역 방식(Interlingua)이 1960년대부터 여러 프로젝트에서 연구되어 왔다 (Vgl. Luckhardt (1987:9ff)). UNL프로젝트가 이전의 프로젝트들과 다른 점은 우선 개발 참가국 수가 훨씬 많고, 따라서 여러 언어들을 통해 시스템의 성능을 검증할 수 있고, 또한 시스템의 개발 목적이 단순히 텍스트의 번역뿐만이 아니라 데이터 베이스(Datenbank) 상에서의 효율적인 정보의 저장, 인덱싱, 검색 등이라는데 있다.

3) HTML이나 SGML은 문서의 포맷 등과 같은 형식만을 제어하고 정의하는 기능을 가진 마크업 언어임에 반하여, UNL은 형식뿐만 아니라 내용까지도 제어하고 정의하는 기능을 가졌다.

그 사용자의 모국어로 전환해 주는 모듈이 필요하다. 이러한 모듈을 UNL-프로젝트에서는 UNL-생성기(UNL-Deconverter)라고 부른다. UNL시스템이 완성되어 상용화되면 이러한 생성기는 현재 대다수의 웹브라우저들이 자바(Java)나 자바스크립트(JavaScript) 형태의 데이터들을 자동으로 해석하여 실행하는 것과 마찬가지로 브라우저(Browser)안에 일종의 플러그인 소프트웨어(plug-in software)로 제공될 것이다. 따라서 일반 사용자는 그 생성기의 존재에 대한 인식 없이 자연스럽게 자신의 모국어로 된 웹문서에 접근할 수 있다.

I.3 본 논문의 구성

II.1 장에서는 먼저 UNL의 정의와 구조가 소개된다. 간단한 독일어 문장이 어떻게 UNL로 표현될 수 있는지 UNL의 정의와 더불어 보여질 것이다.

II.2 장에서는 독일어 UNL-분석기의 구조가 소개된다. 독일어 UNL-분석기를 구성하는 여러 개의 하부 모듈에 대해서도 간단히 언급될 것이다. 독일어 UNL-분석기는 독-영 기계 번역 시스템을 기반으로 구현되었는데, 어떻게 현존하는 시스템이 다른 용도로도 이용될 수 있는지도 논의될 것이다. 또한 UNL-분석기의 역할 중 가장 중요한 것 중의 하나인 입력 독일어 단어에 대한 올바른 개념(Konzept 혹은 UNL용어로는 Universal Word)을 선택하는 방법론도 소개된다.

II.3 장에서는 II.2 장에서 이론적으로 소개된 부분을 보충하기 위하여 한 독일어 예제 문장이 UNL로 변형되는 구체적인 과정을 소개한다.

결론으로 현 시스템의 수준 및 문제점과 앞으로의 과제가 언급될 것이다.

II. 본론

II.1 UNL

독일어나 한국어와 같은 자연 언어가 갖고 있는 표현력을 하나의 중립 언어로 완벽히 대체한다는 것은 거의 불가능한 일이다. 따라서 자연 언어로 된 문장이 내포할 수 있는 미세한 문체적 정보나 함축, 전제와 같은 정보가 UNL로의 변형 과정에서 상실되는 것은 현재의 전산 언어학(Computerlinguistik)적 수준에서 볼 때 거의 불가피한 실정이다. 그러나 한 문장이 지시하는 명제적 의미(propositionale Bedeutung)는 어느 정도 정확하게 형식 언어에 의해 표현될 수 있다. 이론 언어학적인 측면에서 본다면 몬테규 의미론(Montaguesemantik), 상황 의미론(Situationssemantik), 담화 표상 이론(Diskursrepräsentationstheorie)등이 그러한 역할을 할 수 있겠으나, 전산 처리(Elektrische Datenverarbeitung)라는 관점에서 본다면, 이러한 이론들은 부적합한 면이 있다.[4]

UNL에서 자연 언어의 한 문장이 지시하는 명제는 두 개의 개념, 혹은 컨셉트(Konzept)와 그 컨셉트들 사이의 관계라는 이항 관계(binäre Relation)로써 표기될 수 있다. 예를 들어 'Hans küßt Maria'라는 문장이 지시하는 명제 속에는 Hans, Maria, küssen이라는 개념들이 서로 관련을 맺고 있는데, küssen의 행위자는 Hans이며, küssen의 대상자는 Maria라고 볼 수 있다. 이러한 관계를 축약된 UNL식 표기법으로 표현한다면 (1)과 같을 것이다.

---

4) 몬테규 의미론, 상황 의미론, 담화 표상 이론 등이 전산 처리의 목적으로 사용되는 것이 불가능하지는 않다. 현재 DFKI가 중심이 되어 독어와 일어간의 자동번역 시스템의 개발을 위해 진행 중인 VERBMOBIL 프로젝트에서는 독어와 영어, 영어와 일어간의 인터페이스를 위해 중간 레벨로 담화 표상(Diskursrepräsentation)을 사용하고 있다.

(1) agt(kiss,hans)
obj(kiss,maria)

(1)에서 볼 수 있듯이, UNL은 agt, obj와 같은 '관계 레이블'과, kiss, hans, maria와 같은 '개념'들간의 이항 관계로 이루어져 있다. 관계 레이블은 간단히 생각하면 지배 결속 이론의 의미역(thematische Rolle) 개념 혹은 격문법의 격(Kasus) 개념과 유사한 듯하지만, 반드시 그렇지만은 않다. 왜냐하면 실제 세계에서 접하는 언어 정보들에서 각각의 구성 성분이 서로 어떠한 관계를 맺고 있는지를 명시하기 위해서는 일반 언어학에서 주로 사용되는 의미역들, 예를 들어 agent, theme, instrument, location ... 등만으로는 부족하기 때문이다. 다음의 예문 (2)에서 '*ISDN*'과 '*(Integrated Services Digital Network)*'간의 관계는 일반적인 의미역으로써는 표현하기가 어렵다고 할 수 있다.

(2) Dabei stieg die Zahl der Kunden zu Beginn der Geschichte von *ISDN (Integrated Services Digital Network)* nur sehr langsam an.

현재 UNL시스템에서 사용하는 관계 레이블의 수는 41개로, 각각의 레이블들은 명확하게 정의되어 있고 서로 중복되는 부분은 없다. 그러나 이러한 관계레이블의 정의는 대다수의 언어학자들이 짐작할 수 있듯이, 시스템의 개발에서 가장 논란의 여지가 많고 중요한 부분이다.

관계 레이블의 논항(Argument)으로 사용되는 것은 개념(Konzept) 혹은 UW(Universal Word)이다. 이 개념들은 의미적 모호성(Ambiguität)의 해소를 위하여 지식 기반(Wissensbasis)안에 계층적(hierarchisch)으로 나열되어 있다. 예를 들어 독일어의 'Tor'라는 단어는 '이동하는 개체가 드나들 수 있는 공간'이라는 개념과 '축구 경기에서 공이 골라인을 넘은 것'이라는 두 개의 개념을 갖고 있는데, 이들은 UNL-지식 기반 상에서 그림 1과 같이 조직되어 있다.

모든 의미 유형은 지식 기반 속에 계층적으로 나열되어 있고, 모든 개념(혹은 UW)은 역시 지식 기반 속에서 의미 유형에 따라 나열되어 있다. 다시 말하면 독일어 'Tor'에 해당하는 두 가지 개념 'gate', 'goal'은 그것이

갖는 의미 유형에 따라 (3)에서 보는 것처럼 구분된다.[5]

(3) gate(icl⟨aperture)
goal(icl⟨state)

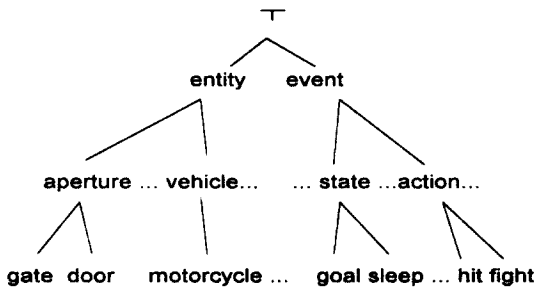

그림 1 UNL-지식 기반(Wissensbasis)

각각의 개념들을 지시하는 스트링(Zeichenkette)들은 위의 예에서 볼 수 있듯이 해당 영어 단어와 동일하지만 모든 영어 단어가 항상 어떠한 개념을 지시하는 것은 아니다. 속어나 여러 단어로 되어 있는 숙어 등의 경우에는 동일한 개념을 지시하는 일반적인 단어가 대표 개념으로 사용된다. 또한 영어에는 존재하지 않지만 다른 언어에 존재하는 개념이 있는 경우에는, 일련의 검토 과정 후에 지식 기반에 추가로 등록된다. 현재 UNL-지식 기반 속에는 약 1백만 개의 UW(개념)들이 의미 자질에 따라 계층적으로 나열되어 있다.

한 문장 안의 기능 형태소들이 지니는 시제 정보(Tempus)나 명사의 수(Numerus), 한정성(Bestimmtheit)등과 같은 정보는 '속성(Attribut)'으로서 @기호 다음에 표기된다.[6] UNL에서 '속성'의 값으로 올 수 있는 것들은 시제와 수, 한정성에 관련된 자질 뿐 아니라 동사의 Aktionsarten,

---

5) 'icl⟩aperture'는 어떤 개념이 지식 기반 상에서 'aperture' 밑에 위치함을 의미한다.

6) 여기서의 'Attribut'는 '자질-값(Attribut-Wert)' 구조의 'Attribut'와는 다른 개념으로, UNL 고유의 용어이다.

화법 조동사의 용법(가능, 필연, 의무, 허가 ...) 등에 관련된 정보가 있다. (1)의 예에서 'küßt'의 시제는 현재 시제이고 이것은 (4)와 같이 표기된다.

(4) kiss(icl>action)@present

지금까지 살펴본 예를 정리하여 UNL의 통사 규칙(Syntax)을 간단히 형식화하면 다음과 같다.[7]

(5) 이항 관계::= 관계 레이블 (개념1, 개념2)
개념::= UW (icl>SEMANTIK)@attribut1@attribut2 ... @attribut5

41개의 관계레이블들은 다음의 예에서와 같이 명시적으로 정의되어 있다.

(6) Definition: 'Via' defines an intermediate place or state of an event
이항 관계의 예: via(go(icl>event), mannheim(icl>city))
예문: "Der Zug fährt von Saarbrücken über Mannheim nach München"

(6)의 예에서 보이는 바와 같이 이항 관계들은 술어 역할을 하는 관계 레이블의 정의에 의해 해석된다.

II.2. 독일어 UNL-분석기

독일어 UNL-분석기는 독일어 문장을 자동으로 UNL로 바꾸어 주는 역할을 한다. 독일어 UNL-분석기는 여러 가지 하위 모듈로 구성되어 있는데, 그 중 가장 중요한 모듈은 CAT2 독-영 기계 번역 시스템이다.[8] CAT2는 1980년대 후반 유럽 연합(EC)의 다국어 자동 번역 시스템 개발

---

7) 저자는 본 논문에서 프로젝트 위임 기관인 IAS연구소와의 계약 사항 준수를 위하여 예로 제시한 내용 이상의 UNL 세부 사항은 기술할 수 없음을 밝힌다.

8) CAT2 형식 문법에 기반한 한-영, 한-독 기계 번역 시스템에 대해서는 Choi (1995), Hong (1998a,b) 참조.

프로젝트인 EUROTRA에서 개발된 CAT시스템의 후속 모델로서, 트랜스퍼(Transfer) 방식의 기계 번역 시스템이다. 현재 CAT2 독-영 번역 시스템은 규칙 기반(regelbasiert) 시스템과 예제 기반(beispielbasiert) 시스템이 연결된 복합적(hybrid) 성격을 띄고 있다. EDGAR라고 불리는 이 예제 기반 시스템은 입력 문장에서 'Jürgen Klinsmann', 'Gerhard Schröder', 'Johannes Brahms' 등과 같은 고유명사를 예제 데이터 베이스에서 찾게 되면 하나의 분석 단위로 인식하여 분석 과정의 복잡성을 덜어 주는 역할을 한다. 번역 엔진은 전통적인 트랜스퍼 방식과 예제 기반 시스템이 상호 보완적으로 작용하여 구동된다. 입력 문장의 고유명사, 숙어나 관용구는 예제 기반 시스템에 의하여 인식, 분석되고 나머지 부분은 규칙 기반 시스템에 의하여 분석된다. 이때 예제 기반 시스템에 의해 분석된 부분은 규칙 기반 시스템에 의해 더 이상 분석되지 않는다. 우리는 현재 UNL 분석기의 구현을 위해 예제 기반 방식의 장점을 좀더 충분히 살리는 방법을 실험 중이다.[9]

독일어 문장이 입력되면 문장을 구성하는 단어들의 형태소가 분석된다. 이 과정은 MPRO라는 형태소 분석기에 의해 이루어진다.[10] 일단 형태소가 분석되면 각 단어는 자질-값(Attribut-Wert) 구조의 정보를 지니게 된다. 예를 들어 'singt'라는 동사는 형태소 분석 후에 {lex=singen, head={tense =present, num=sg}}와 같은 자질 구조를 가지게 된다. 일반적인 형태소 분석 과정의 문제점은 형태소 분석시 통사적 환경을 고려하지 않는다는 데 있다. 예를 들어 관사 'der'는 ({lex=det, head={gen=masc, num=sg, case=nom}};{lex=det, head={gen=fem, num=sg, case=(gen; dat)}};{lex=det, head={num=pl, case=gen}})와 같이 성(gen), 수(num), 격(case)에 대해 잉여적인 정보를 지니고 있다.[11] 이러한 모호성(Ambiguität)은 통사 분석시 불필요한 과정을 유도하여 시스템의 효율성을 떨어뜨린다. 이러한 문제를 해결

9) 규칙 기반 시스템과 예제 기반 시스템의 결합을 위한 실험은 Carl et al. (1998)에 기술되어 있다.

10) MPRO에 대해서는 Maas (1994) 참조.

11) CAT2 포말리즘상에서 ';'은 Disjunktion의 의미를 지닌다.

하기 위해 형태소 분석 후처리기(Post- morphological processor)인 KURD가 사용된다.[12] KURD는 간단한 통사 정보를 가지고 자질 구조의 모호성을 해결한다. 예를 들어 'auf'란 단어는 문장의 맨 끝에 문장 종결 부호 '.'와 함께 사용되면 전치사가 아니라 동사의 분리 전철일 가능성이 높으므로 전치사로서의 자질 구조가 삭제된다.

형태소 분석이 끝나고 형태소 분석 후처리까지 끝난 입력 문장은 통사 구조 분석의 과정으로 전달된다. 현재 필자는 독일어 문장의 분석과 UNL 표현의 생성을 위해 독일어 문장을 문장의 유형에 따라 간략하게 분석하고 번역하는 패턴매칭(Pattern-Matching)의 방식을 따르고 있다. 다시 말하면, 독일어 문장의 분석을 위한 문법 규칙이 핵심어 주도 구조 문법(HPSG)에서와 같이 몇 개의 원리(Prinzipien)와 소수의 언어 개별적 규칙으로 구성되는 것이 아니라, 독일어 말 뭉치(Korpora)를 조사하여 얻게 된 특정 문장 유형(Satztypen)들, 예를 들어 V2-문장(V2-Satz), 기능동사 구문(Funktionsverbgefüge), 관계문(Relativsatz) 등과 같은 문장의 유형별로 구성되어 있다. 이러한 방식은 문법(Grammatik)의 일반화(Generalisierung)와 공리화(Axiomatisierung)를 추구하는 독어학자나 언어학자의 입장에서 보면 불만족스러운 면이 있지만, 전산 처리라는 면에서 보면 여러 가지 장점을 지닌 방식이라고 할 수 있다. 전산 처리를 위한 문법 기술시 어느 정도의 일반화는 반드시 필요하지만 그것이 지나치면 시스템의 불필요한 처리 과정을 유도하게 되어 효율성을 떨어뜨리는 결과를 초래한다. 예를 들어 문법 규칙이 간단하면 그 만큼 많은 정보가 어휘부(Lexikon)에 수록되어야 하는데 어휘부라는 데이터 구조가 커지면 커질수록 처리 속도는 떨어지게 된다. 이러한 문제를 해결하기 위해 나름대로 어휘부 기술을 간략하게 하는 방법, 예를 들어, 매크로(Macro) 혹은 다중 상속(Mehrfachvererbung)등과 같은 방법론이 제안되고 있긴 하지만, 전체적으로 볼 때 처리의 효율성이 떨어지게 되는 점에는 큰 변화가 없다. 물론 마찬가지로 모든 문장 유형에 대해 규칙을 만들게 된다면, 일반화라는 방법을 통하여 얻게 되는 장점을 잃게 되고, 규칙의 관리가 어려워지고

12) KURD에 대해서는 Carl et al. (1997)참조.

규칙을 탐색하는데 시간이 걸리게 되는 문제를 초래하게 된다. 따라서 가장 좋은 방법은 적당한 일반화와 효율성을 떨어뜨리지 않는 범위 내에서 문법 규칙을 작성하는 것이라 보고, 이러한 방법으로 현재 독일어 분석 문법을 관리, 추가하고 있다.

입력 문장이 올바로 분석되면 분석 수형도(Baum)가 생성되고 이 독일어 분석 구조는 UNL쪽으로 전이(Transfer) 된다. Ⅱ.1 장에서 볼 수 있었듯이, UNL의 UW들은 영어와 동일한 스트링(Zeichenkette)을 가지므로 이 과정에서 독-영 기계 번역 시스템이 사용될 수 있다. 각 문장 유형에 대한 독일어 문법 규칙은 그에 상응하는 영어 문법 규칙과 연결되어 있다. 이러한 사상(Abbildung)관계를 통해 입력 독일어 문장에 해당되는 영어 문장 구조가 생성된다. (그림 2 참조)

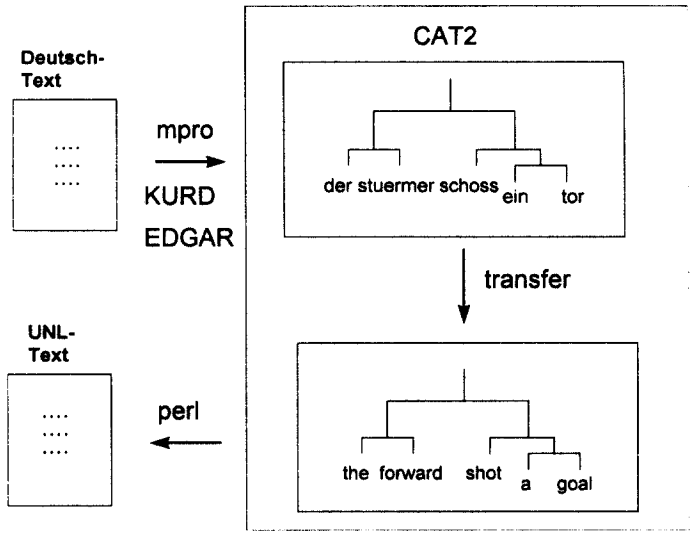

그림 2 독일어 UNL-분석기의 구조

영어 문장 구조의 생성시 가장 중요한 점은 입력된 독일어 문장의 단어들에 대한 올바른 UW(혹은 영어 단어)가 생성 되야 한다는 점이다. 이 문제는 기계 번역에서 가장 중요하고 해결하기 힘든 과제 중의 하나인 역

어 선택의 문제와 크게 다를 바가 없다. 즉, 독일어 단어 'Tor'는 문맥에 따라 영어로는 'goal' 또는 'gate'로 번역될 수 있는데, 어떤 상황에서 어떤 역어를 선택해야 하는가가 중요한 문제이다. 이 문제의 해결을 위해 우선 단어의 통사, 의미 정보가 사용될 수 있다.[13] 그러나 모호성을 해소하기 위해 통사, 의미 정보만으로는 부족한 경우가 많다. 우리는 이러한 문제를 해결하기 위해 텍스트의 주제 분야(경제, 컴퓨터, 의학, 스포츠 등)와 어휘의 빈도 수를 이용한 통계적인 방법을 사용하고 있다.[14] 간단히 말해, X라는 독일어 어휘가 영어에서 A와 B로 번역될 수 있다고 하고, '스포츠'라는 주제 분야에서 A라는 단어가 B라는 단어보다 영어 말 뭉치(Korpus) 상에서 통계적으로 출현 빈도 수가 높을 때, X라는 단어가 '스포츠' 분야에서 사용되었다면, X를 A로 번역하는 것이 옳을 가능성이 높다는 것이다. 독일어 단어 'Tor'의 예를 다시 들자면, '스포츠' 분야에서 'goal'의 출현 빈도 수가 'gate'의 출현 빈도 수보다 높다는 가정 하에 'Tor'의 역어로 'goal'이 선호된다.[15] 이러한 목적을 위해 사용되는 영어나 독일어 말 뭉치는 그 크기가 클수록 정확한 정보를 제공할 가능성이 높은데, 이러한 말 뭉치 또한 주제 분야에 대한 정보와 함께 자동으로 획득될 수 있다. 현재 우리가 채택하고 있는 방법은 인터넷상의 번역 서비스를 통한 말 뭉치의 획득이다. IAI의 홈페이지 (http:// www.iai.uni-sb.de)에서 접근할 수 있는 자동번역 서비스를 통해 사용자는 번역을 원하는 텍스트를 주제 분야에 대한 정보와 함께 입력하게 된다. 이렇게 입력된 정보는 모든 어휘의 빈도수

---

13) 단어의 의미 정보를 이용한 동사의 역어 선택에 관한 방법론에 대해서는 김성묵 (1993) 참조.

14) 역어 선택 문제의 해결을 위한 통계적 접근 방법에 대한 구체적인 내용은 Streiter et al. (1999)에 기술되어 있다.

15) 한 언어로만 된 말 뭉치(monolinguale Korpora)에서 추출한 통계 정보가 오히려 트랜스퍼에 장애가 되는 경우도 있다. 예를 들어 독일어 입력 문장 'Der alte Präsident traf heute in Albanien ein'에서 형용사 alt는 영어로는 'old' 혹은 'aged'로 번역될 수 있는데, 한 주제 분야의 말 뭉치에서 우연히 'aged'가 'old'보다 출현 횟수가 높다면 통계 지식 모듈은 'aged'를 'alt'의 역어로 잘못 제안 할 것이다. 이러한 문제를 해결하기 위해서는 통계 정보를 두 개의 언어로 구성된 말 뭉치(bilinguale Korpora)에서 추출하는 방법이 고려되어야 한다.

계산 과정을 마친 후에 그 결과와 함께 데이터베이스에 저장된다. 이러한 방법으로 상당히 깨끗한 대량의 말 뭉치에 인간의 개입 없이 접근할 수 있게 된다.

역어가 선택되고 나면 독일어 어휘가 지니고 있는 시제(Tempus), 수(Numerus), 한정성(Bestimmtheit), 의미역(thematische Rolle), 문법관계(grammatische Relationen)등과 같은 통사, 의미 정보가 UNL쪽으로 전달된다. CAT2 형식 문법에서 사용하는 통사, 의미 정보의 자질 및 값의 이름이 UNL에서 사용하는 자질과 값의 이름과 다른 경우가 있으므로, 이들을 UNL식으로 변형하는 과정이 필요하다. 이 목적을 위하여 예를 들어 다음과 같은 CAT2 형식 문법의 규칙(f-Regel)[16]이 손쉽게 사용될 수 있다.

(5)

```
f_present = {head = {cat = verb, tense = pres}, unl = {a1 = '. @present'}}. {}.
f_plural = {head = {cat = noun, number = pl}}, unl = {a1 = '. @plural'}}. {}.
```

위의 보기에서 f_present 규칙을 통해 현재 시제를 갖는 동사는 UNL의 첫 번째 Attribut(a1) 값으로 'present'를 갖고, 복수 명사는 f_plural 규칙을 통해 'plural' 값을 a1의 값으로 갖게 된다.

영어 문장 구조의 각 터미널 노드는 unl이라는 자질을 지니고 이 자질은 위와 같은 방식으로 UNL의 생성에 관련된 정보를 값으로 갖는다. unl자질에 각 정보들이 모아지면, 이 정보들은 UNL표현의 생성을 위한 하위 프로그램의 입력이 된다. 이 하위 프로그램은 Perl-프로그래밍 언어로 작성되어 있으며 unl자질의 값을 입력 값으로 받아 관련된 정보만 추출, 출력한다. 이 장에서 이론적으로 제시된 부분들은 다음 장에서 구체적인 예와 함께 설명된다.

---

16) CAT2 시스템에서 f-Regel은 자질 구조(Merkmalstruktur)내의 값(Wert)을 검사하거나 제어하고, 또한 디폴트 값을 할당하는 역할을 한다.

II.3 예문 분석

이 장에서는 앞장에서 언급된 부분들이 실제로 어떻게 적용되는지를 하나의 예문을 통해 살펴보도록 하겠다. 입력 문이 문장의 주제 분야 정보와 함께 주어지면 우선 문장의 형태소 분석이 이루어진다.[17]

(6) Eingabe: Der gefährliche Stürmer schoß in der Saison 20 Tore. (sport-domäne)
Morphologische Analyse: {der, gefaehrlich, stuermer, schiessen, in, der, saison, 20, tor, punkt}

형태소 분석은 입력 문을 분석 단위에 따라 분절하는 역할뿐만 아니라 각 어휘에 대해 적합한 자질들을 부여한다. 위의 예문에서 예를 들어 'Stürmer'는 다음과 같은 자질-값 구조를 갖게 된다.

(7) {string=stuermer, lex=stuermer, head={cat=n, ({num=sing, case=(acc;dat;nom), gen=male};{num=plu, case=(acc;gen;nom)})}}

형태소 분석이 끝난 문장은 형태소 분석 후처리를 위해 KURD시스템으로 보내진다. 위의 예에서 전치사구 'in der Saison'의 'der'는 성, 수, 격 정보에 대해 다음과 같은 정보를 지닌다.

(8) {string=der, lex=det, head={cat=d, ({num=sing, case=nom, gen=male};{num=sing, case=(gen;dat), gen=female};{num=plu, case=gen})}

이러한 모호성의 해소를 위해 KURD 시스템은 우선 입력 문장을 간략하게 분석(shallow parsing)한다.[18] 이 시스템은 'in der Saison'을 전

17) 현재 개발 중인 UNL-분석기 프로토타입(Prototyp)에서는 사용자가 입력 문과 함께 문장의 주제 분야에 관한 정보도 직접 입력을 해야 한다. 그러나 이 시스템이 상용화될 경우, 문장의 주제 분야를 사용자가 직접 입력하지 않더라도 시스템이 스스로 주제 분야를 인식하는 방법도 고려되어야 한다.

치사구로 분석하고, 전치사구 내의 명사구 'der Saison'에서 관사와 명사가 성, 수, 격에 대해 일치(Kongruenz)관계에 있어야 한다는 정보를 이용하여 관사 'der'의 자질 구조 내에서 {num=sing, case=nom, gen=male}와 {num=plu, case=gen}을 삭제한다. 그리고 전치사구의 핵심어(Kopf)인 'in'이 3격(Dativ) 또는 4격(Akkusativ)명사구를 하위범주화한다는 정보를 이용하여 {num=sing, case=(gen;dat), gen=female} 내에서 case=gen값을 삭제한다. 결과적으로 이 문장의 분석을 위하여 불필요한 정보들은 모두 삭제되고 필요한 정보만 통사 분석의 단계로 넘어가게 된다.

위의 예문은 독일어에서 가장 많이 발견할 수 있는 V-2(Verbzweit) 문장이다. 위상적 장이론(Topologische Feldtheorie)에 따르면 동사의 앞 위치인 전장(Vorfeld)에는 일반적으로 문장의 주어나 부사어 혹은 동사의 과거분사형이 등장한다.[19] 그리고 중장(Mittelfeld)에는 명사구(NP)와 전치사구(PP)등 비교적 여러 가지 성분이 자유롭게 올 수 있다. 이러한 규칙성을 이용하여 다음과 같은 통사 규칙[20]이 제안된다.

(9) S—〉 NP, V, *ADVP, NP, punkt
NP —〉 det, ^adj, noun
NP —〉 num, noun
ADVP —〉 prep, NP

이 규칙에 의해 위의 문장은 다음과 같이 분석된다.[21]

---

18) Shallow Parsing은 문장 전체의 구조를 분석하는 것이 아니라 문장 내의 명사구나 전치사구 등과 같은 구구조를 간략하게 인식하여 태깅(Tagging)하는 작업인 점에서 일반적인 Parsing과는 다르다. 이러한 Shallow Parsing은 형태소 모호성의 해소 작업뿐만 아니라, 예제 기반 번역을 위한 예제들의 정열(Alignement)에도 사용될 수 있다.

19) 본 논문에서는 전장 위치에 어떤 성분이 등장할 수 있는가에 대한 문제는 다루지 않기로 한다. 위상적 장이론에 대해서는 신수송 (1988:82) 참조.

20) UNL- 분석기의 실제 구현(Implementierung)에서는 통사 규칙이 예에서 보는 바와 같이 문맥 자유 규칙(Kontextfreie Grammatik)의 형태를 띄는 것이 아니라, CAT2 내부의 규칙을 따른다.

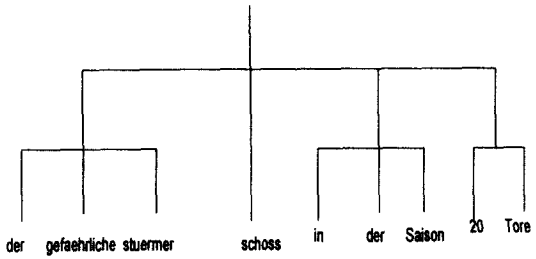

그림 3 독일어 입력 문의 통사 구조 분석 수형도

(그림 3)의 독일어 통사 분석 구조는 해당되는 영어의 구조로 다음의 규칙에 의해 전이(Transfer)된다.

(10) De:(S—〉 NP, V, *ADVP, NP, punkt) 〈—〉 Eng:(S—〉 NP, V, NP, *ADVP, punkt)

톱-다운(Top-Down)방식으로 적용되는 (10)의 규칙은 영어 문장의 생성을 위해 다시 각각의 구성 성분(NP, ADVP ...)들에 대한 전이 규칙(Transfer-Regel) (11)을 필요로 한다.

(11) De:(NP —〉 det, ^adj, noun) 〈—〉 Eng:(NP —〉 det, ^adj, noun)
De:(NP —〉 num, noun) 〈—〉 Eng:(NP —〉 num, noun)
De:(ADVP —〉 prep, NP) 〈—〉 Eng:(ADVP —〉 prep, NP)
De:(V) —〉 Eng:(V)

이러한 규칙들은 독일어 문장의 구성 성분들, der gefährliche Stürmer, schoß, in der Saison, 20 Tore 를 각각 the dangerous striker, shot, in the season, 20 goals로 번역하는데 사용된다. 이 번역 과정에서 가장 중요한 것은 독일어 단어에 대한 올바른 역어를 선택하는 것이다.

---

21) 물론 이 문장에서 'schoß in der Saison 20 Tore'를 하나의 동사구(VP)로 분석할 수도 있다. 이 논문에서는 문장의 주어(der Stürmer)와 목적어(20 Tore)가 같은 층위에 위치하는 것으로 간단히 분석한다.

예를 들어 독일어 단어 'schießen'은 영어로는 'shoot'이나 'fire'등으로 번역될 수 있는데, 위 예문의 경우에는 'shoot'으로 번역되는 것이 적합하다. 올바른 역어 선택을 위하여 앞장에서 언급한 바와 같이 통계적 방법을 사용하는데, 이 경우 스포츠 분야에 관련된 영어 말 뭉치에 대한 조사 결과로 얻어진 통계자료가 사용된다. 이 자료에 따르면 스포츠 분야에서는 'shoot'이라는 단어가 'fire'라는 단어보다 자주 등장하므로 'schießen'을 'shoot'으로 번역하게 된다. CAT2 시스템의 사전에서 각 단어는 역어에 대한 정보가 (12)에서와 같이 trans자질 안에 수록되어 있는데, 통계 정보가 보충되면 데이터 구조가 (13)과 같은 모습을 띄게 된다.[22]

(12) {lex=schiessen, head={cat=v, proc=activity},
subcat={a={AGENT},b={THEME}},
trans={en={t=(fire;shoot;swoop)}}}

(13) {lex=schiessen, head={cat=v, proc=activity},
subcat={a={AGENT},b={THEME}},
trans={en={sbj=spt, t=(shoot;fire;swoop)};
{sbj=soc, t=(fire;shoot;swoop)};
{sbj=ecn, t=(fire;shoot;swoop)}}}

trans자질 안에 값으로 주어진 단어들이 각 주제 분야에서 독일어 단어에 대한 후보 역어들이다. 이들 가운데 가장 왼쪽에 위치한 역어가 제일 먼저 시도된다.

우리가 추구하는 바가 단지 기계 번역이라면 이렇게 생성된 영어 구조를 형태소 생성 과정을 통해 완전한 문장의 형태로 출력하면 끝나겠지만, UNL-표현의 생성을 위해서는 한 단계의 절차가 더 필요하다. 즉, 이 문장의 명제적 의미를 결정하는데 필요한 요소들과 그에 딸린 정보들을 수집하여야 한다. 이 문장의 명제적 의미 결정에 참여하는 세 가지 개념, 'Stürmer(striker)', 'schießen (shoot)', 'Tor(goal)'이 어떠한 관계레이

22) (13)의 사전 목록에서 sbj는 Subjekt-Domäne, spt는 sports, ecn은 economy, soc은 society의 약어이다.

블에 의해 관련을 맺게 되는지 살펴보자. 우선 이 문장의 주어와 목적어인 'Stürmer(striker)'와 'Tor(goal)'은 동사 'schießen(shoot)'의 하위범주화(Subkategorisierung)에 의해 의미역(thematische Rolle)과 함께 UNL의 관계레이블을 할당받는다. 이 경우 관계레이블은 다음의 규칙 (14)에서 보이는 바와 같이 할당받은 의미역에 따라 결정된다.

(14) f_agt_agt={role=agt, head={cat=n}, unl={rel=agt}}.{}.
f_theme_obj={role=theme, head={cat=n}, unl={rel=obj}}.{}.

이 규칙들은 각 명사구들에 적용되는데, 어떤 명사구가 의미역 'agt'를 지니면 UNL 관계레이블로는 'agt'를 갖게 되고, 'theme'의 경우에는 'obj'를 갖게 된다. 이렇게 하여 주어와 목적어는 자신의 unl자질 속에 자신이 문장의 술어와 관련된 관계레이블의 값을 갖게 된다. 그리고 각 관계레이블의 첫 번째 논항(uw1)값으로는 문장의 동사 개념(Konzept), 즉 'shoot'을 'VAR'변수의 공유(variable binding)를 통해 갖게 된다.

(15) striker:{lex=striker, lex=LEX1, unl={rel=agt, uw1=VAR, uw2=LEX1}}
shoot:{lex=shoot, lex=VAR, unl={uw1=VAR, uw2=nil}}
goal:{lex=goal, lex=LEX2, unl={rel=obj, uw1=VAR, uw2=LEX2}}

이 문장의 주어와 목적어를 이루는 명사구에서는 각각 형용사와 수사가 명사와 결합되어 있는데, 부가어적 형용사(attributive Adjektive)와 수사(Zahlwörter)는 문장의 명제적 의미 결정에 있어서 빼 놓을 수 없는 요소들이다. 명사구 내의 부가어적 형용사는 일반적으로 관계레이블 값으로 'mod(ifier)'를 부여받는다.[23] 수사는 관계레이블 'quan(tifier)'를 부여

23) 모든 부가어적 형용사가 수식 받는 명사와 'mod'관계로 연결될 수는 없다. 예를 들어 명사구 'ein starker Raucher'가 UNL로 'mod(strong,smoker)'로 분석된다면, 이 UNL표현은 불어 UNL-생성기를 통하여 'un gros fumeur'가 아니라 'un fort fumeur'로 잘못 생성될 것이다. 따라서 여기서 'stark'의 의미는 단순한 Modifikator가 아니라 수식 받는 명사의 정도를 강조하는 기능어(lexikalische Funktion)로 분석되어야 한다.

받게 된다. 'dangerous'와 '20'의 unl-자질은 (16)과 같은 값을 갖게 된다. dangerous는 변수 LEX1값의 공유를 통해 uw2의 값으로 'striker'를 갖게 되고, 수사 20은 마찬가지로 LEX2를 통해 uw2값으로 'goal'을 가지게 된다.

(16) dangerous: {lex=dangerous, lex=LEX3, unl={rel=mod, uw1=LEX3, uw2=LEX1}}
20: {lex=20, lex=LEX4, unl={rel=quan, uw1=LEX4, uw2=LEX2}}

전치사구 'in the season(= in der Saison)'에서 전치사 'in'은 이 경우 시간을 나타낸다. 전치사들은 단순히 기능적 품사로 쓰일 경우와 (예를 들어 'warten auf den Mann'의 'auf') 의미를 지니는 경우('warten in dem Zimmer'의 'in')로 분류할 수 있는데, UNL-표현의 생성을 위해서는 의미를 지니는 경우만 고려된다. 전치사의 의미를 올바로 분석하는 것도 중요하고 어려운 작업이지만, 현재 우리는 전치사가 취하는 보족어 명사의 의미 정보를 이용하는 방법을 택하고 있다. 즉, 전치사 'in'은 시간, 공간, 방향 등의 의미를 가질 수 있는데, 다음에 오는 명사 'season'은 시간적 의미를 지니므로 전치사 'in'과 보족어 'season'은 시간의 관계를 지닌다고 분석된다. 또한 'in the season'은 의미적으로 문장의 동사 'shoot'을 수식한다((17) 참조).

(17) in: {lex=prep, unl={rel=time, uw1=LEX1, uw2=LEX5}}
season: {lex=season, lex=LEX5, unl={uw1=LEX5, uw2=nil}}

이러한 방식으로 UNL-표현의 생성에 필요한 정보들이 각 터미널 노드들의 unl-자질 내에 저장된다((18) 참조).

(18) dangerous: {lex=dangerous, lex=LEX3, unl={rel=mod, uw1=LEX3, uw2=LEX1}}
striker: {lex=striker, lex=LEX1, unl={rel=agt, uw1=VAR, uw2=LEX1}}
shoot: {lex=shoot, lex=VAR, unl={uw1=VAR, uw2=nil}}
in: {lex=prep, unl={rel=time, uw1=LEX1, uw2=LEX5}}
season: {lex=season, lex=LEX5, unl={uw1=LEX5, uw2=nil}}

20: {lex=20, lex=LEX4, unl={rel=quan, uw1=LEX4, uw2=LEX2}}
goal:{lex=goal, lex=LEX2, unl={rel=obj, uw1=VAR, uw2=LEX2}}

최종적으로 Perl-프로그램이 unl-자질들에 적용되어 다음과 같은 UNL-표현이 생성된다.[24]

(19) mod(dangerous,striker)
agt(shoot,striker)
time(shoot,season)
quan(20,goal)
obj(shoot,goal)

III. 결론

이 논문은 UNL프로젝트에 대해 간단히 소개하고 독일어 UNL-분석기의 구조와 특징을 한 예문을 통해 다루었다. 독일어 UNL-분석의 과정은 대부분의 기계 번역 시스템이 채택하고 있는 분석 과정과 유사하기 때문에 일반적인 기계 번역 시스템이 UNL-분석기로 사용될 수 있음을 CAT2-시스템의 예로 제시하였다.

현재 독일어 UNL-분석기는 V2-문장, 양상 동사 구문(Modalverb-Konstruktionen), 기능 동사 구문(Funtionsverbgefüge), 수동 구문(Passiv- Konstruktionen), 관계문(Relativsatz-Konstruktionen) 등을 분석하여 UNL로 나타낼 수 있다.[25] 이 시스템의 평가(Evaluierung)를 위해 독일어 UNL-분석기를 통해 자동으로 생성된 UNL-표현을 다시 입력으로 하여 독일어 문장을 재생성 해보는 테스트가 현재 DFKI에서 진

24) 실질적으로 각 개념들은 좀 더 자세한 정보들을 지니게 된다. 예를 들어 'shoot'는 'shoot(icl)activity).@past.@entry'와 같이 Aktionsarten, 시제 등에 관련된 정보와 함께 생성된다.

25) 현재 독일어 UNL-분석기는 IAI의 UNL 홈페이지(http://www.iai.uni-sb.de/UNL/ unl-iai.html)에서 온라인으로 테스팅할 수 있으나, 프로젝트의 원활한 진행을 위해 각국의 개발 진들에게만 사용권이 주어져 있다.

행 중이다.

이 시스템의 완성도를 높이기 위하여 해결해야 할 언어학적 난제들은 아직도 산적해 있다. 가장 큰 문제는 문장의 단위를 넘어서는 현상의 수학적 기술, 예를 들어 대용어의 선행어를 자동으로 찾는 알고리즘의 개발(Anaphernresolution), 생략 구문(Ellipsen)[26]과 등위 접속 구문(Koordination)등에서 생략된 요소들을 계산해 내는 알고리즘의 개발 등이다.

또 다른 문제는 사전에 누락된 독일어 어휘와 UW를 적절히 연결하는 것에 대한 문제이다. 컴퓨터 시스템의 사전은 스스로 새로운 어휘를 창출해 내는 능력을 가지고 있지 못한 데에 반하여, 인간은 얼마든지 새로운 어휘들을 어휘 규칙 등을 통해 만들어 낼 수 있다. 실제로 예를 들어 독일어의 경우 복합 명사 생성 규칙 등이 매우 활발하게 적용되어 얼마든지 많은 신조어가 생겨날 수 있다.[27] UNL-시스템은 이러한 어휘들도 그 의미를 분석하여 적당한 UW와 올바르게 연결시켜야 한다. 이러한 문제들의 해결을 위한 여러 독어학자들의 연구가 필요하다고 할 수 있다.

참고문헌

김성묵 (1993): 기계 번역에서의 다의어 동사 처리 연구, 제 27회 어학연구회, 서울대학교 어학연구소

김성묵 (1998): 독일어 명사 합성어의 기계 분석, 서울대학교 독어독문학과 박사학위 논문

신수송 (1988): 현대독어학, 교육과학사

이민행 (1999): 독일어의 어휘부에 대한 연구 -전산 언어학적 접근, 독일문학 제 69집 40권 1호

---

26) Lee (1998)은 독일어에 나타나는 생략(Ellipse)현상을 HPSG이론에 기반하여 기술했다.

27) 독일어의 복합 명사에 대한 기계적 분석은 김성묵 (1998) 참조. 이민행 (1999)은 사전에 누락된 형용사 합성어에 대한 어휘 목록을 자동으로 생성할 수 있는 알고리즘을 제시하였다.

Allied Business (1998), Language Translation, s. 31

Michale Carl, Antje-Schmidt Wigger & Hong, Munpyo (1997): KURD - A Formalism for Shallow Post Morphological Processing, Proceedings of the Natural Language Processing Pacific Rim Symposium 1997 (NLPRS'97) in Phuket, Thailand

Michale Carl, Leonid L. Iomdin & Oliver Streiter (1998): Towards dynamic linkage of Example-Based and Rule-Based Machine Translation, ESSLI '98 Machine Translation Workshop

Choi, Sung-Kwon (1995): Unifikationsbasierte Maschinelle Übersetzung mit Koreanisch als Quellsprache, IAI Working Papers 34, Saarbrücken

Hong, Munpyo (1998a): Treating Multiple-Subject Construction in a Constraint-Based MT-System, Tagungsband der 4. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS -98) in Universität Bonn, Peter Lang

Hong, Munpyo (1998b): Multiple-Subject Construction in a Multilingual MT-System CAT2, Third Conference of the Association for Machine Translation in the Americas, AMTA'98, Langhorne, USA, Springer Verlag

Hong, Munpyo & Oliver Streiter (1999): Overcoming the Language Barriers in the Web: The UNL-Approach, 11. Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung (GLDV'99), Frankurt am Main, Deutschland

Lee, Hae-Yun (1998): Ellipsen in Satzkoordinationen, Peter Lang

Luckhardt, Heinz-Dirk (1987): Der Transfer in der maschinellen Übersetzung, Sprache und Information, Niemeyer

Maas, Heinz-Dieter (1994): Analysis and Translation of

Compound Nouns in Mpro, Proceeding of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition, S.162-172, Geneva

Oliver Streiter, Leonid L. Iomdin, Munpyo Hong & Ute Hauck (1999): Learning, Forgetting and Remembering: Statistical Support for Rule-Based MT, 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI99), Chester, England

Zusammenfassung

Die Implementierung eines UNL-Enconverters für das Deutsche auf der Basis eines maschinellen Deutsch-Englisch Übersetzungssystems

Hong, Munpyo(Saarland Univ.)

In der vorliegenden Arbeit wurden das UNL-Projekt und der sogenannte UNL-Enconverter für das Deutsche vorgestellt. UNL ist ein internationales Kooperationsprojekt, an dem Forschungsinstitute und Universitäten unter anderem aus Deutschland, Frankreich, Japan, China, Italien, und Spanien beteiligt sind. Zur Zeit nehmen insgesamt 13 Länder weltweit an dem UNL-Projekt teil. Um die Sprachbarrieren im Internet zu überwinden, wurde die formale Sprache UNL (Universal Networking Language) konzipiert. Sie ist eine Bedeutungsrepräsentationssprache, die die propositionale Bedeutung von Sätzen zum Ausdruck bringen soll. Die Umwandlung von einer natürlichen Sprache wie z.B. vom Deutschen oder Französischen in die UNL und

umgekehrt werden jeweils von dem sogenannten UNL-Enconverter und dem UNL-Deconverter übernommen.

Der UNL-Enconverter für das Deutsche wird am Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung (IAI) an der Universität des Saarlandes entwickelt und getestet. Das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI) an der Universität des Saarlandes entwickelt den UNL-Deconverter für das Deutsche.

Der UNL-Enconverter für das Deutsche basiert auf dem maschinellen Deutsch-Englisch Übersetzungssystem CAT2. CAT2 ist ein regelbasiertes System, das dem Transfer-Ansatz folgt. Außer dem CAT2-System enthält der UNL-Enconverter verschiedene Submodule: Mpro für die morphologische Analyse, KURD für die post-morphologische Analyse, EDGAR für die Eigennamenerkennung und ein Perl-Program für die UNL-Generierung.

Da UNL das Vokabular für die Konzepte (Universal Words) dem Englischen entnimmt, kann ein maschinelles Deutsch-Englisch Übersetzungssystem verwendet werden. Die wichtigste und schwierigste Aufgabe besteht darin, ein richtiges UW (Universal Word) für deutsche Wörter zu finden. Da ein Wort zuweilen mehrdeutig sein kann, ist es wichtig, den richtigen Sinn des Wortes zu bestimmen. Um die relevante Bedeutung des Wortes in einem Kontext zu ermitteln, können unter anderem seine syntaktischen und semantischen Informationen mit einbezogen werden. In zahlreichen Experimenten hat sich jedoch herausgestellt, daß diese Methode nur begrenzt eingesetzt werden kann. Deswegen verwenden wir zusätzlich noch statistische Infor-

mationen aus einem schon bestehenden monolingualen Korpus. An einem Beispiel verdeutlicht, sähe dies folgendermaßen aus: Angenommen, ein Wort X einer bestimmten Sprache läßt sich mit Y und Z in einer anderen Sprache übersetzen und die Übersetzung Y kommt in einer Subjekt- Domäne S1 häufiger als das Wort Z vor, so ist es sehr wahrscheinlich, daß das Wort X in der Subjekt-Domäne S1 mit Y übersetzt wird.

Der UNL-Enconverter für das Deutsche kann im Moment verschiedene Satztypen wie Funktionsverbgefüge, V2-, Relativsatz-, Modalverb-, und Passiv-Konstruktionen analysieren sowie darüber hinaus richtige UNL-Ausdrücke generieren. Um das System zu evaluieren, werden die von dem Enconverter automatisch erstellten UNL-Ausdrücke an den vom DFKI entwickelten UNL-Deconverter für das Deutsche für die Generierung deutscher Sätze weitergereicht.