독어학 제35집 (2017.6)
PP. 101~123

웹기반 독일어 구문분석기의 활용과 평가

이민행
(연세대)

한 국 독 어 학 회

웹기반 독일어 구문분석기의 활용과 평가*

이민행 (연세대)

국문요약

본 연구는 인지적인 관점에서 마법의 수로 알려져 있는 7개 어휘로 구성된 문장들이 나타내는 독일어 구문들의 스펙트럼을 확인하고자 하는 일반언어학적인 목적과 웹을 기반으로 수행되는 구문분석기(parser)가 어느 정도의 정확도를 보이는 지를 평가하고자 하는 전산언어학적인 목적으로 수행되었다. 전자의 경우, 연구성과를 독일어 교육에 활용할 수 있다는 데서 연구의의를 찾을 수 있으며, 후자의 경우, 독일어 구문분석기의 성능평가에 대한 본격적인 선행연구가 없기 때문에 새로운 연구영역의 개척이라는 점에서 그 의의를 발견할 수 있다.

본 연구를 통해 몇 가지 사실을 새롭게 발견할 수 있었다. 언어심리학자 George Miller의 연구결과를 받아들여 마법의 수인 7개 어휘로만 구성된 원시 코퍼스 StammMI7에 포함된 구문들의 유형을 살펴 본 결과, 그 유형의 스펙트럼이 매우 넓음을 확인했다. 이에 따라서 7개 어휘로 구성된 문장들을 이용하여 독일어 교육을 시도할 경우, 어떤 효과를 거둘 수 있는 지에 대한 후속연구가 필요하다고 생각된다. 이어 의존구조 코퍼스 StammMI7-DpS를 분석한 결과를 살펴보면 분석오류의 비율이 20%에 가까울 만큼 상대적으로 오류가 많음을 확인했다. 이러한 결과는 영어 파서의 경우, 오류율이 9% 미만으로 보고된 것과 비교하면 상대적으로 높은 오류율로 간주된다.

핵심어: 코퍼스, 구문분석기, 의존구조, 오류분석, 오류율

1. 문제제기

스위스 작가 Peter Stamm의 어떤 작품은 다음과 같은 문장들로 시작한다.

(1) Agnes ist tot. Eine Geschichte hat sie getötet.

* 이 논문은 연구년(2016.9-2017.8) 기간에 수행한 연구결과물이다.

1998년에 발표된『Agnes』라는 장편소설에 나오는 이 문장들은 소설을 이끄는 Leitmotiv를 담고 있다. 아래에 제시된 문장들도 Peter Stamm의 작품들 속에 나오는데, 모두 7개 어휘로 구성된 문장이라는 공통점을 지닌다.

(2)

Agnes stand auf und kam zum Sofa.

Als es dämmerte, machten sie ein Feuer.

Andreas war froh, dass nicht gesungen wurde.

Da schlug Beatrice vor, Verstecken zu spielen.

Er wusste, was sie von ihm erwartete.

이 용례들을 통해 등위접속 구문, 부사절, 종속절, zu-부정사 구문 및 간접의문문 등 다양한 구문이 7개 어휘로 표현될 수 있음을 확인하게 된다.

본 연구는 인지적인 관점에서 마법의 수로 알려져 있는 7개 어휘로[1] 구성된 문장들이 나타내는 독일어 구문들의 스펙트럼을 확인하고자 하는 일반언어학적인 목적과 웹을 기반으로 수행되는 구문분석기(parser)가 어느 정도의 정확도를 보이는 지를 평가하고자 하는 전산언어학적인 목적을 수행한다. 전자의 경우, 연구 성과를 독일어 교육에 활용할 수 있다는 데서 연구의의를 찾을 수 있으며, 후자의 경우, 독일어 구문분석기의 성능평가에 대한 본격적인 선행연구가 없기 때문에 새로운 연구영역의 개척이라는 점에서 그 의의를 발견할 수 있다.

이 연구를 위해 7개 어휘로 구성된 문장 700개를 원시코퍼스 StammMI7로[2] 구축한 다음에, 웹을 기반으로 작동하는 Weblicht의[3] 의존구조분석기를 이

1) 1950년대 중반에 프린스턴 대학의 인지심리학자 George Miller가 평균적으로 사람들이 단기기억에 저장하거나 불러내 처리할 수 있는 개체의 수가 7+/-2라고 주장한 이래, 숫자 7이 기호의 인지처리와 관련하여 마법의 수(magic number)라고 불린다. Miller(1956) 참조.

2) 본 연구에서 분석대상으로 삼은 원시코퍼스가 마법의 숫자(magic number) 7과 깊숙이 연관되어 있기 때문에 코퍼스의 명칭을 StammMI7으로 명명하기로 한다. MI는 MagIc의 줄임말이다.

3) Weblicht의 웹사이트 주소는 http://weblicht.sfs.uni-tuebingen.de/이다.

용하여 이 코퍼스에 속한 700 문장의 의존구조를 추출하였다. 의존구조들을 모은 코퍼스를 편의상 StammMI7-DpS로 명명한다. 문장마다 의존구조에 나타난 여러 가지 정보들을 분석하여 해당 분석의 타당성을 검토하고, 오류가 발견될 경우에 그 오류가 어떤 유형의 것인지, 맥락으로 작용한 구문은 무엇인지를 살펴보았다.

2. 원시코퍼스의 구성 및 구문유형

원시코퍼스 StammMI7은 Peter Stamm이 쓴 장편소설 세 편을 스캔한 다음에, OCR로 해독하여 텍스트 파일로 만들고, 이로부터 7개 어휘로 구성된 문장 700개를 선별하여 구축한 것이다. 따라서 원시코퍼스의 규모는 4,900개 어휘이다. 코퍼스의 토대가 된 Stamm의 세 작품은 다음과 같다.[4)]

Stamm, Peter (1998). Agnes. Roman. Arche, Zürich.

Stamm, Peter (2001). Ungefähre Landschaft. Roman. Arche, Zürich.

Stamm, Peter (2007). An einem Tag wie diesem. Roman. S. Fischer, Frankfurt am Main.

세 작품으로부터 추출한 문장은 모두 10,000개가 넘는데, 이 문장들을 길이를 기준으로 오름차순으로 정렬한 다음에 10,000번째 문장까지를 살펴볼 때에 가장 짧은 문장은 하나의 어휘로 되어 있는 반면, 가장 긴 문장은 44개 어휘로 되어 있다. 문장의 길이에 따른 분포는 다음의 표와 같다.

4) 작가 Peter Stamm의 작품과 작품세계 및 줄거리가 위키피디아에 자세히 소개되어 있다: https://de.wikipedia.org/wiki/Peter_Stamm.

문장길이	빈도	누적빈도	문장길이	빈도	누적빈도
1	55	55	23	112	9,453
2	192	247	24	87	9,540
3	473	720	25	82	9,622
4	649	1,369	26	59	9,681
5	852	2,221	27	47	9,728
6	835	3,056	28	49	9,777
7	797	3,853	29	41	9,818
8	719	4,572	30	30	9,848
9	622	5,194	31	27	9,875
10	588	5,782	32	23	9,898
11	546	6,328	33	14	9,912
12	444	6,772	34	9	9,921
13	442	7,214	35	20	9,941
14	399	7,613	36	4	9,945
15	349	7,962	37	12	9,957
16	309	8,271	38	12	9,969
17	234	8,505	39	6	9,975
18	222	8,727	40	6	9,981
19	191	8,918	41	6	9,987
20	160	9,078	42	3	9,990
21	134	9,212	43	3	9,993
22	129	9,341	44	7	10,000

표 1. Peter Stamm 작품의 길이별 분포

이 표를 보면 5개부터 8개 어휘로 구성된 문장들이 많은 것으로 분석된다. 곧, 마법의 수 7에 관한 G. Miller의 주장이 타당하다는 것을 입증하는 좋은 예라고 할 수 있다.

코퍼스 StammMI7에 나타난 구문의 유형을 정리하면 다음 (3)과 같다.

(3)

- 복합명사구
- (단순) 명령문
- 관계문
- 부정사구문
- 생략문
- 외치구문
- 동격구문
- 완료구문
- (단순) 도치구문
- 종속문
- 간접의문문
- 등위접속구문
- 분연속구문
- 비교구문
- 수동구문
- (일반) 단순문

구문의 분포에 대해 살펴보자면, 도치구문이 가장 많이 나타나는 반면, (단순)명령문이 출현빈도가 가장 적다.[5] 다시 말하여 이 코퍼스는 구문의 균형성은 고려하지 않은 코퍼스라고 할 수 있다. 이러한 불균형성은 코퍼스에 포함한 문장의 길이를 7개 어휘로 한정한데서 비롯된다.

3. 의존구조 코퍼스의 구축 및 활용

아래 (4)에 열거된 네 가지 유형의 정보가 의존구조 코퍼스 StammMI7-DpS를 구성하는 700개의 의존구조 각각에 담겨져 있다.

(4)

- 어휘형태(word form) 정보 [Aber, sie, hatte, ...]
- 품사(pos) 정보 [KON, PPER, VAFIN, ...]
- 의존기능(dependency function)/표지(label) [JU, SB, OC, ,,,]
- 레마(lemma) 정보 [aber, sie, haben, ...]

예를 들어 문장 "Aber sie hatte keine Antwort gegeben."의 의존수형도는 그림

5) 코퍼스 StammMI7내 개별 구문의 분포는 후에 구문별 오류율을 논의하는 과정에서 함께 제시된다 [표 7 참조].

1.과 같다.

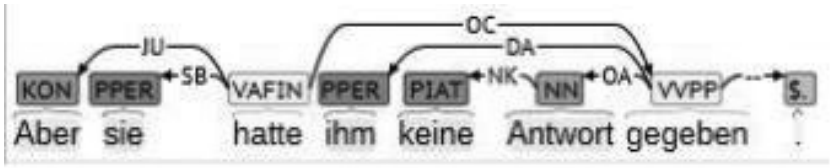

그림 1. 의존구조 수형도 (Weblicht)

위 수형도를 살펴보면, Aber, sie, hatte 등 어휘형태 정보가 맨 아래층에 나타나고, KON, PPER, VAFIN 등 품사 정보가 아래로부터 두 번째 층에 위치해 있으며, JU, SB, OC 등 의존기능 표지가 맨 위층에 자리하여 품사들간의 의존관계를 화살표를 통해 표시하고 있다. 이 수형도에서 화살표가 출발하는 품사가 핵어이고 화살촉이 맞닿는 품사가 의존어이다. Weblicht에서 제공하는 뷰어(viewer)의 제한성 때문에 레마에 대한 정보는 이 의존수형도에는 나타나 있지 않으나 코퍼스 StammMI7-DpS의 데이터 구조에는 포함되어 있다. 반면, 아래의 그림 2에서 확인할 수 있듯이 Tündra에서 활용되는 의존수형도에는 레마에 대한 정보가 명시적으로 표현된다.

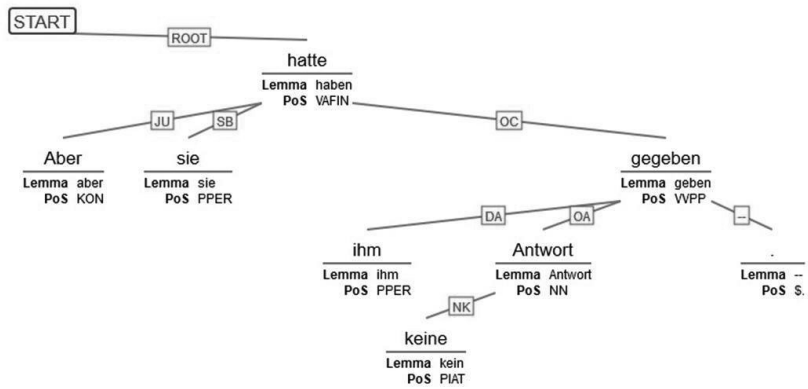

그림 2. 의존구조 수형도 (Tündra)

위에 제시된 수형도에서 교점간의 의존관계를 표현하기 위해 의존기능 표지를 나타내지만, 의존관계의 방향은 수형도상의 관할(domination) 관계가 Weblicht 수형도의 화살표를 대신한다. 다시 말하여 임의의 두 교점이 의존기능 표지가 부착된 가지를 통해 연결되어 있을 경우에 상위에 위치한 교점이 핵어로, 하위에 위치한 교점이 의존어로 간주된다.

이제 핵어와 의존어간의 의존관계들 가운데 자주 나타나는 유형들을 정리해 보면 다음 표 2와 같다.

핵어	의존어	예
동사	명사 (주어, 목적어 등)	gebraucht -> Punkte (s4) (VVPP -OA-> PPER)
조동사	동사	hätte -> gebraucht (s4) (VAFIN -OC-> VVPP)
동사	부사	ließ -> wohl (s5) (VVFIN -MO-> ADV)
동사	(서술)형용사	sein -> froh (s2) (VAINF -PD-> ADJD)
동사	전치사 (보충어,부사어)	führte -> an (s43) (VVFIN -MO-> APPR)
동사	대용부사	bestand -> darauf (s1) (VVFIN -OP-> PROAV)
동사	관계부사	fahre -> wohin (s36) (VVFIN -MO-> PWAV)
명사	형용사	Oberflächen -> stumme (s11) (NN -NK-> ADJA)
(서술) 형용사	부사	froh -> wirklich (s2) (ADJD -MO-> ADJD)
형용사	명사	[pos=ADJA [pos=NN]] 3 Fälle ~~~> Fehler
명사	관사	Abenden -> den (s43) (NN -NK-> ART)
명사	전치사 (보충어)	Streit -> um (s204) (NN -OP-> APPR)

표 2. 핵어-의존어 유형

위에 정리된 의존관계 유형 가운데 ‘동사’가 핵어인 의존관계들이 많은데, 여기서 ‘동사’는 완전동사 정동사(VVFIN), 조동사 정동사(VAFIN), 과거분사(VVFIN), 완전동사 부정형(VVINF) 등을 일괄지칭한 상위범주이다. 보다 세밀하게 완전정동사(VVFIN)가 핵어가 되는 의존관계들을 일종의 행렬형식으로 정리하면 다음 표 3과 같다.

	Pos		KON	PPER	NN	NE	PIS	APPR
VVFIN	262	JU	13					
VAFIN	140	OA				1		
VVPP	4	SB		123	1	64	7	
NN	108	CP						1
APPR	38	--						
NE	9	MO			1			
APPRART	17	DA						
VMFIN	20	EP		4				
PROAV	2	PH		1				

표 3. 의존관계들의 행렬

위 행렬은 Stuttgart 대학에서 개발한 ICARUS 시스템을 이용하여 코퍼스 StammMI7-DpS의 의존관계를 검색한 결과의 일부를 보여준다. 이 행렬에서 왼편의 'VVFIN'을 선택하면 완전정동사가 핵어 기능을 하는 의존관계를 한눈에 볼 수 있다. 위의 표 3은 ‘VVFIN’이 중심이 된 의존관계 262 사례 중 일부만을 보인 것이다. 이 표를 통해 우리는 ‘VVFIN’이 핵어로서 의존어인 ‘접속사(KON)’와 의존관계를 가지며, 이들간의 관계는 ‘접속어(JU)’라는 것과 이 의존관계 유형의 출현빈도가 13이라는 사실을 확인할 수 있다. 마찬가지로 의존어인 ‘인칭대명사(PPER)’는 핵어 ‘VVFIN’와 123차례 의존관계를 가지는 데 이들간의 관계가 ‘주어(SB)’인 것을 알 수 있다.

이제 개별 의존기능이 코퍼스 StammMI7-DpS내에서 어떠한 통계적인 분포를 보이는 지를 살펴보기로 하자. Tündra에서 제공하는 검색엔진을 이용하여 의존기능의 출현빈도를 추출할 수 있는데, 검색을 위해 다음과 같은 검색식을 실행시켰다.

(5) #1:[pos=/.*/] >#2. #3:[pos=/.*/]

이 검색식은 TIGERSearch의 검색문법에 의거하여 생성한 것이며, 그 의미는 #1로 지시되는 교점과 #3으로 지시되는 교점이 있을 때, #1 교점이 #3 교점의 상위에 위치하며, 두 교점간의 의존관계를 나타내는 의존기능은 #2로 나타낸다는 것이다. 우리의 관심은 변수 #2의 값을 채우는 의존기능들의 출현빈도이다. 의존기능의 출현빈도를 백분율과 함께 제시하면 아래의 표 4와 같다.

의존기능	빈도	백분율(%)	의존기능	빈도	백분율(%)
NK	914	19.50	CM	14	0.30
SB	815	17.39	RC	12	0.26
MO	714	15.23	PNC	12	0.26
--	490	10.45	PG	12	0.26
OC	405	8.64	RE	8	0.17
OA	365	7.79	CC	7	0.15
PD	146	3.11	AMS	7	0.15
CJ	121	2.58	UC	6	0.13
CD	97	2.07	PH	6	0.13
SVP	91	1.94	APP	6	0.13
NG	85	1.81	PAR	3	0.06
DA	79	1.69	DM	3	0.06
CP	61	1.30	OG	2	0.04
JU	43	0.92	OA2	2	0.04
OP	34	0.73	AC	2	0.04
MNR	33	0.70	VO	1	0.02
EP	29	0.62	SBP	1	0.02
PM	26	0.55	RS	1	0.02
AG	19	0.41	AVC	1	0.02
CVC	14	0.30			

표 4. 의존기능의 분포

위 표를 보면, 의존기능 '명사핵(NK)'의 출현빈도가 가장 높고, '주어(SB)'

와 ‘수식어(MO)’가 그 뒤를 따른다는 것을 알 수 있다.[6]

한편, 코퍼스로부터 품사(pos)의 분포를 추출하기 위해 필요한 검색식은 다음 (6)과 같다.

(6) #1:[pos=/.*/]

이 검색식을 실행하여 얻은 결과는 아래의 표 5이다.

품사	빈도	백분율(%)	품사	빈도	백분율(%)
$.	702	11.54	VMFIN	52	0.85
NN	675	11.09	PIAT	34	0.56
PPER	664	10.91	PDS	27	0.44
VVFIN	612	10.05	PTKZU	25	0.41
ART	360	5.91	PWAV	21	0.34
ADV	313	5.14	VAPP	16	0.26
VAFIN	306	5.03	PWS	16	0.26
$,	302	4.96	PROAV	16	0.26
NE	260	4.27	KOKOM	16	0.26
APPR	247	4.06	CARD	14	0.23
$(	184	3.02	VAINF	11	0.18
ADJD	151	2.48	PRELS	9	0.15
KON	139	2.28	PTKA	8	0.13
VVPP	115	1.89	PDAT	7	0.11
VVINF	104	1.71	VVIZU	6	0.10
ADJA	100	1.64	PTKANT	6	0.10
APPRART	92	1.51	VVIMP	5	0.08
PTKVZ	91	1.49	KOUI	5	0.08
PIS	88	1.45	FM	5	0.08
PTKNEG	85	1.40	VMINF	3	0.05
PRF	70	1.15	XY	1	0.02

6) 네 번째 순위를 차지한 의존기능 ‘--’은 절 단위의 최상위 핵어가 콤마($,)나 마침표($.) 등 구두점과 갖는 의존관계를 나타낸다.

PPOSAT	68	1.12	APPO	1	0.02
KOUS	54	0.89			

표 5. 품사의 분포

이 표를 살펴보면, 마침표($.)에 이어 일반명사(NN), 인칭대명사(PPER) 및 완전동사 정동사(VVFIN) 순으로 출현빈도가 높은 것을 확인할 수 있다.

의존구조 코퍼스 StammMI7-DpS의 구축절차는 다음에 제시된 몇 가지 절차를 따른다.

(7)

- 원시코퍼스 준비
- Weblicht내에서 원시코퍼스를 불러들임
- 파싱 도구들의 선택 및 수집
- 수집된 도구들의 실행

다음 그림 3은 위의 네 단계 가운데 세 번째 단계를 마친 상태를 보여준다. 7)

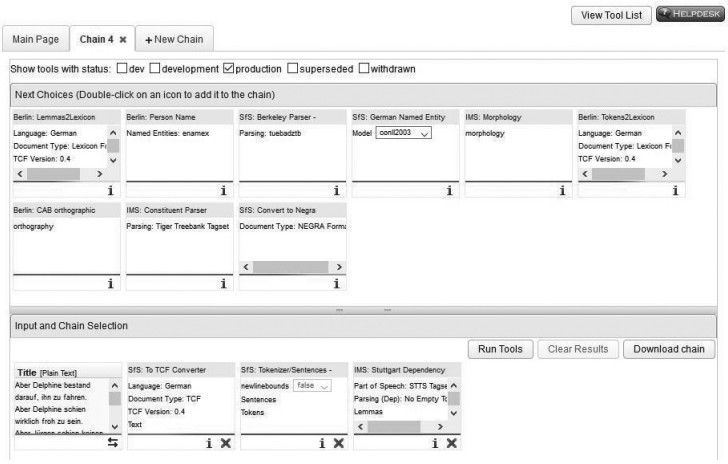

그림 3. 파싱 도구 모음

7) Weblicht의 주소는 다음과 같다: https://weblicht.sfs.uni-tuebingen.de/weblicht/

위 그림을 통해 우리는 의존구조 코퍼스의 생성을 위해 TCF 포맷으로 변환하기 위한 도구(To TCF Converter)와 문장 및 토큰분할기(Tockenizer/ Sentences) 및 의존구조 파서(Stuttgart Dependency)가 쓰임을 확인할 수 있다. 이처럼 파싱에 필요한 도구들이 모두 모아지면 우측 하단의 [Run Tools] 단추를 클릭함으로써 코퍼스를 구축하게 된다. 그 결과로 생성 되는 것은 아래 (8)에 제시되는 형태의 데이터구조인데, 이 구조는 TCF0.4 포맷을 취하고 있다.

(8) 코퍼스 StammMI7-DpS의 데이터 구조

```
<tc:tokens xmlns:tc="http://www.dspin.de/data/textcorpus">
    <tc:token ID="t1">Aber</tc:token>
    <tc:token ID="t2">Delphine</tc:token>
    <tc:token ID="t3">bestand</tc:token>
        .....
    <tc:token ID="t7">zu</tc:token>
    <tc:token ID="t8">fahren</tc:token>
    <tc:token ID="t9">.</tc:token>
        .....
<tc:sentences xmlns:tc="http://www.dspin.de/data/textcorpus">
    <tc:sentence tokenIDs="t1 t2 t3 t4 t5 t6 t7 t8 t9" ID="s1"></tc:sentence>
      .....
<namedEntities type="ENAMEX">
    <entity ID="e1" tokenIDs="t2" class="person"></entity>
      .....
<tc:lemmas xmlns:tc="http://www.dspin.de/data/textcorpus">
    <tc:lemma ID="l_0" tokenIDs="t1">aber</tc:lemma>
    <tc:lemma ID="l_1" tokenIDs="t2">Delphine</tc:lemma>
    <tc:lemma ID="l_2" tokenIDs="t3">bestehen</tc:lemma>
                .....
    <tc:lemma ID="l_6" tokenIDs="t7">zu</tc:lemma>
    <tc:lemma ID="l_7" tokenIDs="t8">fahren</tc:lemma>
```

```
    <tc:lemma ID="l_8" tokenIDs="t9">--</tc:lemma>
                    .....
  <tc:POStags xmlns:tc="http://www.dspin.de/data/textcorpus" tagset="stts">
    <tc:tag tokenIDs="t1">KON</tc:tag>
    <tc:tag tokenIDs="t2">NE</tc:tag>
    <tc:tag tokenIDs="t3">VVFIN</tc:tag>
                    .....
    <tc:tag tokenIDs="t7">PTKZU</tc:tag>
    <tc:tag tokenIDs="t8">VVINF</tc:tag>
    <tc:tag tokenIDs="t9">$.</tc:tag>
                    .....
   <tc:depparsing xmlns:tc="http://www.dspin.de/data/textcorpus" tagset="tiger"
multigovs="false" emptytoks="false">
    <tc:parse>
     <tc:dependency func="JU" govIDs="t3" depIDs="t1"></tc:dependency>
                    .....
     <tc:dependency func="--" govIDs="t8" depIDs="t9"></tc:dependency>
    </tc:parse>
                    .....
```

이 데이터구조는 널리 통용되는 XML 포맷을 따르는 것이 특징인데, 이 데이터구조 안에는 어휘형태(token)에 대한 정보, 레마(lemma)에 대한 정보, 품사(pos)에 대한 정보 및 의존관계(dependency)에 대한 정보들이 포함된다.

이 장을 마무리하면서 의존구조 코퍼스를 어떻게 활용할 수 있는 지에 대해 몇 가지 활용방안에 대해 논의하기로 한다. 일차적으로 코퍼스 StammMI7-DpS를 Tündra(Tübingen aNnotated Data Retrieval Application) 검색 시스템에 넣어 의존구조 수형도를 생성하고 적절한 검색식을 세워서 필요한 통계를 추출할 수 있다. 다음으로, 의존구조 코퍼스를 ICARUS(Interactive platform for Corpus Analysis and Research tools, University of Stuttgart) 분석도구에 넣어 의존구조 수형도를 생성하고 수정하거나 관심있는 언어현상에 대한

통계를 추출할 수 있다(Gärtner et. al. 2013 참조). 마지막으로, StammMI7-DpS를 구성구조 코퍼스 StammMI7-CsS로 확장할 수 있는데, 이 작업은 Weblicht에서 제공하는 해당 모듈을 이용하여 수행할 수 있다. Weblicht는 의존구조를 구성구조로 변환하는 모듈을 제공하고 있기 때문이다.

4. 웹기반 의존구조 파서의 평가

이 장에서는 웹기반 파서의 평가방법과 평가결과에 대해 논의한다. 네 단계로 구성된 평가방법에 대해 먼저 기술하자면, 코퍼스 StammMI7-DpS의 700문장에 대한 의존구조 수형도를 개별적으로 분석한 다음에, 오류가 존재하는 지 여부를 먼저 판단하고, 오류가 있는 경우 어떤 유형의 오류가 나타나는 지를 식별하고 더불어 오류는 어떤 구문에서 일어나는 지를 관찰하여 종합적으로 오류 분석데이터 파일로 정리한다. 수형도 700개를 모두 살펴본 결과 131개 수형도에서 분석오류가 발견되었는데, 이를 비율로 환산하면 18.7%에 이른다.

이제, 분석오류의 사례들을 몇 가지 들어보기로 한다.

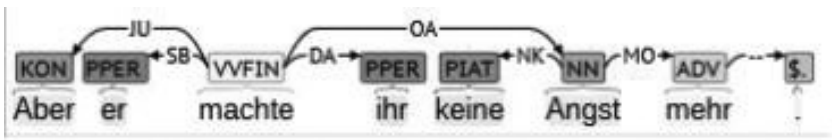

그림 4. 분석오류 사례 1

이 수형도는 문장 10의 의존구조를 보여주는데, 핵어의 선택에 있어 오류가 생겼고 연관구문은 '단순문'이다. 의존구조에 따르면 부사(ADV)의 핵어가 명사(NN)인 'Angst'로 분석되어 있는데, 올바른 분석이라면 정동사(VVFIN)인 'machte'를 핵어로 선택했어야 한다.

두 번째 사례를 보자.

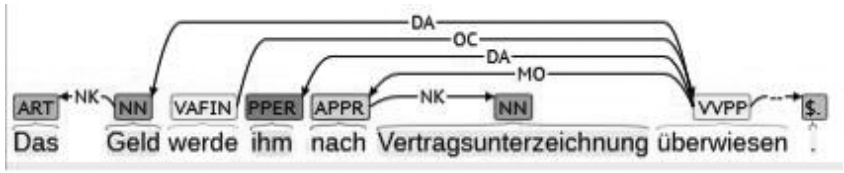

그림 5. 분석오류 사례 2

이 의존구조는 109번 문장을 분석한 결과인데, 앞서의 경우와 마찬가지로 구조상에 핵어선택 오류가 나타나 있고, '수동문'이 오류가 발생한 환경이다. 의존어 'Geld'의 핵어는 과거분사(VVPP)인 'überwiesen'가 아니라 정동사(VAFIN)인 'werde'로 분석되는 것이 타당하다.

다음 오류 사례는 문장 167의 분석과 관련된다.

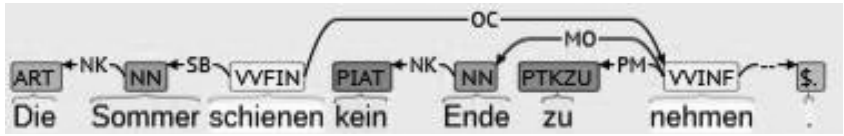

그림 6. 분석오류 사례 3

이 수형도에서는 부정동사(VVINF)인 'nehmen'과 일반명사(NN) 'Ende'간의 의존관계를 수식어(MO) 관계로 분석하고 있는데, 이들 간의 관계는 4격목적어(OA) 관계로 분석되어야 한다. 이 오류와 연관되어 있는 구문은 '부정사구문'이다.

다음 분석도 의존관계의 오류를 보여준다.

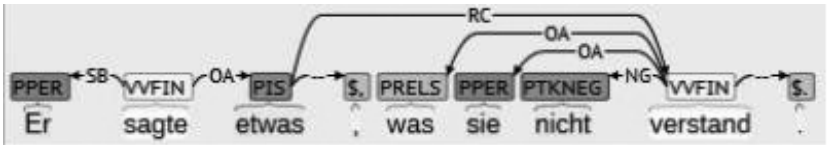

그림 7. 분석오류 사례 4

이 의존구조는 242번 문장을 분석한 것인데, 이 구조는 정동사 'verstand'와 인칭대명사 'sie'간에 4격목적어(OA)라는 의존관계가 성립하는 것으로

잘못 분석한 결과를 보여주고 있다.둘 간에 주어(SB)라는 의존관계가 성립하는 것으로 분석하는 것이 옳으며, 이런 오류가 일어난 구문적 환경은 '관계문'이다.

다음으로는 다시 핵어선택의 오류가 나타난 사례를 살펴보기로 한다.

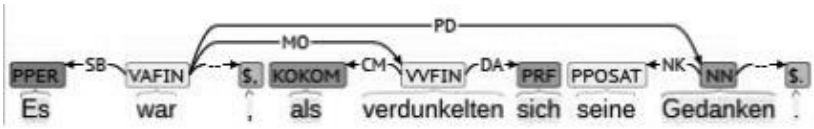

그림 8. 분석오류 사례 5

이 수형도는 288번 문장에 대한 분석을 담고 있다. 이 분석의 경우, 일반명사 'Gedanken'의 핵어를 종속절의 정동사 'verdunkelten'가 아닌 주절의 정동사 'war'로 파악한 것이 오류이다. 이 오류를 유발한 구문은 '생략문'이다.

다음은 다시 의존관계의 오류를 보여준다.

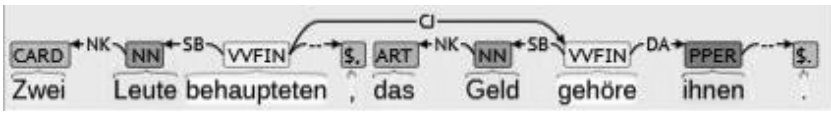

그림 9. 분석오류 사례 6

이 구조는 문장 611을 분석한 결과를 담고 있다. 이 분석에 따르면 주절의 정동사 'behaupteten'와 종속절의 정동사 'gehöre'간에 접속성분(CJ)이라는 의존관계가 성립하는데, 사실은 이와 다르다. 이들 간의 의존관계를 목적절(OC)로 파악하는 것이 올바른 분석이다. 이 오류가 발생한 환경은 종속접속사 dass의 '생략문'이다.

이상 여러 사례를 통해 살펴본 바와 같이 자동구문분석 결과 구축된 의존구조 코퍼스 StammMI7-DpS에서는 적지 않은 분석오류가 발견되는데, 이들 오류를 두 가지 유형으로 분류할 수 있다. 곧 의존관계의 오류와 핵어선택의 오류이다. 이 둘의 통계적 분포는 다음의 표 6과 같다.

오류유형	빈도	백분율(%)
의존관계 오류	118	90.08
핵어선택 오류	13	9.92

표 6. 오류유형의 분포

이 표에 나타난 분포를 보면 '의존관계 오류' 유형이 압도적으로 많이 나타나는 것으로 알 수 있다.

이제 오류가 발생하는 환경을 제공한 구문의 분포에 대해 살펴보자. 오류와 연관된 구문의 분포는 다음 표 7과 같다.

구문	출현빈도	MI7 백분율 (%)	오류 백분율(%)	누적 백분율(%)
도치구문	18	9.00	13.74	13.74
완료구문	14	11.71	10.69	24.43
생략문	13	3.57	9.92	34.35
종속문	12	11.00	9.16	43.51
복합명사구	11	6.29	8.40	51.91
외치구문	9	9.86	6.87	58.78
(일반) 단순문	9	12.43	6.87	65.65
불연속구문	8	5.57	6.11	71.76
간접의문문	7	3.43	5.34	77.10
부정사구문	7	9.43	5.34	82.44
관계문	7	1.71	5.34	87.79
등위접속구문	6	11.29	4.58	92.37
동격구문	5	1.43	3.82	96.18
(단순) 명령문	2	0.57	1.53	97.71
수동구문	2	1.14	1.53	99.24
비교구문	1	1.57	0.76	100.00

표 7. 구문의 분포

위의 표에 정리된 바, 오류와 연관된 개별 구문의 분포와 해당 구문이 코퍼스 StammMI7-DpS에서 출현하는 분포(백분율로 표시됨)와의 비교를 통해

우리는 다음과 같은 사실을 확인할 수 있다.

(9)
- 도치구문이 오류를 가장 많이 유발함[8)]
- 생략문과 동격구문이 오류를 많이 유발함
- 관계문과 간접의문문도 오류를 많이 유발함
- 부정사구문과 등위접속구문은 구문의 복잡성에도 불구하고 오류를 상대적으로 적게 유발함

위 명시적으로 드러난 (9)의 내용가운데 세 번째의 관찰을 좀 더 일반화시켜 아래 (10)과 같이 정리할 수 있다.

(10)
- 의존거리가 길수록 오류를 많이 유발함

어떤 문장내에 관계문이나 간접의문문 등이 포함되면, 핵어와 잠재적인 의존어사이의 거리가 늘어나게 되어 오류를 유발할 개연성이 증가한다.

5. 맺음말

본 연구를 통해 몇 가지 사실을 새롭게 발견할 수 있었다. 언어심리학자 George Miller의 연구결과를 받아들여 마법의 수인 7개 어휘로만 구성된 원시 코퍼스 StammMI7에 포함된 구문들의 유형을 살펴 본 결과, 그 유형의 스펙트럼이 매우 넓음을 확인했다. 이에 따라서 7개 어휘로 구성된 문장들을 이용하여 독일어 교육을 시도할 경우, 어떤 효과를 거둘 수 있는 지에 대한 후속연구가 필요하다고 생각된다. 이어 의존구조 코퍼스 StammMI7-DpS를 분석한 결과를 살펴보면 분석오류의 비율이 20%에 가까울 만큼 상대적으로

8) '도치구문'은 오류백분율이 13.74%인데, 이 구문이 코퍼스내에서 차지하는 백분율은 9%에 불과하다.

오류가 많음을 확인했다. 이러한 결과는 영어 파서의 경우, 오류율이 9% 미만으로 보고된 것과 비교하면 상대적으로 높은 오류율로 간주된다(Charniak/Johnson 2005). 또한, 독일어와 중국어를 평가대상으로 삼은 Berkeley 파서도 오류율이 10% 미만으로 보고된 결과와 비교해도 코퍼스 StammMI7-DpS의 오류율이 상대적으로 높다(Petrov/Klein 2007). 그런데, 앞서 비교한 영어파서나 독일어 파서 및 중국어 파서의 경우 모두 구성구조 파서의 평가결과이기 때문에 의존구조 파서의 결과와 직접 비교하는 것은 정당하지 않은 것으로 볼 수도 있다. 왜냐하면 의존구조의 경우, 구성구조와 달리 의존관계에 대한 정보가 추가되고 의존구조 파서의 오류는 대부분 의존관계의 오류를 나타내기 때문이다. 의존관계에 대한 정보가 포함된 Malt 파서의 경우 영어 자동 구문분석의 정확률이 86.3%에 불과하다는 Khmylko/Menzel(2010: 661)의 보고가 이 주장을 뒷받침한다.

참여수업의 관점에서 의존구조 코퍼스를 독문법 강의나 독어 통사론 강의에서 활용할 수 있는 여지가 있으나 교수자의 많은 노력이 필요할 것으로 보인다. 왜냐하면 자동 의존구조 분석 결과의 오류율이 20%에 이르기 때문에 교수자가 분석결과를 수업에 활용하기 위해서는 오류분석을 제거하는 수고를 들여야 하기 때문이다. 이 문제의 해결책의 하나로 제안할 수 있는 것은 바른 분석결과만을 담은 정선된 의존구조 코퍼스를 구축하기 위해 개인적 차원에서 바른 결과를 낸 문장들만을 반복적으로 수집하고, 집단차원에서 문장분석 결과를 공유할 필요가 있다는 것이다.

코퍼스 StammMI7을 의존구조 코퍼스 평가세트(Testsuite)로 활용할 수도 있는데, 이를 위해서는 구문의 분포에 대한 균형적인 설계가 필요하다.[9] 파서의 분석 오류율이 평가문장의 길이와 상관성을 갖는다는 보고가 있기 때문에 길이가 다른 문장들을 평가세트안에 포함시켜 오류율을 측정할 필요도 있다고 보여진다(McDonald/Nivre 2007).

9) 영-한 기계번역의 객관적 평가를 위해서 평가세트 및 평가방안을 이민행 외(1998)에서 제안한 바 있다. Maier et. al.(2014)에서는 독일어 불연속구문의 평가를 위한 평가방안에 대해 논의하고 있다.

참고문헌

이민행, 지광신, 정소우(1998): 기계번역 시스템 측정 장치 연구. 언어와 정보 제2권 2호: 185–220.

Charniak, E./ Johnson, M.(2005): Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. In: Proceedings of the 43rd Annual Meeting of the ACL, 173–180,

Cowan, N./ Morey, C./ Chen, Z.(2007): The legend of the magical number seven. In: Della Sala, S.(Ed.): Tall tales about the mind & brain: Separating fact from fiction. Oxford, U.K.: Oxford University Press, 45-59.

Foth, K.(2006): Eine umfassende Constraint-Dependenz-Grammatik des Deutschen. Tech. rep.. Universität Hamburg.

Gärtner, M./ Thiele, G./ Seeker, W./ Björkelund, A./ Kuhn, J.(2013): ICARUS – An Extensible Graphical Search Tool for Dependency Treebanks. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Sofia, Bulgaria, August 5—7.

Kakkonen, T.(2007): Framework and Resources for Natural Language Parser Evaluation. PhD Dissertation, Department of Computer Science and Statistics, University of Joensuu, Finland.

Kakkonen, T./ Sutinen, E.(2008): Coverage-based Evaluation of Parser Generalizability. In: Proceedings of the 3rd International Joint Conference on Natural Language Processing, Hyderabad, India.

Khmylko, L./ Menzel, W.(2010): Parsing as Classification. In: Locarek-Junge H./ Weihs C.(eds): Classification as a Tool for Research. Studies in Classification, Data Analysis, and Knowledge Organization. Springer. 657-664.

Kübler, S./ Maier, W./ Rehbein, I./ Versley, Y.(2008): How to compare treebanks. In: Proceedings of LREC 2008, Marrakech, Morocco.

Maier, W./ Kaeshammer, M./ Baumann, P./ Kübler, S.(2014): Discosuite–a parser test suite for German discontinuous structures. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),

2905–2912.
McDonald, R./ Nivre, J.(2007): Characterizing the Errors of Data-Driven Dependency Parsing Models. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 122-131.
Miller, G.(1956): The magical number seven, plus or minus two: Some limits on our capacity for processing information. In: Psychological Review Vol. 101(2): 343-352.
Petrov, S./ Klein, D.(2007): Parsing German with Latent Variable Grammars. In: Proceedings of the ACL-08: HLT Workshop on Parsing German (PaGe-08), 33–39.
Weblicht: http://weblicht.sfs.uni-tuebingen.de/

Zusammenfassung

Die Anwendung und Bewertung eines Web-basierten Parsers des Deutschen

Lee, Minhaeng (Yonsei Univ.)

Die vorliegende Abhandlung zielt sich darauf ab, folgende Fragen zu beantworten:

erstens, wie weit ist das Spektrum der deutschen Sätzen, die lediglich aus 7 Wörtern bestehen, in Hinblick auf die Konstruktionssorten?

zweitens, wie gut und korrekt sind die automatischen Analysen des web-basierten deutschen Dependenzparsers, der durch WebLicht bedient wird?

Die erste Frage bezieht sich auf die kognitiven Aspekte der grammatischen Konstrukitonen, denn die Lerner sollen Müller (1956) zufolge die Sätze mit 7 Wörtern ohne große Schwierigkeiten verarbeiten können. Die zweite Frage steht im engen

Zusammenhang mit der Performanzbewertung des Parsers. Wenn die Performanz des Parsers genug hoch wäre, dann könnte man ihn vielfach anwenden.

Was die zweite Aufgabe betrifft, wurde zunächst das Korpus ‘StammMI7’ erstellt, das sich ausschließlich aus 700 Sätzen ohne linguistische Informationen konstituiert, die jeweils 7 Wörter enthalten. Dann wurden 700 Sätzen durch den vom IMS der Uni. Stuttgart entwickelten Dependenzparser geparst und ihre Dependenzstrukturen wurden infolgedessen gesammelt. Anhand jeder Dependenzstruktur wurde geprüft, ob sie irgendeinen Strukturfehler aufwies und was für ein Fehler auftrat. Durch solchen Bewertungsverfahren konnten die Stärken swoie Schwächen des Parsers erhellt werden. Durch die Untersuchung wurden einige neue Befunde zum kognitiven Aspekt des Deustchen und zur Charakteristik festgestellt. Das Korpus StammMI7 vertritt verschiedenartige Konstruktionen, z.B. Passiv-Konstruktion, indirekte Frage-Konstruktution, Relativsatz-Konstruktion, Zu-Infinitiv-Konstruktion usw. In der Untersuchung wurde die Häüfigkeitsstatistik zu Konstruktionen erstellt. Durch die Fehleranlyse konnte festgestellt werden, dass der Stuttgarter Parser fast 20% Fehler generierte. Die Fehlerratio scheint relativ hoch im Vergleich mit Performanzen der anderen Parsers zu sein. Dazu wurde beobachtet und als Statistik erfasst, welche Konstruktionen häufiger Fehler verursachten. Dabei wurde festgestellt, dass die Konstruktion mit markierter Worstellung einen schlechten Umfeld bietete, und auch indirekte Frage-Konstruktion kein guter Kontext war und auch ellptische Konstruktion fehlerfördernde Konstruktion war.

Um den Dependenzparser in den Kurs der deutschen Syntax oder der Grammatik des Deutschen einzusetzen, muss man viel Mühe geben, durch das Bewertungsverfahren wiederholt nur richtige Dependenzstukturen aufzubereiten.

[Schlüsselwörter] Korpus, Parser, Dependenzstuktur, Performanzbewertung, Fehleranalyse, Fehlerratio

이민행 03722
서울특별시 서대문구 연세로 50

연세대학교 문과대학 독어독문학과
leemh@yonsei.ac.kr

논문투고일: 2017.04.25.
심사완료일: 2017.06.05.
게재확정일: 2017.06.20.