언어자료 검색을 위한 계층구조형 형태소 분석 프로그램

姜龍熙 東京大学大学院 総合文化研究科 言語情報科学
kang@phiz.c.u-tokyo.ac.jp

The Layered Structural Tagging Program for Seaching

Yong Hee Kang Language Information Scinece Tokyo University

1999년 제1회 형태소 분석기 및 품사태거 평가 워크숍 이후 표준안에 대한 새로운 대안이나 문제제기등을 제시한 논문은 전무하다. 본 연구에서는 평가대회 참가 이후 표준안을 수정한 새로운 유형의 형태소 분석 프로그램을 제작하여 그 실용성과 앞으로의 발전 가능성과 문제점을 밝혀, 계층구조형의 형태소분석 시스템을 채택하고 있는 일본의 JUMAN을 참조 새로운 유형의 형태소 분석형식을 제시한다. 본 연구는 일본방송협회 방송기술연구소(이하 NHK기술 연구소)의 의뢰에 인한 것이며 어절단위의 표준안과 다른 형태소 단위를 기본요소로 삼고 있으며 활용형을 갖고 있는 용언에 대해서는 활용형의 전개를 하고 있다. 어절단위로 탈피한 이유는 형태소 분석의 기본요소로써 어절단위 보다는 형태소 단위를 기준으로 삼는 것이 생산성이 높다고 생각된다. 어절정보와 문장정보는 XML(extensible makrup language)등의 별도의 정보를 주는 방법을 채택했다. 음절말음이 자음인지 모음인지의 음운정보에 따라 활용형을 차별했으며 표준안과 달리 명사의 종류와 개념을 세분화 했다. 아울러 조사와 어미등의 검색어와 함께 음절을 형성하고 있는 비검색어 대상은 배제하는 프로그램과 표준안의 어절방식으로 출력하는 3가지 프로그램을 작성했다. 본 연구에서는 계층구조의 형태소분석 프로그램의 가능성과 한국어의 특성을 고려한 출력항목등을 고찰하는 것을 목적으로 한다.

목차

1. 序論

한일기계번역시스템의 형태소 해석 프로그램의 중간버퍼를 표준안에 맞추어 명사추출 및 품사부착을 시도한 1999년의 연구에 이어 NHK기술연

구소의 의뢰에 준하여 새로운 형태의 검색용 형태소 분석 프로그램을 제작하였다.

위 프로그램은 3가지 형태의 출력방식이 가능하다.

1)1999년의 표준안을 기본으로 한 출력방식(약간의 변형은 있지만 기본적인 출력방식은 같다)

2)1999년의 명사추출과 같이 품사부착을 하지 않고 특정한 어휘만(체언과 용언의 어간)을 추출하는 출력방식

3)출력방식에 있어서 기본단위를 어절단위로 삼지않고 형태소 단위로 출력하는 방식

1.1. 시스템의 사양

· 기본모델 :solalis 2.8 intel		
· XMKNDIC	10363136	사전
· atfile2.dat	22	속성정의
· kan_kjp.tbl	9965	한자 변환용 테이블
· nkeiskis.tbl	9608	한글의구성을 파악하는 테이블
· kszhin.tbl	9036	품사접속 테이블
· distkfil.tbl	7431	표제어 최대길이 테이블
· HAN16.FNT	353440	한국어 폰트
· kidiom.tbl	32	관용구 변환용 테이블
· ktnhk[1]	225428	실행프로그램

(이하 본 프로그램을 KTNHK라 한다.)

품사의 분류: 28종의 품사와 125개의 등록형으로 사전이 이루어져 있다. 사전의 검색방법은 순차검색방법과 비순차(램덤)검색방식을 병용하고 있다. 사전의 기본 형식은 다음과 같다.

1표제어 2번역어 3품사 4품사 5의미정보 6의미정보
한국어 일본어 한국어 일본어

1999년의 연구에서는 입력문의 제한[2]을 어절단위와 문장단위로 나누었으나, 2001년의 연구에서는 문장의 길이를 8192바이트 이하로 설정했다. 아울러 문장에 있어서 표제어가 128개를 넘는 문에 대해서는 문장의 종료기호가 없더라도 문장을 끊어서 처리하도록 했다.

1.2. 품사

품사의 종류[3]및 분류방법은 1999년의 연구와 같은 방식의 사전을 사용하였기에 내용은 같다. 그러므로 본 연구에서는 사전의 표제어만 밝히며 그 밖의 의미법주(semantic marker)및 품사접속 테이블에 관한 정보는 1999년의 연구와 같다. 그러나 본 연구의 골격이 되는 사항이므로 중복되는 것을 감안하고 다시 밝힌다.

1. 명사류
   1.1 일반명사
   1.2 고유명사
   1.3 인명(姓)
   1.4 인명(名)
   1.5 지명
   1.6 법인명
   1.7 동사어간 가능명사(~하다型)
   1.8 대명사
2. 용언류
   2.1 동사
   2.2 형용사
3. 조사류
   3.1 조사
   3.2 終조사(어미)
   3.3 接續조사
4. 부사
   4.1 부사(일반부사)
   4.2 접속사 (접속부사)
5. 접사
   5.1 접두사
   5.1.1數詞선두어
   5.2 접미사
   5.2.1數詞用접미사
6. 관형사
7. 기호류
   7.1 숫자
   7.2 英字
   7.3 기호(4종)
8. 連語형
   8.1 조동사(어미+보조어간;10종)
   8.2 連用수식어
   8.3 連體수식어.

1) ktnhk는 (Kang-Tanaka-NHK)의 뜻이다.

2) · 입력어절의 최대길이는 512문자
   · 입력문장 최대길이도 512문자

3) 품사의 개념이라기 보다 기계번역시스템의 표제어 단위 혹은 등록형 단위라는 표현이 적적하다고 판단된다. 일본[히다치]에서 계발된 관계로 품사의 개념 및 용어는 일본어와의 관계로 한국에서 사용하는 문법용어와는 다르다.

9. 기타류
  9.1 인사말
  9.2 제목
10. 등록형
  10.1 동사8형(어간+어미형 포함[4])
  10.2 변칙동사5형
  10.3 조사 6형
  10.4 형용사 7형
  10.5 접속조사 7형
  10.6 조동사 8형

또한 등록단위 중 활용형과 동음이의어의 환경에 대비한 2중 등록형[5]을 포함하고 있다. 등록형은 파라다임[6](paradigme)으로 입력하며 동사, 동사, 형용사, 조동사(어미+보조용언)의 기본형과 그 활용형을 인덱스형으로 등록하는 방법이다. 그러나 용언의 명사형은 등록형이 아닌 단독형으로 처리하고 있다.

표제어 번역어 품사 품사 의미정보 의미정보
활용형(한) 어간(한) 인덱스
예)
기본 달리 はしる(번역어)
연용 달려 달리(어간) 인덱스정보
연체 달린 달리(어간) 인덱스정보
연용 달렸 달리(어간) 인덱스정보
연체 달리는달리(어간) 인덱스정보
연체 달릴 달리(어간) 인덱스정보
연용 달린 달리(어간) 인덱스정보
명령 달려라 달리(어간) 인덱스정보
*명사형 달림 ----------

1.3. 사전의 크기

1.3.1. 표제어[7]

기본적으로 표제어(한국어)와 번역어(일본어)의 숫자가 다르며, 하나의 표제어에 보통 여러가지 번역어를 가지고 있는 경우가 일반적이다.

1.3.2. 기본적인 분류와 크기

사전 전체	표제어	189,747
	번역어	257,973
명사류	표제어	106,436
	번역어	146,916
용언류	표제어	8,561
	번역어	12,506
기타류	표제어	1,396
	번역어	2,181

1.3.3. 세부적인 품사와 숫자

형식 : 품사	표제어	번역어
명사	83,474	118,740
한자어+하다의 명사	10,433	12,907
고유명사(地名등)	24,689	26,780
동사(어간)	6,005	7,969
동사(변칙)	44	79
부사	1,907	2,512
조사(조사+수식어포함)	225	285
어미	332	518
접사	759	981
보조용언[8]	545	818
연용수식連語[9]	173	194
연체수식連語[10]	73	81
접속조사	440	562

이하 생략

2. 本論

2.1. 어절단위의 프로그램

다음은 표준안에 의거한 품사태거의 예이다.

여러분 여러분/np
안녕하십니까? 안녕/nc+하/xsm+시/ef+ㅂ니까/ef+?/s
KBS KBS/f
9시 9/nn+시/nb
뉴스입니다. 뉴스/nc+이/co+ㅂ니다/ef+./s

그리고 다음의 예는 같은 문장을 새로운 어절형 방식으로 출력한 것이다.

<s>

---

4) 파라다임의 [달려]는 [달리+어]로 분리되나 번역상의 이점을 살려 단독형으로 처리한다.
5 예)나(대명사)+는(조사) : 나는(난다)
6) [합니다]형은 본 시스템에서는 번역어의 특성상 패러다임에 포함하지 않고 별도의 처리를 한다.
  합니다 → します
  예쁩니다 → きれいです
7) 1999년 연구이후 약간의 표제어를 수정보완했으나 사전의 내용이 크게 변하지 않았으므로 1999년 연구를 기본으로 삼는다.
8) 조동사(어미+용언)의 일부분
  예) 먹고 싶다= 먹+고 싶다 -> 싶다
9) 예)~에 관하여
10) 예)~에 관한

여러분 여러분/nc
안녕하십니까? 안녕/nc+하/xsm+시/ep+ㅂ니까/px+?/s
</s>
<s>
KBS 9시 KBS 9/f+시/xsn
뉴스입니다. 뉴스/f+이/co+ㅂ니다/ef+./s
</s>

본 연구는 1999년의 연구를 기본으로 하고 있기에 품사의 태거의 재현율과 정확율은 1999년 논문의 내용과 같다. 표1)은 기계번역스시템의 품사를 표준안의 품사정보로 변환시키는 변환표이다.

표1)

계층1	계층2	계층3	1999년 연구
1. s			7.1, 7.2, 7.3
2. f			NO (S)
3. n			
	3.1 nc		1.1-1.7
	3.2 nb		5.2.1
4. np			1.8
5. nn			5.1.1 (S)
6. pv			2.1, 10.1, 10.2
7. pa			2.2, 10.4
8. px			10.6
9. co			NO (2.2;S)
10. ma	10.1 mag		4.1
	10.2 maj		4.2
11. mm			6.
12. ii			NO(9.1;S)
13. x	13.1 xp		5.1.1(S)
	13.2 xs	13.2.1 xsn	1.1
		13.2.2 xsv	10.1
		13.2.3 xsm	10.4
14. j	14.1 jc		3.1, 10.3 S
	14.2 jx		3.1, 10.3 S
	14.3 ji		3.1, 10.3 S
	14.4 jm		3.1, 10.3 S
15. ep			NO 8.1,10.6
16. e	16.1 ef		3.2
	16.2 ec		3.3, 10.5
	16.3 et	16.3.1 etn	NO 1.1
		16.3.2 etm	10.1

NO = 해당하는 품사개념 없음.
S = 품사처리가 아닌 문자열처리임.

기본적인 태거방식은 같으나 문장의 처음(<s>)과 마지막(</s>)을 구분하기 위한 XML정보를 덧붙이고 있다. 어절형단위의 프로그램 결과는 검색이나 2차 처리등을 하기에 어렵기에 이상적인 출력방식이라고는 생각하기 어렵고, 다만 문장이 어절로 이루어져 있고 그 어절을 구성하는 품사단위를 부가하는데 의의가 있는 방식이라고 판단된다.

2.2. 형태소 단위프로그램

다음은 위의 예문을 형태소 단위로 출력한 예이다.

```
<s>
<p>
여러분    nc
</p>
<p>
안녕하    xsm    9    안녕하    안녕하여
안녕하였  안녕한  안녕하던  안녕할  안녕하
안녕합니  안녕합
시ㅂ니까  px
?         s
</p>
</s>
<s>
<p>
KBS 9     f
시        xsn
</p>
<p>
뉴스      f
이ㅂ니다  ef
.         s
</p>
</s>
```

위의 출력방식의 특징은 어절의 단위를 형태소 단위로 나누고 그 어간을 밝히고 용언에 그 활용형을 덧붙이는 방식이다.

예를 들어 [안녕하십니까]라는 문장에 대해 동사의 어간인 [안녕하]를 밝히고 [십니까]를 불리하여 어절정보는 XML(,)로 참고하는 방식이다. 이 방식에서 용언의 어간을 밝히는 것은 일반적으로 사전의 표제어는 용언의 기본형을 등재하고 있으므로 그 어간을 밝힐 필요가 있다고 판단했기 때문이다.

활용형에 대해서는 기본적으로 연용, 연체형을 기본으로 하고 시제형도 동시에 밝히며, 어절말음이 모음인가 자음인가에 따라 차별출력하는 방식을 택하고 있다. [안녕합니]와 같은 경우는 어간의 모음[하]의 종성[ㅂ]이 붙어 코드상 별도의 코드로 바뀌어 버리므로 활용형의 일부로 출력했다.

그러나 [먹습니다]와 같이 자음어간[먹]에 대해 [습니다/까]가 붙는 경우 별도의 처리를 하지 않았다. 명사형[안녕함]도 같은 이유로 출력을 하였다.

또한 [하다]와 같은 어간은 [하여/해]의 변이적인 활용형이 있는 경우에 대비해 축약형으로 출력할 수 있는 프로그램도 제작하였다.

2.3. 검색어 추출 프로그램

다음의 예는 검색어만을 추출하는 프로그램의 예이다. 1999년의 연구에서는 명사추출을 응용한 프로그램이다. 즉 품사태거를 하지 않고 체언과 용언의 어간 및 활용형만을 추출한다.

```
<s>
<p>
여러분
</p>
<p>
안녕하    9          안녕하    안녕하여  안녕하였
안녕한    안녕하던   안녕할    안녕하    안녕합니
안녕함
?
</p>
</s>
<s>
<p>
KBS 9
시
</p>
<p>
뉴스
</p>
</s>
```

검색어의 대상에서 제외되는 항목은 조사와 어미이다. 물론 형태소 단위로 추출하며 어절정보 및 문장의 처음과 마지막에는 XML정보를 덧붙인다. 1999년의 명사추출 프로그램도 형태소 단위로 추출하는 것이므로 본 연구의 취지와 일맥상통한다고 생각된다

2.4. 계층형 프로그램

표준안의 품사분류은 일부 품사를 카테고리로 묶어 계층형으로 분류하고 있다. 예를 들어 명사[n]를 일반명사[nc]와 의존명사[nb]로 분류하고 있고 부사[ma]도 일반부사[mag]도 접속부사[maj]로 나누고 있으며 접사[x]는 접두사[xp]와 접미사[xs]로 나누고 다시 접미사를 명사파생접미사[xsn]와 동사파생접미사[xsv] 그리고 형용사파생접미사[xsm]로 분류하고 있다. 그러나 품사의 분류를 계층분류하였다 하더라도 품사태거의 방식이 계층적 분류를 반영하는 출력방식이 아니고 소속된 품사만을 태거하는 방법은 품사의 전체적인 모습을 파악하려 할 때에는 재구성을 해야하는 문제가 있다고 본다. 예를 들어 다음과 같이 출력하는 방안도 고려해 볼 필요가 있다.

```
여러분    여러분/np
안녕하십니까?    안녕/[n[nc]]+하/[x[xs[xsm]]]+시
/ep+ㅂ니까/[e[ef]]+?/s
```

위의 예는 표준안의 출력방식에 따라 계층적 구조를 나타낸 방식이다. 이를 형태소 단위로 바꾸어 표현하면 다음과 같이 될 것이다.

```
여러분    [np]
안녕      [n      [nc]]
하        [x      [xs     [xsm]]]
시        [ep]
ㅂ니까    [e      [ef]]
?         [s]
```

다시 형태소 단위로 바꿔어 놓았을 경우 어절정보와 문장정보를 덧붙이면 다음과 같이 될 것이다.

```
<s>
<p>
여러분    [np]
</p>
<p>
안녕      [n      [nc]]
하        [x      [xs     [xsm]]]
시        [ep]
ㅂ니까    [e      [ef]]
?         [s]
</p>
</s>
```

위와 같이 계층적방방식으로 출력을 하게 되면 [발음정보]와 [음운현상] 그리고 [활용정보]등도 추가로 나타낼 수 있을 것이다.

발음정보와 활용정보는 다음의 일본의 형태소 분석 프로그램에서 구체적으로 다루기로 하겠다.

2.5. 일본어 형태소 프로그램

일본에서는 여러 단체에 의해 형태소 분석 프로그램이 10여년 전에 공개되어 그들 프로그램을 활용한 연구도 풍부하다.

본 연구에서는 1999년의 연구에서도 언급을 한 바와 같이 본 연구에서는 구체적으로 한국어 형태소 분석 프로그램과 다른점을 중심으로 기술하겠다.

다음은 일본의 형태소 분석 시스템의 대표하는 JUMAN[1]을 사용하여 형태소를 분석한 예이다.

```
[형태소][발음정보][기본형][품사정보][활용정보][활용형태정보]
これ (これ) これ 名詞形態指示詞
に (に) に 格助詞
對して (たいして) 對する 動詞 サ變動詞 タ系連用テ形
、 (、) 、 讀點
專門家 (せんもんか) 專門家 普通名詞
は (は) は 副助詞
どれほど (どれほど) どれほど 副詞
やさし (やさし) やさしい 形容詞 イ形容詞イ段 語幹
そうに (そうに) そうだ 形容詞性述語接 ナ形容詞 ダ列基本連用
見える (みえる) 見える 動詞 母音動詞 基本形
人 (ひと) 人 普通名詞
でも (でも) でも 副助詞
、 (、) 、 讀點
ひょっとする(ひょっとするひょっとす 副詞
恐ろしい (おそろしい) 恐ろしい 形容詞 イ形容詞イ段 基本形
ところ (ところ) ところ 副詞的名詞
が (が) が 格助詞
ある (ある) ある 動詞 子音動詞ラ行 基本形
かも (かも) かも 接續助詞
知れ (しれ) 知れる 動詞 母音動詞 未然形
ない (ない) ない 形容詞性述語接 イ形容詞アウオ 基本形
、 (、) 、 讀點
と (と) と 格助詞
思う (おもう) 思う 動詞 子音動詞ワ行 基本形
。 (。) 。 句點
EOS
```

일본어 형태소 분석 결과는 우선 형태소 단위로 출력하는 점이 다르다. 본래 우리말과 같은 띄어쓰기가 없으므로 어절정보를 필요로 하지 않고 문장의 종료시에는 [EOS]를 붙인다. 또한 입력문에 대해 계층적으로 정보를 부여하고 있고 품사정보는 생략형 혹은 축약형을 사용하지 않고 인지도가 높은 문법용어를 사용하고 있는 점이 크게 다르다. 출력방식의 구체적인 형식은 제1필드는 입력문을 형태소 단위로 나타내고 제2필드는 발음정보를 제3필드는 기본형이 있는 경우 기본형을 밝힌다. 제4필드는 품사정보를 제5필드는 활용형의 그룹을 제6필드는 활용형의 현태를 밝히고 있다. 품사별로 예를 통해 살펴보면 다음과 같다.

```
1)見える (みえる) 見える 動詞 母音動詞 基本形
[보이다] [발음] [원형] [품사] [품사그룹][활용형]
2)知れ (しれ) 知れる 動詞 母音動詞 未然形
[알 수있다] [발음] [원형] [품사] [품사그룹][활용형]
3)そうに(そうに)そうだ 形容詞性述語接 ナ形容詞 ダ列基本連用
[그렇다] [발음] [원형] [품사] [품사그룹] [활용형]
4)恐ろしい (おそろしい) 恐ろしい 形容詞 イ形容詞イ段 基本形
[무섭다] [발음] [원형] [품사] [품사그룹] [활용형]
5) 、 (、) 、 讀點
[기호[ 、 ] ] [발음] [원형] [품사]
6) 。 (。) 。 句點
[기호[ 。 ] ] [발음] [원형] [품사]
```

1)과 2)의 예는 동사의 경우로 3)과 4)는 형용사의 예이며 5)와 6)은 기호들의 예이다.

활용형을 갖고 있는 용언은 활용형의 형태와 그룹을 알 수 있도록 표시하고 있다.

위와 같은 경우는 우리말에서도 다음과 경우, [먹고 있다.] 의 [고]는 어미로 어간에만 붙는다. 그러나 [걸으면]의 [(으)면]은 불규칙용언의 기본형이 아닌 쪽에 붙는다. [걷다]에 대한 [*걷으면]과 같이 같은 품사라고 활용정보를 붙이는 것과 안 붙이는 것에는 큰 차이가 있다. 또한 기호,부호에 관한 태거도 쉼표인지, 마침표인지 또는 괄호인지를 개별적으로 나누는 방법을 채택하고 있다.

그리고 무엇보다 중요한 것은 출력결과를 눈으로 확인할 때 품사의 약칭을 사용하지 않고 일반 문법용어를 사용한다는 것은 알기쉽고 이해하는데 도움이 되며 비전문가가 결과를 볼 수 있다는 점을 들 수 있겠다. 한편 한국어 형태소 분석의 출력방식은 언어학적 지식이 없으면 결과를 눈으로 보고 바로 인식하기는 어렵다고 생각한다.

시각정보를 논리적으로 바뀌어야 하며 계층적 품사분류를 하였음에도 불구하고 출력방식에서 쉽게 확인하기 어렵다는 문제점이 있다. 위의 점을 비교할때 일본의 형태소 분석기의 출력방식은 사용자가 쉽게 사용할 수 있도록 궁리를 한 흔적을 충분히 엿볼수 있다.

2.6. 문제점

1)[일본어형태소분석시스템JUMAN version 3.5](*http://www-lab25.kuee.kuoto-u.ac.jp/nl-resource/juman.html*)

본 연구의 문제점에 관해서는 첫째 표준안의 품사분류의 문제점과 둘째 KTNHK의 문제점으로 나누어 논술하고저 한다.

2.6.1. 표준안의 문제점

표준안의 문제점으로는 먼저 품사분류의 세분화를 들 수 있다. 또한 한국어 형태소 분석 프로그램이면서 외국어라는 이질적인 품사개념을 도입하고 있다는 점을 들 수 있다.

2.6.1.1. 명사

KTNHK에서는 명사의 개념을 세분화하여 일반명사와 고유명사와 인명, 그리고 지명등을 분리하여 독립된 형태소 단위로 삼았다.

nd (人名 (姓))
ne (人名 (名))
nf (法人名)
ng (地名)
nh (他の固有名詞)

특히 뉴스를 중심으로 하는 코퍼스를 검색대상으로 삼는 NHK의 입장에서 보면 기사를 구성하는 중심요소로 인명과 지명등의 고유명사의 독립은 검색프로그램에 있어서 중요한 요소라고 할 수 있다. 그 때문에 다른 일반명사와 구별할 필요가 있기 때문이다. 아울러 소설과 수필등의 인명의 출현빈도가 적은 코파스를 대상으로 삼는다고 해도 고유명사를 명사와 불리하여 독립된 품사로 다루어야만 타당하다.

2.6.1.2. 외국어

표준안의 품사개념중 유일하게 문법범주에서 취급하지 않는 개념이다. 1999년 연구에서도 언급한 바 있으나 대상어휘가 외국어인지 일반명사인지를 판단하는 구체적 지시가 없으면 인지도에 따라 외국어가 되기도 하고 일반명사가 될 수도 있는 어휘가 많다.
KTNHK에서는 외국어라는 개념보다 외래어라는 개념으로 사용했다.
즉 고유어와 외래어라는 이분법을 사용했으며, 참고로 일본의 형태소 분석 프로그램에서는 외국어 외래어의 개념은 없고 미등록 어휘로 취급하고 있다. 즉 일본어 이외의 형태소는 품사항목을 두고 있지 않다.

2.6.2 본 프로그램의 문제점

2.6.2.1. 조동사의 처리

KTNHK의 문제점으로 1999년 연구에서도 밝힌 바와 같이 기계번역 시스템을 응용한 관계로 일반 품사개념과 조금 다른 어미와 보조용언의 결합형을 채택하고 있기 때문에 표제어에서는 독립된 하나의 형태소이지만 두어절의 결합된 형태소는 완전한 형태로 분리하지 못한 문제점이 남아 있다. 예를 들어 [~고 있다]에 대한 출력이 어절형 프로그램에서는 다음과 같이 나타나는 문제점이 있다.

<s>
재욱이 재욱/ne+이/jc
기다리고 있었다. 기다리/px+고/ec+있/px+었/ep+다/ef+./s
</s>

2.6.2.2. 불규칙 활용의 처리

검색용 프로그램을 위해 활용형을 추가하는 방식에 있어서 불규칙동사의 활용형의 처리는 가장 큰 문제이다. 불규칙활용의 활용은 어간이 바뀌는 [ㄹ변칙동사]나 [으변칙동사] 특정어미와 결합하는 [하다동사]와 [가다,오다]동사 그리고 어간과 어미가 함께 변하는 [ㅎ동사]등이 있다. KTNHK에서는 위의 변칙활용중에서 일부분의 활용에는 대응하고 있지만 완전한 형태로 활용형을 재현하지 못하는 예도 있다. 이 문제는 활용형 테이블을 작성하여 테이블을 사용한 처리를 모색하고 있는 실정이다.

2.7. 새로운 방식의 대안

위의 형식과 정보의 양을 비교해 볼 때 표준안의 형식과 내용에 좀더 한국어의 특성을 살리는 품사부착 정보를 부가할 필요가 있다. 예를 들면 발음, 교착, 굴절, 파생, 복합, 합성, 어원정보등을 들 수 있다.

예)그는 덧신을 신고 지붕으로 올라갔다.

표준안1)그/+는/ 덧/+신/+을/ 신/+고/ 지붕/+으로/ 올라가/+았/+다//

제안1)
그는/ 더씨늘/ 신꼬/ 지붕으로/ 올라갓따/.

제안2)
집-우ㅎ(어원정보) 덧신(합성정보) 덧/+신/ 더/+ㅅ/+신/ 올라/르변칙(불규칙동사정보)

제안3)
[어절단위][발음정보][어원정보,활용전보]
<s>

그는 (그는)

덧신을 (더씨늘)/(합성정보) 덧/+신/ 더/+ㅅ/+신/

신고 (신꼬)/[경음화][신(어간)+고(기본형+접속어미)]

지붕으로(지붕으로)[집+우ㅎ (어원정보)]

올라갔다(올라갓따)[경음화][올라/르변칙(불규칙동사)]

</s>

위의 제의를 포함하여, 과연 어떤 형식으로 품사부착을 하는 것이 보다 합리적이며, 품사정보에 의한 보다 용이한 언어검색이 가능할 것인지는 충분한 시간과 토의를 거쳐 검토해 나가야 될 것이다.

3. 結論

본 연구를 통해 현재의 표준안의 문제점과 계층구조형의 새로운 형태소 분석 방식을 모색해 보았다.

표준안의 방식은 한국어의 특성을 충분히 검토하여 만들어졌다고 보기 어렵고 경음화 격음화 구개음화등의 음운정보 및 활용정보에 대한 검토가 필요한 것은 주지의 사실이다. 또한 형태소 단위로 출력하는 새로운 타입의 형태소 분석 방법도 고려해볼 필요가 있다.

謝禮: 위의 연구가 있기까지 함께 프로그램의 제작을 도와준 日立의 田中光一씨를 비롯하여 프로그램의 의뢰자인 NHK의 江原박사님과 지도교수인 東京大學의 稲井교수님 그리고 많은 조언을 주시는 神田外國語大學의 管野교수님께 謝意를 표합니다.

참고문헌

[1]강용희 [한일기계번역에 있어서의 오역 및 고찰], 제8회 한글 및 한국어 정보처리 학술대회, pp351-366, 1996

[2] 姜龍熙 [韓日機械飜譯における助詞の誤譯の問題], 言語處理學會第3回年次大會發表論文集, pp.47-50, 1997(日本)

[3]강용희 [일본의 한일 기계 번역 시스템에 있어서의 오역과 그 언어환경], 제9회 한글 및 한국어 정보처리 학술대회, pp.303-310, 1997

[4]姜龍熙 [日本語と韓國語との言語の相異点と機械飜譯における問題点], AAMT Vol No22 April, pp.28-32,1998.(日本)

[5]松田純一,河野勝也 [構文ダイレクト方式による日韓機械飜譯システム],情報處理學會全國大會論文集, pp.139-140, 1993,(日本)

[6]강용희 외[한일기계번역시스템의 사전을 사용한 한국어 형태소분석시스템], 제11회 한글 및 한국어 정보처리 학술대회 및 제1회 형태소 분석기 및 춤사태거 평가 워크숍, pp.106-316, 1999