품사 부착 코퍼스 수정 방안에 대하여

김은혜 최기선
한국과학기술원 전문용어언어공학연구센터 KORTERM
{eunhye, kschoi}@world.kaist.ac.kr

On Correction Guideline of Tagged Corpus

Eun-Hye Kim, Key-Sun Choi

요 약

품사 부착 코퍼스를 구축하기 위해서는 일반적으로 형태소 분석, 자동 품사 태깅, 수동 또는 자동 오류 수정의 단계를 거친다. 이 글은 그 마지막 단계의 일환인 수동으로 오류를 수정하는 과정에서 요구되는 여러 가지 정보의 필요성과 문제점에 대해 기술하고자 한다. 조사와 어미의 처리 문제, 접두사/접미사 처리 문제, 다품사 문제 등은 정밀도 높은 코퍼스를 구축하는 데 중요한 열쇠가 되기 때문이다. 자연 언어 자료인 코퍼스에 일관성 있는 품사 정보가 부착된다면 정보 검색이나 사전 구축 등 언어 정보 처리 연구에 중요한 자료로 사용될 수 있을 것이다.

1. 연구의 목적 및 개요[1)]

자연언어처리나 정보 검색, 국어 관련 정보 처리의 중요한 기초 자료인 품사 태그 부착 코퍼스의 품사 태깅 오류를 수동 작업을 통해 교정하고, 분석의 일관성을 확보하여 국어 정보 처리를 위한 언어 데이터베이스 구축에 기여하는데 이 연구의 목적이 있다.

현재 진행중인 코퍼스에 대한 태깅 수정 작업은 '96년에 이미 만들어진 54개 「국어 품사 태그」를 기본으로 하였다. 이 품사 태그는 '①한국어 자료를 대상으로 함(고유어, 한자어, 외래어 및 각종 기호를 포함) ②학교 문법에 기반을 둔 품사 분류(실용성과 생산성은 감안) ③품사론을 기반으로 하고 형태론, 문장론 및 기호까지 포함하여 의미, 형태, 기능을 감안해서 설정된 것이다. 국어 품사 태그 집합은 대분류 9부문과 중분류 24부문 및 소분류 54부문으로 되어 있다. 지금의 수정 작업은 이 54개 「국어 품사 태그」를 그대로 따르며 다만, 기존의 매뉴얼은 분석의 일관성이나 생산성에 바탕을 두어 수정, 추가해 나간다.

2. 연구의 내용

본 연구의 주요 내용은 태그 부착 코퍼스에 나타난 오류의 유형을 분석하여 결정한 정답과 54개 품사 태그 집합을 토대로 정제된 코퍼스를 구축하는 것이다.

먼저 이중 태깅 코퍼스 분석을 통한 정답 작성에 관한 내용을 설명하고, 54개 품사 태그 집합, 그리고 태깅의 전반적인 내용을 설명하고자 한다.

2.1 이중 태깅 코퍼스 분석 및 정답 작성

이전에 가공된 태그 부착 코퍼스를 살펴본 결과 아래와 같이 이중으로 태깅된 목록이 6400어절 정도였다. 이중에는 교정자의 오분석으로 인한 오류도 있지만, 문맥에 따라 달라지는 태그로 인한 경우도 있다.

먼저 오분석에 의해 이중으로 태깅된 경우에는 그중 현재의 지침에 맞는 정답을 선정하여 정답 목록을 만든다. 예를 들면 다음과 같다.

오분석목록　　오분석 예

1) 이 연구는 과학기술부 특정연구과제 "핵심소프트웨어연구개발"중 "국어정보처리"의 세부과제 "대용량 국어정보 심층처리 및 품질관리 기술개발"에 의한 결과입니다.

- 가까이 가까이/mag
  가까이/nbn
  가까이/ncn
  가깝/paa+이/xsa 정답⇒가깝/paa+이/xsa
- 거론되고 거론/ncn+되/xsv+고/ecx
  거론/ncpa+되/xsv+고/ecx
  정답⇒거론/ncpa+되/xsv+고/ecx
  거론/ncpa+되/xsv+고/ecs
  거론/ncpa+되/xsv+고/ecc
- 거머쥔 거머/ncn+쥐/pvg+ㄴ/etm
  거머쥐/pvg+ㄴ/etm
  정답⇒거머쥐/pvg+ㄴ/etm
- 거미줄처럼 거미줄/ncn+처럼/jca
  거미줄/ncn+처럼/jxc
  정답⇒거미줄/ncn+처럼/jca
- 거부하고 거부/ncpa+하/xsv+고/ecc
  거부/ncpa+하/xsv+고/ecx
  거부/ncna+하/xsv+고/ecx
  거부/ncna+하/xsv+고/ecc
  정답⇒거부/ncpa+하/xsv+고/ecc
  거부/ncpa+하/xsv+고/ecx
  거부/ncpa+하/xsv+고/ecs
- 건설교통부 nq[건설/ncn+교통부/ncn]nq
  건설교통부/nq
  정답⇒건설교통부/nq
- 검토중이다. 검토/ncn+중/xsn+이/jp+다/ef+./sf
  검토/ncn+중/nbn+이/jp+다/ef+./sf
  정답⇒검토/ncn+중/nbn+이/jp+다/ef+./sf
- 기잡니다 기자/ncn+이/jp+ㅂ니다/ef
  기자/ncn+ㅂ니다/ef
  정답⇒기자/ncn+ㅂ니다/ef+./sf
- 법률안을 법률안/ncn+을/jco
  법률/ncn+안/ncn+을/jco
  정답⇒법률/ncn+안/ncn+을/jco

다음은 문맥에 따라 달라지는 태그로 인해 이중으로 태깅이 되는 예이다.

틀	틀/pvg+ㄹ/etm 트/pvg+ㄹ/etm 틀/ncn	팔	팔/pvg+ㄹ/etm 파/pvg+ㄹ/etm 팔/ncn
팀과	팀/ncn+과/jcj 팀/ncn+과/jct	적은	적/paa+은/etm 적/pvg+은/etm 적/nbn+은/jxc 적/ncn+은/jxc
파리	파리/nq 파리/ncn	주겠다는	주/pvg+겠/ep+다/ef+는/etm 주/px+겠/ep+다/ef+는/etm

2.2 품사 태그 집합

여기에서는 전체 54부문으로 되어 있는 품사 태그 집합을 각각의 부문별로 간단히 소개하고자 한다.

2.2.1 기호(s)

- sp(쉼표)-반점(,) 가운뎃점(·) 쌍점(:) 빗금(/)
- sf(마침표)-온점(.) 물음표(?) 느낌표(!)
- sl(여는 따옴표 및 묶음표)-여는 따옴표(") 여는 작은따옴표(') 여는 소괄호(( ) 여는 중괄호({) 여는 대괄호([ )
- sr(닫는 따옴표 및 묶음표)-닫은 따옴표(") 닫은 작은따옴표(') 닫은 소괄호( )) 닫은 중괄호(}) 닫은 대괄호( ])
- sd(이음표)-줄표(—) 붙임표(-) 물결표(~)
- se(줄임표)-줄임표(……)숨김표(××, ○○) 빠짐표(□)
- su(단위 기호)-m, cm, mm, ft, yd, kg, g, ℓ, dℓ %, ₩, ¥, $
- sy(기타 기호)-드러냄표(˙ , ˚) 밑줄(_____,_____) +, -, ×, ÷, ∮, ※, ↑, →, ↕

2.2.2 외국어(f)

- 국어 문장에서 외국 글자로 표기된 경우 외국어(f)로 태깅한다
  예) Esperanto/f 는 국제적인 인공어이다.
  Zigzag/f 로 달리며 thrill/f 을 느끼던 차가 전복되었다.
- 외국어를 음차 표기한 경우-고유성 여부를 가려 일반명사(ncn) 또는 고유명사(nq)로 태깅한다.
  예) 펜/ncn과 잉크/ncn는 학습의 필수품이다.
  윌리암/nq 선교사는...로 태깅한다.:
- 한자로 표기된 경우-국어와 동등하게 취급한다.
  예) 學校/ncn와 刑/ncn을 소중히 여겨야 한다.
  無關心/ncn도 罪/ncn가 된다.

2.2.2 체언(n)

- 서술성 명사(ncp)-ncpa(동작성 명사): '하, 되,시키'의 동사 파생 접미사와 결합
  -ncps(상태성 명사): '하, 스럽, 답,롭'의 형용사 파생 접미사와 결합
- 비서술성 명사(ncn)-ncn(비서술성 명사)
- 고유명사(nq)-nq(고유명사)
- 의존명사(nb)-nbu(단위성 의존 명사)
  -nbn(비단위성 의존 명사)
- 대명사(np)-npp(인칭 대명사)
  -npd(지시 대명사)
- 수사(nn)-nnc(양수사)
  -nno(서수사)

2.2.4 용언(p)

- 동사(pv)-pvg(일반 동사)
  -pvd(지시 동사)

• 형용사(pa)-paa(성상 형용사)
-pad(지시 형용사)
• 보조 용언(px)-px(보조 용언):말다, 못하다, 하다, 지다, 가다, 나가다, 오다, 있다, 계시다, 버리다, 주다, 드리다, 보다, 대다, 두다, 놓다, 내다, 나다, 싶다, 않다

2.2.5 수식언(m)
• 관형사(mm)-성상 관형사(mma)
-지시 관형사(mmd)
• 부사(ma)-일반 부사(mag)
-지시 부사(mad)
-접속 부사(maj)

2.2.6 독립언(i)
• 감탄사(ii)-감탄사(ii):그래, 그럼, 아, 아이구, 암, 어쩜,

2.2.7 관계언(j)
• 격조사(jc)-주격 조사(jcs): 이/가, 께서
-목적격 조사(jco): 을/를, ㄹ
-보격 조사(jcc): 이/가
-관형격 조사(jcm): 의
-호격 조사(jcv): 아, (이)야,(이)여,(이)시여
-부사격 조사(jca): 에, 으로, 보다, 처럼..
-접속격 조사(jcj): 와/과, 다, (이)랑, 며
-공동격 조사(jct): 랑, 와/과, 하고
-인용격 조사(jcr): 라고, 하고, 고
• 서술격조사(jp)-술격 조사(jp):'-이'
• 보조사(jx)-통용 보조사(jxc)
-종결 보조사(jxf): 요, 마는, 그려, 그래

2.2.8 어미(e)
• 종결 어미(ef)-종결 어미(ef)
• 선어말 어미(ep)-선어말 어미(ep)
• 연결어미(ec)-대등적 연결 어미(ecc)
-종속적 연결 어미(ecs)
-보조적 연결 어미(ecx)
• 전성 어미(et)-명사형 어미(etn)
-관형사형 어미(etm)

2.2.9 접사(x)
• 접두사(xp):생산성이 많다고 판단되는 것만을 선별하여 인정함.
예) 대/xp일본, 가/xp건물, 고/xp성능, 과/xp보호, 대/xp가족, 반/xp혁명, 비/xp공개, 신/xp제품, 총/xp인구, 친/xp정부, 피/xp지배,재/xp2과...
• 접미사(xs)-명사 파생 접미사(xsn)
-동사 파생 접미사(xsv): 하, 되, 시키
-형용사 파생 접미사(xsm): 하, 답, 스럽, 롭, ㅁ직하, ㅁ직스럽
-부사 파생 접미사(xsa): 이, 히

2.3 태깅 방식에 관한 전반적인 내용
태깅에 관한 전반적인 내용은 복합 형태의 분석, 서술성 명사의 태깅, 보조 용언의 태깅, 중의성을 띤 형태의 분석, 여러 가지 분석이 가능한 유형으로 나누어 설명하고자 한다.

2.3.1 태깅 방식
태깅은 띄어쓰기를 중심으로 어절별로 하되 하나의 어절로 구성되어 있을지라도 형태소별로 태깅하는 것을 원칙으로 한다. 또한 사용된 언어의 모습 그대로를 반영하면서 태그를 부착하는 것을 원칙으로 한다.

● 서술격 조사의 생략 : 서술격 조사 '-이'는 체언의 어말음이 모음인 경우에 생략된다. 이런 경우에 서술격 조사를 복원하지 않는다.
예: 이것은 소다 → 소/ncn+다/ef
∵ '소이다'에서 '이'가 생략된 형태이지만 '이'를 복원하지 않고 그대로 태깅한다.

● 의존 명사 '것'의 준말 '거' : 의존 명사 '것'은 뒤에 모음이 연결될 때 '거'의 형태로 나타나는 경우가 많다. 이 경우에도 '거'를 '것'으로 복원하지 않는다.
예: 먹을 게 많이 있다 → 거/nbn+이/jcs
그게 무엇이냐 → 그거/npd+이/jcs

● 지시대명사 '뭐', 인칭대명사 '누' 역시 원형을 복원하지 않고 그대로 태깅한다.
예: 뭐가 제일 좋을까? → 뭐/npd+가/jcs
누가 왔느냐? → 누/npp+가/jcs

● 불규칙 동사의 어간 : 불규칙 동사는 활용을 하거나 파생을 할 때 뒤의 어미나 접사의 어두음이 모음인지 자음인지에 따라 어간이 바뀌는데 그 원형을 복원하여 태깅한다.
예: 고기를 구워 먹었다. → 굽/pvg+어/ecs
끈과 끈을 이었다. → 잇/pvg+었/ep+다/ef+./sf

● 용언의 어간의 끝 모음과 어미의 첫 번째 모음이 동일하여 나타나는 축약 현상은 어미를 복원하여 분석한다.
예: 학교에 가 공부를 해라 → 가/pvg+아/ecs
감자에 싹이 나 못 먹게 되었다.→나/pvg+아/ecs

● 어미 '-지' 뒤에 '않-'이 어울려 '-잖-'이 되는 경우와 '-하지' 뒤에 '않-'이 어울려 '-찮-'이 되는 경우에는 준대로 적는 것이 표준안으로 되어 있다. 이들은 이미 줄어진 형태가 굳어져 하나의 단위로 다루어지고 있어 그 원형을 복원하지 않고 줄어진 형태 그대로 태깅한다.
예: 이번에는 적잖은 손해를 보았다.→적잖/paa+은/etm
∵ '적잖은'은 '적지 않은'이 줄어서 된 말이지만 '적잖

-'을 하나의 단위로 취급한다.

문제가 만만찮다. → 만만찮/paa+다/ef+./sf
인물이 변변찮다. → 변변찮/paa+다/ef+./sf

● 서술성 명사 가운데 접미사 '-하-'가 결합하여 사용될 때 접미사 '-하-'의 'ㅏ'가 탈락하여 'ㅎ'이 다음 음절의 첫소리와 어울려 거센 소리로 되는 경우가 있다. 이 경우에도 원형을 복원하지 않는다.

예: 일을 간편캐 하여라. → 간편/ncps+캐/ecx
대책을 마련키 위해 → 마련/ncpa+키/etn

서술성 명사일지라도 단독으로 사용되는 경우에는 비서술성 명사(ncn)로 태깅하기로 하였지만 '간편캐, 마련키'의 경우에는 '하'의 존재가 남아 있는 것으로 '간편, 마련'을 서술성 명사 중 상태성 명사로 태깅한다.

2.3.2 복합 형태의 분석

복합 형태는 각각을 분석함을 원칙으로 한다. 복합 형태로는 복합명사, 고유명사, 복합동사, 복합조사, 복합어미, 조사와 어미의 결합 등이 있다.

● 복합명사의 태깅 : 융합복합어와 종속복합어는 각각을 분석할 경우 그 의미를 상실하므로 그대로 하나의 단위로 다루고 병렬복합어의 경우에는 각각의 의미를 가지고 있으므로 분석하여 태깅한다.

예: 눈물 → 눈물/ncn
물거품 → 물거품/ncn
학교생활 → 학교/ncn+생활/ncn
경제정책 → 경제/ncn+정책/ncn

* 사이 'ㅅ'이 들어간 합성어는 분석하지 않고 하나의 단위로 취급하여 태깅한다.

예: 어젯밤 → 어젯밤/ncn
고깃국 → 고깃국/ncn

● 고유명사의 태깅 : 고유명사는 단일어로 되어 있는 경우도 있지만 다어절이 하나의 고유명사를 이루고 있는 경우가 많다. 다어절이 하나의 단위를 이루고 있을 때 이들이 복합형태로 붙여써 있는 경우에는 이들을 하나의 단위로 취급하여 고유명사로 태깅하면 되지만 다어절이 띄어써 있는 경우에는 이 어절 뒤에만 태그를 붙인다면 그 전체에 대한 정보를 얻을 수가 없을 것이다. 그래서 다어절로 구성된 고유명사의 경우에는 하나로 묶어 줄 필요가 있다.

예: 한국과학기술원 → 한국과학기술원/nq
추락하는 → **nq[**추락/ncpa+하 /xsv+는/etm
것은 것/nbn+은/etm
날개가 날개/ncn+가/jcs
있다 있/paa+다/ef**]nq**

● 복합조사의 경우 : 복합조사는 각각을 분석하여태깅한다.

예: 대전에서부터 → 서울/nq+에서/jca+부터/jxc
학교엘 갔다 → 학교/ncn+에/jca+ㄹ/jco

● 복합어미의 경우 : 복합어미 역시 각각을 분석하는 것을 원칙으로 한다.

예: 잡았겠군요 → 잡/pvg+았/ep+겠/ep+군/ef+요/jxf
불리었다는군→불리/pvg+었/ep+다/ef+는군/ef

● 종결어미와 관형형 어미의 결합된 형태 : 종결어미와 관형형 어미의 결합된 형태는 국어현상 가운데 많이 사용되는 예이다.

예. 책을 보았다는 사실-책/ncn+을/jco 보/pvg+았/ep+다/ef+는/etm 사실/ncn

과 같이 종결어미와 관형형 어미를 분석하여 각각 태깅한다.

● 조사와 어미의 결합된 형태 : 조사와 어미의 경우에는 각각을 분석한다.

예' 밥을 먹어야만 한다 → 밥/ncn+을/jco 먹/pvg+어야/ecs-만/jxc 하/px+ㄴ다/ef

어떻게 돈을 쓰느냐가 중요하다. → 어떻/pad+게/ecs 돈/ncn+을/jco 쓰/pvg+느냐/ef+가/jcs 중요/ncps+하/xsm+다/ef+./sf

● 복합동사의 태깅 : 복합동사는 모두 분석하여 태깅하는 것을 원칙으로 한다.

예: 꺼내입은→ 꺼내/pvg+어/ecs+입/pvg+은/etm
내려앉아 → 내리/pvg+어/ecs+앉/pvg+아/ecs
흘러나오고 → 흐르/pvg+어/ecs+나오/pvg+고/ecc

● 주격이나 목적격 조사가 생략되고 '명사+동사/형용사'형으로 구성된 유형도 역시 각각을 분석하여 태깅한다.

예: 대책없이 → 대책/ncn+없/paa+이/xsa
면제받은 → 면제/ncn+받/pvg+은/etm
밑빠진 → 밑/ncn+빠지/pvg+ㄴ/etm
인기있는 → 인기/ncn+있/paa+는/etm
앞세워 → 앞/ncn+세우/pvg+어/ecs

2.3.3 서술성 명사의 태깅

서술성 명사는 뒤에 '하다, 되다, 시키다'가 결합하여 서술어를 만들 수 있는 명사를 말한다. 서술성 명사의 구성은 'X하-'의 구성이다. 서술성 명사의 태깅은 다음과 같다.

● X가 2음절 이상인 경우 : 'X/ncpa+하/xsv+다/ef' 또는 'X/ncps+하/xsm+다/ef'로 분석할 수 있다.

예: 공부하다 → 공부/ncpa+하/xsv+다/ef
깨끗하다 → 깨끗/ncps+하/xsm+다/ef

● X가 단음절일 경우 : 'X'가 독립적으로 쓰일 수 있는 명사인 경우에는 분석하고 그렇지 않은 경우에는 'X하/pvg+다/ef' 또는 'X하/paa+다/ef'로 분석한다.

예: 말하다 → 말/ncpa+하/xsv+다/ef
가하다 → 가하/pvg+다/ef
착하다 → 착하/paa+다/ef
약하다 → 약하/paa+다/ef

● 그러나 서술성 명사가 '-하-, -되, -시키-'의 결합 없이 홀로 쓰일 때는 비서술성 명사로 태깅한다.

예: 공부를 열심히 하였다. - 공부/ncn+를/jco

2.3.4 보조 용언의 태깅

보조 용언은 본용언에 기대어 쓰이는 것을 말한다. 그러나 본용언과 보조 용언이 띄어 써 있는 경우에는 한 어절씩 태깅을 하면 되겠지만 본용언과 보조용언이 붙여 써 있는 경우가 있다. 이런 경우에도 본용언과 보조용언을 분석하여 태깅한다.

● 보조 용언이 본용언과 결합하여 붙여 써 있는 경우 본용언과 보조 용언의 의미를 나누어 생각할 수 없는 경우에는 분석하지 않고 태깅한다.

예: 떨어지다 → 떨어지/pvg+다/ef
사라지다 → 사라지/pvg+다/ef

● 본용언이 독립적으로 사용될 수 있으면 비록 붙여 써 있더라도 본용언과 보조 용언을 분석하여 태깅한다.

예: 이루어지다 → 이루/pvg+어/ecx+지/px+다/ef

● 본용언과 보조 용언을 띄어 쓴 경우 : 어절 단위로 태깅하게 되어 있어 문제가 없다.

예: 먹고 있다 → 먹/pvg+고/ecx 있/px+다/ef

● 문장 뒤에 쓰이는 보조 용언

예: 책을 쓸까 한다 → 책/ncn+을/jco 쓰/pvg+ㄹ까/ef 하/px+ㄴ다/ef
지갑을 찾았나 보다 → 지갑/ncn+을/jco 찾/pvg+았/ep+나/ef 보/px+다/ef

● 보조 용언으로 분석했던 '되다'의 처리

1차 가공시에는 '되다'를 보조 용언으로 분석했으나 현 작업에서는 '되다'를 보조 용언으로 처리하지 않고 본용언으로 분석한다.

학자에 따라 '어머니는 아이에게 밥을 먹게 하였다.'의 '-게 하다'를 보조 용언으로 보는 것 처럼 '그 사람이 돌아오게 되었다'의 '-게 되다'의 구성을 피동을 만드는 보조 용언으로 보기도 한다. 그러나 전통 문법과 학교 문법에서는 '-게 되다'를 보조 용언으로 보지 않는다. 그 이유는 '되다'라는 동사는 '산이 바다로 되었다, 공사가 잘(기막히게) 되었다.'에서처럼 기본적으로 부사어를 요구하는 동사로, '그 사람이 돌아오게 되었다'의 '돌아오게'도 '되다' 동사가 요구하는 부사어로 보기 때문이다. 이것을 본동사로 보고 뒤에 오는 '되다'를 보조 동사로 할 이유가 없다는 것이다. 그러므로 이번 수정 작업에서는 '되다'를 보조 용언으로 분석하지 않고 본용언으로 태깅하기로 한다.

2.3.5 중의성을 띤 형태의 분석

하나의 형태가 둘 이상의 기능을 가지고 있어 이것에 태그를 부착하려면 형태 그 자체보다는 문장 내에서의 기능을 고려하여야 한다. 비록 형태가 같다 하더라도 문장 내에서의 쓰임에 따라 다르게 태깅할 수 있기 때문이다.

● 명사와 부사의 구분

명사와 부사 가운데 동일한 형태이면서도 문장 내에서의 기능은 다른 것이 많다. 이러한 경우 문장 내에서의 기능을 고려하지 않을 수가 없다.

예: 그 일은 모두에게 책임이 있다 → 모두/ncn+에게/jca
그릇에 담긴 소금을 모두 쏟았다 → 모두/mag

● 의존 명사와 조사

의존명사와 조사의 구분은 관형형 어미 뒤에 사용되었느냐 체언 뒤에 사용되었느냐로 구분할 수 있다.

예: 먹을 만큼만 가져가거라. → 만큼/nbn
너만큼만 공부를 잘 했으면 좋겠다. → 너/npp+만큼/jxc+만/jxc
시키는 대로 열심히 한다. → 대로/nbn
나는 나대로 계획이 있다.→ 나/npp+대로/jxc

● 대명사와 관형사

예: 그는 학교 선생님이다. → 그/npp
그 책의 표지는 붉다. → 그/mmd

● 대명사와 부사

대명사와 부사의 구분은 격조사의 결합 여부로 판단할 수 있다. 격조사가 결합할 수 있으면 대명사이고, 격조사가 결합할 수 없으면 부사로 본다.

예: 여기가 어디인가? → 여기/npd
그가 여기 있다. → 여기/mad

● 인용격 조사, 종속적 연결어미, 종결어미의 '-라고'

• 인용격 조사의 '-라고' : 인용격 조사 '-라고'는 직접 인용의 경우에 사용되는 조사이다.

예: "그가 나를 사랑한다"라고 말했다. → 라고/jcr

• 종결어미와 인용격 조사가 결합하여 '-라고'의 형태를 띠기도 한다. 이 경우에는 종결어미 '-라'와 인용격 조사 '-고'로 분석한다.

예: 그걸 말이라고 하느냐? → 말/ncn+이/jp+라/ef+

고/jcr

• 종속적 연결어미의 '-라고-'

'-라고'가 인용격 조사와 형태가 동일하여도 인용격 조사로 사용되지 않고 종속적 연결어미로 사용되는 경우에는 '-라고'를 분석하지 않는다. '-라고'가 종속적 연결어미로 사용되는 경우에는 앞의 말이 뒤에 오는 말의 원인이나 근거가 됨을 나타낸다.

예: 몸이 정상이 아니라고(라고/ecs) 비관해서는 안 된다.

• '-라고'가 종결어미로 사용되는 경우 : '-라고'가 반문할 때 쓰인다.

예: 저것이 사슴이라고(이/jp+라고/ef)?

2.3.6 여러 가지 태깅이 가능한 경우

● 보조사(jxc)와 어미 - '란'

• '란'이 받침 없는 체언 뒤에 붙어 어떤 대상을 특별히 집어서 화제로 삼을 때 쓰는 경우에는 통용 보조사로 태깅한다. 이 때는 '란'을 주격 조사(이, 가)나 통용 보조사(은, 는)로 대치할 수 있다.

예문: 친구란/jxc 어려울 때 도와주는 것이 참다운 친구야.

(=친구는 어려울 때 도와주는 것이 참다운 친구야.)

• '란'이 '라고 하는'의 줄어든 말로 쓰인 경우에는 '라/ef+ㄴ/etm'으로 분석하여 태깅한다. 이 경우에는 '란'대신 주격 조사(이, 가)나 통용 보조사(은, 는)로 대치하면 문장이 성립되지 않는다.

예문: 철수란(라/ef+ㄴ/etm) 아이가 이 연극의 주인공이다.

(≠ 철수는 아이가 이 연극의 주인공이다.)

가란(라/ef+ㄴ/etm) 소리 못 들었니?

● 못하다

• 보조동사로 쓰인 경우

예문: 민수는 그 영화를 보지 못했다. (못하/px+었/ep+다/ef+./sf)

• 형용사로 쓰인 경우

예문:영희는 민수보다 못하다.(못하/paa+다/ef+./sf)

• '못하다'가 일정한 수준에 못 미치거나 할 능력이 없음을 나타내는 본동사로 쓰인 경우

예문:민수는 노래를 못한다.(못하/pvg+ㄴ다/ef+./sf)

• '못' 뒤에 나오는 동사가 본동사로 쓰인 경우에는 '못'을 부사(mag)로 태깅한다.

예문: 민수는 잠을 통 못 잤다. (못/mag+자/pvg+았/ep+다/ef+./sf)

● 안되다

• 형용사로 쓰인 경우

예문: 마음이 안됐다. (안되/paa+었/ep+다/ef+./sf)

얼굴이 안됐다. (안되/paa+었/ep+다/ef+./sf)

• 동사로 쓰인 경우

예문: 경기가 안 좋아서 장사가 잘 안된다. (안되/pvg+ㄴ다/ef+./sf)

공부가 잘 안된다. (안되/pvg+ㄴ다/ef+./sf)

• '안' 뒤에 나오는 동사가 본동사로 쓰인 경우에는 '안'을 부사(mag)로 태깅한다.

예문: 비가 안 온다. (안/mag+오/pvg+ㄴ다/ef+./sf)

다시는 그 사람을 안 만나겠다. (안/mag+만나/pvg+겠/ep+다/ef+./sf)

참고) 동사적 용법의 '안되다'는 흔히 '안 되다'와 혼동된다. 이런 경우에는 '잘되다'와 상반되는 개념인 경우에만 동사로 보아 붙여쓰고, 그 나머지는 구로 보아 띄어쓰면 된다.

※ 동사로 보아 붙여 써야 하는 경우

예문: 장사가 잘 안된다 ↔ 장사가 잘된다. (안되/pvg+ㄴ다/ef)

자식이 안되기를 바란다 ↔ 자식이 잘되기를 바란다. (안되/pvg+기/etn+를/jco)

안되어도 세명은 합격할 것이다 ↔ 잘되어야 세명 정도 합격할 것이다.

※ '부사+동사'로 분석해야 하는 경우

예문: 담배 크기의 절반이 채 안 되는... (안/mag+되/pvg+는/etm)

자신의 이익만을 생각해서는 안 되겠죠. (안/mag+되/pvg+겠/ep+지/ef+요/jxf)

1주일도 안 돼서...(안/mag+되/pvg+어서/ecs)

3. 문제시된 태깅의 예와 처리기준

어절	교정전	교정후	기준
켰다	키/pvg+었/ep+다/ef	켜/pvg+었/ep+다/ef	기본형이 '키다'가 아니라 '켜다'임
자괴감도	자괴감/ncn+도/jxc	자괴/ncn+감/xsn+도/jxc	'감'을 접미사로 인정한다
코웃음을	코/ncn+웃음/ncn+을/jco	코웃음/ncn+을/jco	'코'와 '웃음'을 분석했을 때 본래의 의미를 상실함
그럼	그렇/pad+ㅁ/etn	그럼/mag	'그럼'자체를 부사로 인정 cf)'그럼, 당연하지'의 경우에는 '그럼'이 감탄사(ii)
이봐요	이봐/npp+요/ef	이봐/ii+요/jxf	
냄비국수를	냄비/ncn+국수/ncn+를/jco	냄비국수/ncn+를/jco	음식이름은 분석하지 않음 분석하면 본래의 의미를 상실함
웃긴댄다	웃기/pvg+ㄴ/etm+대/nbn+ㄴ다/ef	웃기/pvg+ㄴ댄다/ef	'ㄴ단다'의 잘못
감동어리	감동/ncps+어리/xsm	감동/ncn+어리/pvg+ㄴ/etm	
고려인삼의	고려/nq+인삼/ncn+의/jcm	고려인삼/nq+의/jcm	'고려인삼'자체가 고유명사
꺼내입은	꺼내입/pvg+은/etm	꺼내/pvg+어/ecs+입/pvg+은/etm	복합동사는 분석하는 것을 원칙으로 한다
걸음이 (빠르다)	걷/pvg+음/etn+이jcs	걸음/ncn+이/jcs	문장안에서 체언의 역할을 할 때는 명사로, 서술의 역할을 할 때는 동사의 명사형으로 분석한다
(차도를) 걸음은	걷/pvg+음/etn+이jcs	걷/pvg+음/etn+이jcs	
바꾸곤 (한다)	바꾸/pvg+곤/ecs	바꾸/pvg+곤/ecs	반복을 나타낼 때는 종속적 연결어미로, '고는'의 줄임꼴일 때는 연결어미+통용보조사로 분석한다
바꾸곤 (가버렸다)	바꾸/pvg+곤/ecs	바꾸/pvg+고/ecs+ㄴ/jxc	
화젯거리	화제/ncn+거리/ncn+로/jca	화젯거리/ncn+로/jca	사이'ㅅ'과 결합한 합성어는 분석하지 않는다.

4. 결론

자동 검색이나 기계 번역 시스템 등의 연구가 활발해짐에 따라 대용량 코퍼스 구축의 중요성이 점점 커지고 있다. 물론 양적인 면에서의 발전도 중요하겠지만 좀더 일관성있는 유용한 코퍼스를 구축하는 것이 그 못지않게 중요하다. 본 연구는 이러한 점에서 보다 체계적인 태그 규정을 제시하였다.

본 연구에 이은 향후 과제로 접사 사전의 구축 및 조사 어미에 관한 연구를 들 수 있다. 사실 한국어에 있어서 명사를 처리하는데 접사는 매우 중요한 부분을 차지한다. 접사에 관한 연구가 좀더 체계적으로 이루어진다면 코퍼스에 나타난 명사를 처리하는데 유용한 정보가 될 것이다.

5. 참고문헌

[1] 한국어정보베이스를 위한 형태.통사 태그 표준에 관한 연구 최기선, 남영준, 김진규, 한영균, 박석문, 김진수, 이춘택, 김덕봉, 김재훈, 최병진 인지과학 1996, 12 Vol. 7, No. 4 P. 43-61

[2] 구문 트리 부착 코퍼스 구축을 위한 한국어 구문 태그 이공주, 김재훈, 최기선, 김길창 인지과학 1996, 12 Vol. 7, No. 4 P. 7-24

[3] 한국어 텍스트 분석을 통한 지식베이스 구축 기법 양기철, 최기선 인지과학 1996,12 Vol. 7, No. 4 P.203-216