형태소 깎는 노인: 국어사 자료를 위한 형태분석 보조기

김미경○, 박수지, 이상아
서울대학교 언어학과
saltpeanuts@snu.ac.kr, mam3b@snu.ac.kr, visualjan@snu.ac.kr

The POS Elderly: Semi-automatic annotation tool for Historical Korean

Migyeong Kim○, Suzi Park, Sana Lee
Seoul National University, Dept. of Linguistics

요 약

'형태소 깎는 노인'은 국어사 자료를 처리하는 고성능 자동 형태분석기의 개발이 난항을 겪고 있는 상황에서 수동으로 형태분석 작업을 하는 연구자들을 지원하기 위하여 개발된 형태분석 보조기이다. 인간과 기계의 분업을 통해 인간의 피로를 최대한 줄이고, 단순 반복 형태에 대해서는 정답을 확실하게 제안할 수 있다는 것이 특징이다. 국어사 자료에는 한국어 정보처리를 위해 필요한 어휘 사전이 없으므로, 문법형태소 사전을 만들어 이를 단서로 조사/어미부와 어간부를 구분하도록 하였다. 이를 통해 구축된 소규모 형태분석 말뭉치들이 장기적으로는 자동 형태분석기의 성능 개선에 일조할 수 있을 것으로 기대한다.

주제어: 국어사 자료, 말뭉치 언어학, 형태분석

1. 서론

지금까지 한국어 정보처리를 위한 다양한 패키지가 공개되어 널리 사용되고 있지만, 그 중에서 현대 이전의 한국어 자료를 처리할 수 있는 것은 아직 공개의 수준에 이른 것이 없다. 현대 이전의 한국어 자료에는 형태음운 규칙으로 기술하기 어려운 패턴의 자료가 다수 출현하고, 한자와 옛한글 처리 문제가 있으며, 텍스트 유형이 매우 다양한 데 비해 자동화된 형태분석기를 개발하기 위해 필요한 학습용 형태분석말뭉치의 양은 너무 적다.

특히 텍스트 유형에 따른 질 좋은 학습용 말뭉치의 양이 적다는 문제가 현대 이전의 한국어를 처리하는 데에 가장 큰 걸림돌이라고 할 수 있다. 국어사를 전공한 사람의 손과 시간을 들이지 않으면 해결되지 않는데, 현 시점에서 그러한 자원을 확보하기가 쉽지 않기 때문이다. 이 문제를 해결하려면 사람의 형태분석 작업을 도와주는 형태분석 보조기가 필요하다.

이 글은 현대 이전의 한국어 자료(이하 국어사 자료)에 형태분석 정보를 부착하는 작업을 돕는 형태분석 보조기 '형태소 깎는 노인'의 설계 및 구현에 대하여 기술한다. 현 단계에서는 19세기 말, 이른바 개화기 자료를 처리하기 위한 도구로서 개발되었으나 사전을 어떻게 구축하느냐에 따라서 19세기 이전의 한국어 자료나 방언 자료 등에도 사용이 가능하다.

2장에서는 국어사 자료에 대한 형태분석 시도와 현황에 대해 요약한다. 3장에서는 '형태소 깎는 노인'의 설계 사상을 소개한다. 4장에서는 '형태소 깎는 노인'의 설계 및 구현에 대하여 기술하고, 5장에서 요약 및 결론을 내린다.

2. 관련 연구

지금까지 보고된 국어사 자료를 처리하는 형태분석기로는 연세대학교 서상규 교수팀의 Histag[1, 2], 전북대학교 황용주 박사의 고어형태소분석 프로그램[1, 2], 경희대학교 김진해 교수 팀의 역사자료 형태분석 프로그램[2]이 있었다. 그러나 이 셋은 모두 공개되지 않고 있으며, 가장 최근에 개발된 역사자료 형태분석 프로그램의 경우에도, 독립신문 논설을 이 프로그램을 이용하여 1차로 주석한 후 수작업으로 상당한 부분을 교정했다는 사례[3]를 볼 때 만족할 만한 성능에 도달하려면 앞으로도 시간이 좀 더 걸릴 듯하다.

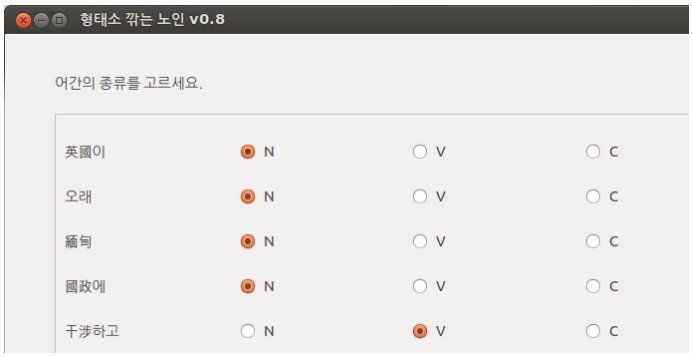
요약하면 국어사 자료를 위한 자동 형태분석기의 개발은 십 년 이상 난항을 겪고 있는 상황이다. 현 상황에서 국어사 자료에 대한 형태분석 말뭉치 구축은 매우 어려운 일이 되었으며, 지금까지 형태분석을 수행하여 공개된 자료도 21세기 세종계획에서 제공된 것이 전부이다. 이 때문에 국어사를 말뭉치 언어학의 방법으로 접근하고자 하는 연구자들은 대개 유니콩크(서울대 박진호 교수 개발), 깜짝새(전주대학교 소강춘 팀 개발) 등의 검색기를 최대한 활용하여 연구 대상 어형을 원시 말뭉치 안에서 찾는 방법을 쓰고 있다.

그러나 검색기를 활용하여 어형을 찾는 방식은 연구자가 알고 있는 어형 및 그 변이형만 찾을 있다는 점, 그리고 검색하기 쉽고 검색 결과가 지나치게 많이 나오지 않을 어형으로 연구가 쏠리게 된다는 점 등에서 한계가 있다. 국어사 연구에 계량적 방법론을 정식으로 도입하기 위해서는 역시 형태분석 말뭉치가 필요하고, 장기적으로는 구문 분석 말뭉치도 구축 대상으로 삼아야 한다고 본다. '형태소 깎는 노인'는 이러한 상황에서 전산화된 대량의 국어사 자료를 다루며 분투하는 개별 연구자들이 소규모로라도 형태분석 말뭉치를 구축하여 연구를 수행할 수 있도록 지원하고, 그렇게 개별적으로 만들어진 형태분석 작업물이 궁극적으로는 상호 호환되어 더 큰 형태분석 말뭉치로 합쳐질 수 있도록 기반을 제공하는 것을 목표로 한다.

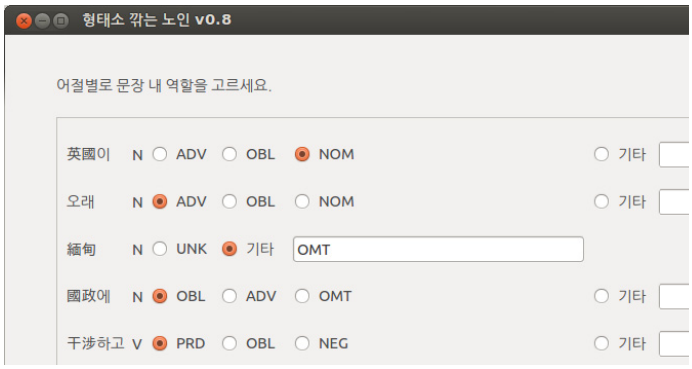
3. 설계 사상

'형태소 깎는 노인'을 떠받치고 있는 것은 적정기술의 관점이다. 상황에 따라서는 최신 기술보다 이전 수준의 기술이 더 활용하기에 좋을 수도 있다는 것이다. 최근의 현대 한국어 정보처리 도구들은 기계학습을 활용하는 경우가 많지만, 우리의 형태분석 보조기는 일부러 기계학습을 쓰지 않고 사전 대조와 통계적 방법에 따른 추정을 이용하였다. 현재 국어사 자료를 처리하는 수준으로는 학습을 아무리 하더라도 틀린 답이 나올 경우가 많아서, 주석자의 입장에서는 그 틀린 답을 보고 정정하는 것이 처음부터 직접 주석하는 것보다 더 피로하다. 따라서 처리할 문자열에 대응하는 형태분석 정보가 사전에 있으면 확률을 계산하여 후보를 제시하고, 없으면 추정하지 않고 사용자의 입력을 받도록 하였다.

그리하여 이 보조기는 인간이 잘 하는 처리는 인간에게, 기계가 잘 하는 처리는 기계에 맡기는 방식으로 설계되었다. 현 단계에서 국어사 자료에 등장하는 조사-어미 동형성을 기계가 처리하기는 어렵지만, 인간이 처리하기는 쉽다. 따라서 대상 어절이 명사류를 포함하는지 용언류를 포함하는지 인간이 판정하면, 기계가 그 정보를 받아서 문법형태소 실현 여부 및 그 종류를 추천한다. 기계가 대응할 수 있는 어절을 늘리는 것보다 처리하기 쉬운 어절에 대해서 확실하게 정답을 제시할 수 있도록 하는 것에 중점을 두었다.



[그림 1] 인간에 의한 어간 품사 판정



[그림 2] 문법형태소 종류/실현 여부 후보 제안

형태분석 보조기의 가장 중요한 기능은 주석자의 피로를 줄이는 것이다. 작은 단위로 작업하도록 유도하기 위하여, 원시말뭉치 파일을 불러오는 방식이 아니라 처리할 문단을 복사하여 붙여 넣는 방식으로 시작하게 하였다. 국어사 자료는 문헌 하나가 한 파일에 담겨 있는 경우도 많아서, 그런 긴 자료 파일을 한 번에 불러와서 작업하려고 하면 주석자가 금방 소모되기 때문이다. 다음으로 분석 단계를 쪼개서, 먼저 어간의 종류를 판정하고 문법 형태소의 종류를 고른 다음 이 정보를 가지고 형태소 분석을 하도록 하였다. 인간은 선택지가 많으면 쉽게 피로해지므로 단계별로 나누어서 선택지를 줄이고, 부가적인 효과로 기계의 추정도 더 쉬워지게 한 것이다. 구체적인 작동 방식은 4장에서 설명하겠다.

4. '형태소 깎는 노인'의 설계 및 구현

4.1. 사전 목록

국어사 자료를 처리할 때에는 현대 한국어 자료와 달리 기구축된 어휘 사전을 이용할 수 없다. 그러나 문법형태소는 어휘보다 목록이 짧고 반복적으로 사용되므로 사전을 구축하기가 상대적으로 쉽다. 현재 '형태소 깎는 노인'은 19세기 말, 이른바 개화기 자료에 맞춰서 개발되고 있어서, 21세기 세종계획에서 제공하는 형태분석 말뭉치 중에서 경향신문 29호-52호 4만5천여 어절을 대상으로 명사류(명사, 부사), 용언류(동사, 형용사), 명사+계사 어절에 실현된 문법형태소 목록을 뽑아서 조사/어미 사전을 구축하였다. 이 사전은 기본 버전과 사용자 입력으로 업데이트되는 버전의 두 가지로 존재하며, 불러올 때에는 두 버전을 모두 불러온다. 따라서 이 보조기로 형태분석 작업을 꾸준히 수행할수록 사전 항목이 늘어나므로 더 좋은 성능을 기대할 수 있다.

어간 사전	(접미사+)조사/어미 사전			
STEM_MAG	C_ACC	N_ACC	V_ACC	VC_ACC
STEM_NNG	C_ADJ		V_ADJ	VC_ADJ
STEM_NNP	C_ADV		V_ADV	VC_ADV
STEM_NP	C_CRD	N_CRD	V_CRD	VC_CRD
STEM_NR	C_GEN	N_GEN	V_GEN	VC_GEN
STEM_NNB	C_NEG		V_NEG	VC_NEG
STEM_VV	C_NOM	N_NOM	V_NOM	VC_NOM
	C_OBL	N_OBL	V_OBL	VC_OBL
	C_OMT	N_OMT	V_OMT	VC_OMT
	C_PRD		V_PRD	VC_PRD
	C_TPK	N_TPK	V_TPK	VC_TPK

[그림 3] 사전 목록

어간유형	통사적맥락	표제어	형태소분석
N	ACC	ㅅㅏㄹ·ㅁㅇ-	사ㄹ、ㅁ/NNB/을/JKO
N	ACC	ㄷㅡㄹㅇㅡㄹ	들/XSN/을/JKO
N	ACC	ㅈㅣㅗㅇㅇㅡ	죵/NNB/을/JKO
N	ACC	ㅇㅜㅓㄴㅇㅡ	원/NNB/을/JKO
N	ACC	ㄹ·ㄹ	ㄹ、ㄹ/JKO
N	ACC	ㄷㅏㅣㅅㅣㄴ	대신/NNG/을/JKO
N	ACC	ㅅㅅㅣㄹ·ㄹ	씨/XSN/ㄹ、ㄹ/JKO
N	ACC	ㅇㅣㅓㅊ·ㄹ-	여/XSN/ㅊ、/NNB/를/JKO
N	ACC	ㅇㅣㅓㅁㅣㅓ	여/XSN/명/NNB/을/JKO
N	ACC	ㅇㅣㅓㅊ·ㄹ	여/XSN/ㅊ、/NNB/ㄹ、ㄹ/JKO
N	ACC	ㄹㅣㅏㅇㅇㅡ	량/NNB/을/JKO

[그림 4] 사전의 항목 구조 예시 - N_ACC

주석자에게 제공되는 형태분석 후보는 기본적으로 이 조사/어미 사전을 이용하여 추정된다. 조사 사전에는 조사뿐만 아니라 접미사-조사나 자주 나오는 일반명사 조합까지 생산적으로 명사와 결합하는 것들은 가능한 한 모두 등록한다. 'STEM_'으로 시작하는 어간 사전을 둔 것은 형태분석의 결과 추출되는 어휘를 저장함으로써 향후 자동 형태소분석기를 개발할 때 필요할 사전 구축 자료를 수집하려는 목적이다.

4.2. 전처리

옛한글을 한양PUA 방식으로 입력해온 오랜 관행 때문에, 국어사 자료들은 새로 입력되는 자료들도 한양PUA로 입력 및 공개되는 경우가 많다. 본 개발팀은 장기적으로 옛한글 입력 방식이 첫가끝으로 통일되어야 한다고 생각하기 때문에, 보조기에 텍스트를 붙여 넣으면 우선 첫가끝으로 변환되게 하였다. 다음으로 국어사 자료의 대부분은 띄어쓰기가 없거나 띄어쓰기가 일관적이지 않다. 현대 한국어를 처리하듯이 공백을 기준으로 어절을 구분하면 처리 난이도가 상승한다. 따라서 사람이 띄어쓰기를 수정하고 어절을 구분함으로써 형태소 분석의 대상이 되는 어절 목록을 보조기에 넘기도록 하였다.

3장에서 잠시 언급하였듯이, 국어의 조사-어미 동형성은 국어사 자료의 처리 난이도를 올리는 원인 중의 하나이다. 국어사 자료 처리를 위한 어휘 사전이 없고, 어휘 및 문법형태소가 표현되는 철자법이 통일되지 않아서 변이형이 매우 많다. 이러한 상황에서 의지할 수 있는 것은 문법형태소일 확률이 매우 높은 어절 끝부분의 자소열이다. 그런데 '삼고'의 '고'가 용언에 결합한 어미 '-고'인지 명사의 일부를 이루는 '고'인지 어휘 사전 없이 기계가 고르기는 쉽지 않다. 그러나 인간은 맥락을 통해 이를 어려움 없이 구분할 수 있다. 만일 인간이 '삼고'가 용언류라고 알려주면, 기계 역시 어려움 없이 어미 사전과의 대조만으로도 '삼고'를 '삼/VV'과 '고/EM'로 구별할 수 있을 것이다. 따라서 어간의 종류를 인간이 판정하여 분석 대상 어절 문자열과 함께 보조기에 입력으로 넘기도록 하였다.

4.3. 통사적 맥락 추정

'빅셩이(백성/NNG + 이/JKS)'라는 어절을 예로 우리의 보조기가 어떻게 작동하는지 보도록 하자.

어절	빅셩이
어간종류	N
ㅣ	ADV: 62, OBL: 39, NOM: 12, OMT: 11, GEN: 6, CRD: 3
ㅇㅣ	NOM: 11, ADV: 3, OMT: 1
ㅇㅇㅣ	NOM: 2, OMT: 1
ㅕㅇㅇㅣ	NOM: 2, OMT: 1
ㅣㅕㅇㅇㅣ	NOM: 2
ㅅㅣㅕㅇㅇㅣ	None
ㄱㅅㅣㅕㅇㅇㅣ	None
ㅣㄱㅅㅣㅕㅇㅇㅣ	None
ㆍㅣㄱㅅㅣㅕㅇㅇㅣ	None
ㅂㆍㅣㄱㅅㅣㅕㅇㅇㅣ	None
합	ADV: 65, OBL: 39, NOM: 29, OMT: 14, GEN: 6, CRD: 3
문맥 후보	ADV, OBL, NOM
선택된 문맥	NOM

인간이 '빅셩이, N'이라는 입력을 넘기면, 우리의 보조기는 분석 대상 어절의 끝에서부터 한 단위씩 늘려가며 대조용 문자열을 만들어 사전과 대조한다. 이 경우 어간이 명사이므로 'N_' 접두사를 가진 사전에서 찾게 될 것이다. 'ㅣ'를 예로 보면, 'N_' 접두사를 가진 모든 사전에서 'ㅣ'로 끝나는 항목의 통사적 맥락 정보를 모두 추출하여 빈도를 센다. 다음으로는 'ㅇㅣ'로 끝나는 항목의 통사적 맥락 정보를 모두 추출하여 빈도를 센다. 이렇게 어절 전체가 대조될 때까지 숫자 세기를 반복한 다음, 추출되었던 통사적 맥락을 빈도가 높은 순서로 위에서 3개만 제시한다. 사전과 대조할 때에는 완전 일치와 후방 부분 일치를 모두 시도해 보았는데, 개화기 자료의 경우 후자가 성능이 가장 좋았기 때문에 기본값으로 설정하였다.[1)]

여기서 예로 든 '빅셩이'에서는 부사어(ADV), 기타성분(OBL), 주어(NOM)가 후보로 제시되었다. 이렇게 선택지가 제시되면 인간이 'NOM'을 선택하여 '빅셩이, N, NOM'을 다음 입력으로 넘기게 된다. 이 통사적 맥락 주석은 그 다음 단계인 문법형태소 분석의 정확도를 올리기 위한 장치이지만, 동시에 이 정보 자체가 한국어 문법사 연구자가 봐야 할 자료의 양을 줄여주고 형태소 검색을 넘어 문법관계를 직접 찾아볼 수 있게 해 주는 단서이기도 하다. 연구 목적에 따라서는 이 단계에서 주석을 끝낼 수도 있다.

4.4. 조사/어미부 문법형태소 추정

인간에게 '빅셩이, N, NOM'이라는 입력을 넘겨받은 보조기는 앞서 후방 부분 일치로 대조했던 조사/어미 사전을, 이번에는 완전 일치로 대조하여 일치하는 형태소 분석 정보를 후보로 제시한다. 완전 일치로 검색하는 것은 각종 변이형에 대응하면서 어간과 조사/어미를 정확히 구분하기 위한 것이다. 계속하여 '빅셩이'의 예를 보면 다음과 같다.

```
ㅂㆍㅣㄱㅅㅣㅕㅇㅇㅣ
ㆍㅣㄱㅅㅣㅕㅇㅇㅣ
ㅣㄱㅅㅣㅕㅇㅇㅣ
ㄱㅅㅣㅕㅇㅇㅣ
ㅅㅣㅕㅇㅇㅣ
ㅣㅕㅇㅇㅣ
ㅕㅇㅇㅣ
ㅇㅇㅣ
ㅇㅣ              Matched: [ㅇㅣ]   이/JKS
ㅣ                Matched: [ㅣ]    이/JKS
```

1) 성능을 측정할 때에는 보조기가 추천하는 후보 셋 중에서 정답이 있으면 맞춘 것으로 보았다. 성능을 확인할 때에는 한성주보 1, 2호와 제국신문 1898년 9월 4일의 논설 일부를 썼다.

문법형태소 후보 1 : [ㅇㅣ] 이/JKS
문법형태소 후보 2 : [ㅣ] 이/JKS

선택된 문법형태소: [ㅇㅣ] 이/JKS

이번에는 분석 대상 어절의 시작부터 한 단위씩 줄여가며 대조용 문자열을 만들고, 이를 지정된 종류의 조사/어미 사전과 대조하여 완전 일치 항목을 찾는다. 이 경우 N_NOM 사전과 대조했을 때 'ㅇㅣ'와 'ㅣ' 항목이 일치한 것이다. 대조용 문자열 중 사전 항목과 일치하는 것이 여럿일 경우, 대조용 문자열의 길이가 가장 긴 것부터 순서대로 세 개까지를 후보로 제시해 준다. 그러면 인간이 '[ㅇㅣ] 이/JKS' 항목을 골라서 '빅셩이, N, NOM'에서 어절 끝의 '이'는 '이/JKS'로 분석되어야 한다고 기계에게 알려줄 수 있다. 이렇게 객관식으로 만듦으로서 인간의 피로를 줄일 수 있고, 선택지의 일관성이라는 측면에서 기계의 도움을 받을 수 있게 된다.

4.5. 어간부 형태소 분석

이제 분석 대상 어절에서 어미를 제외한 나머지가 미분석 문자열로 남게 될 것이다. 우리의 보조기는 '빅셩이'에서 '이'를 제외한 나머지 문자열 '빅셩'을 어간으로 인지하고, 이 어간이 혹시 사전에 등록되어 있는지 확인한다. 이 경우 어간이 'N'이고 주격조사(JKS)가 결합할 수 있는 종류라고 인간이 정보를 주었으므로 NNG, NNP, NP, NR, NNB 중에서 검색한다. 없는 경우가 더 많을 것이므로, 이 단계에서는 없는 어휘를 등록하고, 혹시 접두사나 전처리 단계에서 분절하지 못한 관형사가 있을 경우 형태소분석을 추가로 수행하여 저장하는 작업이 주로 이뤄질 것이다. 사전에 있는 '빅셩'과 사전에 없는 '친구'를 비교하여 이 단계의 처리를 보이면 다음과 같다.

ㅂ·ㅣㄱㅅㅣㅓㅇ Matched:[빅셩] 빅셩/NNG
·ㅣㄱㅅㅣㅓㅇ
ㅣㄱㅅㅣㅓㅇ
ㄱㅅㅣㅓㅇ
ㅅㅣㅓㅇ Matched:[셩] 셩/NNG
ㅣㅓㅇ
ㅓㅇ
ㅇ
어간 후보 1 : [빅셩] 빅셩/NNG
어간 후보 2 : [셩] 셩/NNG
선택된 어간: [빅셩] 빅셩/NNG

ㅊㅣㄴㄱㅜ
ㅣㄴㄱㅜ
ㄴㄱㅜ
ㄱㅜ Matched:[구] 구/NR
어간 후보 1 : [구] 구/NR
새로 입력한 어간: [친구] 친구/NNG

어간에 대해서도 조사/어미 형태분석 단계와 마찬가지로 대조 문자열을 생성하여 완전히 일치하는 항목이 있는지 검색한다. 이를 통해 사전에 등록되어 있지 않은 접두파생, 관형사 결합, 복합명사나 파생명사 등도 부분적으로라도 포착할 수 있게 하였다. 그리하여 '빅셩'은 '빅셩'과 '셩'이 모두 후보로 제시된다. 이 경우는 선택을 하여 저장하는 것으로 끝났지만, '친구'처럼 '친구'가 사전에 없어서 찾아내지 못하고 '구'만 제시하는 경우 사람이 '친구'를 사용자 사전에 등록할 수 있다. 이러한 사용자 사전이 축적되면 결과물로서 분석 대상 텍스트의 어휘-품사 목록이 구축되므로 국어사 자료를 처리하기 위한 자동 형태분석기의 성능 개선에 한 발 더 다가갈 수 있을 것이다.

5. 결론 및 추후 과제

'형태소 깎는 노인'은 현재 기본 골격을 완성하고 개선 및 안정화 작업을 진행 중이다. 사전에 없는 조사/어미 문자열이 등장했을 때 조사/어미 사전에 등재하기 위한 처리가 아직 완료되지 않아서, 이를 개선할 필요가 있다. 또한 파이썬을 잘 모르는 사용자도 이용할 수 있어야 하므로 다양한 입력 오류에 대비하고 여러 플랫폼에서 작동할 수 있도록 조정해야 할 것이다.

'형태소 깎는 노인'은 국어사 자료를 위한 자동 형태분석기의 성능 개선을 한없이 기다릴 수 없는 연구자들을 위하여 수동 형태분석 작업을 돕기 위해 개발된 과도기적 도구이다. 향후 고성능의 국어사 자료 형태분석기가 만들어지면 용도를 다하고 잊히게 될 것이며, 그렇게 되어야 한다. 다만 그 시점이 언제인지 알 수 없기 때문에, 당장 활용할 수 있고 단순 반복 작업을 확실하게 줄여주는 도구를 개발한 것이다. 만의 하나 실용적인 자동 형태분석기의 출현을 앞으로도 상당 기간 기대할 수 없다는 것이 판명될 경우, 인간의 노력을 결합하기 위하여 '형태소 깎는 노인'을 CorA[4]처럼 웹 기반의 주석 도구로 발전시키는 것도 생각해 볼 수 있다. 현 단계에서는 시간과 인력과 자금의 문제로 로컬에서 작동하도록 만들었다.

지금까지 국어사 연구에는 특정한 용례의 출현을 바탕으로 한 논증이 많았지만, 형태분석 말뭉치를 구축하여 연구할 수 있다면 특정한 현상의 빈도를 바탕으로 한 논증도 시도해 볼 수 있게 된다. 이는 국어사에서 일어났던 다양한 변화들을 좀 더 입체적으로 포착하고 변화의 전조나 변화와 관련이 있는 요소들에 대하여 좀 더 적극적으로 검토할 수 있게 됨을 뜻한다. 국어사 자료의 형태분석기는 어떻게 개발하더라도 초기에는 국어사 전공자의 손과 시간을 요구하는 주석 작업이 들어가지 않을 수 없는 만큼, 주석자의 피로를 줄이고 작업 효율을 올리는 것은 큰 의미가 있다.

참고문헌

[1] 이태영, "국어국문학의 정보화 수용에 대한 논의의 반성과 전망", 語文硏究(어문연구학회), 제52집, pp.29~50, 2006.
[2] 김진해, 차재은, 김건희, 이의철, "歷史資料 형태

분석 프로그램 개발의 國語學的 意義와 活用 硏究-活字本 古小說을 중심으로”, 어문연구(한국어문교육연구회), 제37권, 제4호, pp.137~162, 2009.
[3] 강남준, 이종영, 최운호, “『독립신문』 논설의 형태 주석 말뭉치를 활용한 논설 저자 판별 연구 - 어미 사용빈도 분석을 중심으로”, 한국사전학, 제15호, pp.73~101, 2010.4
[4] M. Bollmann, F. Petran, S. Dipper, J. Krasselt, “CorA: A web-based annotation tool for historical and other non-standard language data”. In Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014, pp.86~90, Gothenburg, Sweden, April 26 2014.