상대빈도를 이용한 문법형태소의 인식 방법[1)]

A Method of Function-word Recognition by Relative Frequency

강 승 식*
(Seung-Shik Kang)

요 약 한글 문서에서 일부 조사와 일부 어미들은 출현 빈도가 높은 반면에 그 외의 조사와 어미는 출현 빈도가 낮을 것으로 추측되고 있다. 본 연구에서는 실험을 통해서 이러한 사실을 확인하고 한국어 분석 시스템에서 활용하기 위하여 조사와 어미의 상대적 출현 빈도를 조사하였다. 조사의 상대적 출현 빈도를 조사한 결과, 말뭉치 분야에 따라 약간의 차이가 있으나 일반적으로 빈도수가 높은 9개의 조사가 전체 조사의 70%를 차지하고 상위 20개, 32개, 69개의 조사가 각각 90%, 95%, 99%를 차지하고 있음을 확인하였다. 어말어미는 빈도수가 높은 10개의 어말어미가 전체 어말어미의 70%를 차지하고 상위 33개, 54개, 117개가 각각 90%, 95%, 99%를 차지하고 있다. 본 논문에서는 조사와 어미의 상대적 출현 빈도에 따라 문법형태소 사전을 구성하는 방법을 제안한다. 조사와 어미의 상대적 출현 빈도는 미등록어 추정이나 형태론적 중의성을 해결할 때도 활용된다.

주제어 조사/어미의 출현빈도, 문법형태소 사전 구성

Abstract It is expected that some Josa/Eomi's are frequently used and others are not in the Korean documents. In this paper, we confirm it through the experiment and show that such information is very useful for Korean language processing. In case of Josa, most frequent 9 Josa's occupied 70% of total Josa's and 20, 32, 69 Josa's occupied 90%, 95%, and 99% respectively. Similarly, most frequent 10 numbers of Eomi's occupied 70% of total Eomi's and 33, 54, 117 Eomi's occupied 90%, 95%, and 99% respectively. We propose a dictionary construction method for Josa/Eomi dictionary that is classified by the frequency information. Furthermore, Josa/Eomi frequency results are very useful for the identification of unregistered morphemes and the disambiguation of lexical ambiguities.

1. 서론

한국어의 어절은 어휘형태소와 문법형태소로 이루어지는데 어휘형태소의 수에 비해 문법형태소의 수는 많지 않다[1,2,3]. 또한, 어휘형태소에는 복합어와 신조어, 외래어 등 사전 미등록어가 많지만 문법형태소에는 미등록어가 드물다. 문법형태소의 수가 많지 않고 미등록 문법형태소는 드물다는 사실은 미등록어를 인식하거나 형태론적 중의성을 해결하는데 유용한 정보로서 활용될 수 있다[4,5,6].

문법형태소는 한국어 분석 시스템에서 문장 단위의 구문분석이나 의미분석에서도 중요한 역할을 한다. 그 동안 조사와 어미의 기능에 관한 많은 연구가 있었으나, 복합조사와 복합어미의 종류와 빈도수 등 한국어 정보처리에 필요한 자료는 많지 않다. 조사와 어미의 출현 빈도를 조사함으로써 고빈도 문법형태소에 가중치를 부여하거나 형태소 분석 등 한국어 분석 시스템에서 조사, 어미와 관련된 정보로서 유용하게 활용할 수 있다.

미등록어를 인식할 때는 음절수가 적고 저빈도 문법형태소가 분리된 분석결과의 우선순위를 낮추어 정확한 형태소 분석을 하는데 빈도특성이 사용되며, 한글 자동 띄어쓰기에서 어절 블록을 인식할 때 고빈도 조사/어미 특성이 유용하게 활용된다[7,8].

본 논문에서는 현대 국어의 말뭉치에서 조사와 어

* 한성대학교 정보전산학부
School of Information and Computer Engineering
Hansung University
Seoul 136-792, Korea

1) 이 논문은 1999년도 과학기술부 원자력 중장기 연구개발 사업(과제명: 치료방사선분야 정보관리 시스템 개발)의 지원을 받았음.

말어미를 추출하여 상대적인 출현빈도를 조사하여 문법형태소 사전을 구축하는 방법을 제안한다. 실험에 사용된 말뭉치는 약 160만 어절로 전산학 및 문헌정보학 논문 요약집 14만 어절(KT Test Set)[9], 신문기사에서 추출한 어휘 66만 어절, 초등학교 교과서에서 추출한 어휘 49만 어절, PC 통신에 공개된 소설 등 문학 작품에서 추출한 어휘 34만 어절이다.

2. 문법형태소의 특성

국어사전에 수록된 조사는 140여 개이고, 어말어미는 약 500여 개이다[4,10].[2] 단일조사뿐만 아니라 2개 이상의 조사가 결합된 복합조사를 포함한 조사의 수는 2,000여 개로 추정되고 있으며, 어말어미의 경우에도 복합어미와 조사가 결합된 유형을 포함했을 때 2,000~3,000여 개로 추정된다[11].

문법형태소는 조사와 어말어미, 선어말어미 등 허사(虛辭)들로 구성되는데, 언어는 시대에 따라 조금씩 변하므로 정확한 숫자를 파악하기가 어렵다. 그러나 문법형태소는 어휘형태소에 비해 상대적으로 변화 속도가 느린 편이며, 국어사전에 수록된 어휘의 수를 비교해 볼 때 문법형태소의 수는 매우 적다.

일반적으로 유한집합인 알파벳 T와 문법 G에 의해 생성되는 언어 L(G)는 무한집합이다[12]. 그러나 L(G)를 구성하는 스트링(string)의 최대 길이를 N으로 제한했을 때 T_N와 T_N의 부분집합인 언어 $L_N(G)$는 모두 유한집합이 된다. 실제로 조사와 어미는 음절수가 20 이상인 경우가 거의 없으므로 복합조사를 포함한 조사집합과 어말어미에 조사가 결합된 것을 포함한 어미집합의 경우에 최대 음절수가 유한하므로 유한집합이다. 즉, 임의의 조사 α와 임의의 어말어미 β, 그리고 스트링 최대길이 k에 대하여 스트링 α, β의 길이 $|\alpha|$, $|\beta|$는 $|\alpha| \leq k$, $|\beta| \leq k$이다.

(1) 복합조사와 복합어미를 포함한 조사와 어미는 미등록어가 거의 없다.

조사와 어미는 일반적으로 단일조사와 단일어미를 의미하는 경우가 많다. 그러나 본 논문에서 조사와 어미는 단일형과 복합형을 모두 포함하는 의미로 사용한다. 문법형태소의 개수는 유한하며, 어휘형태소에 비해 그 수가 많지 않으므로 대부분의 문법형태소를 나열할 수 있다. 또한, 조사와 어미의 결합형을 포함하더라도 조사집합과 어미집합은 유한집합이므로 이들을 모두 나열할 수 있다. 따라서 모든 문법형태소를 사전에 수록하는 것이 어렵지 않으며 미등록어가 거의 발생하지 않는다.

(2) 조사와 어미는 사용빈도에 따라 고빈도어와 저빈도어로 구분된다.

조사와 어미는 한글 문서에서 일부 조사와 일부 어미는 자주 사용되지만, 나머지 대부분의 것들은 드물게 사용된다. 한국어의 특성을 살펴보면 음절, 형태소, 어절 단위로 빈도수를 조사했을 때 말뭉치 유형에 따라 조금씩 차이가 있지만 각각 출현빈도가 비교적 일정하게 나타난다. 조사의 경우에 '은/는/이/가/을/를/의/에/로/...' 등 사용빈도에 따라 순서대로 나열할 수 있다. 어미는 조사에 비해 집중도가 약간 떨어지지만 '아/어/는/다/ㄴ/ㄹ/...'와 같이 사용빈도에 따라 고빈도어와 저빈도어로 구분된다. 이러한 현상은 3장의 말뭉치에 대한 실험 결과에 의하여 확인된다.

(3) 문법형태소는 형태소 분석 등 한국어 분석 시스템에서 중요한 역할을 한다.

어절은 어휘형태소와 문법형태소로 구성되고 형태소 분석은 각 어휘형태소와 문법형태소들을 분리하여 인식하는 과정이다. 그런데 어휘형태소는 그 수가 많고 복합명사와 신조어들로 인하여 모든 어휘를 사전에 수록하는 것이 불가능하기 때문에 미등록어가 다수 발생한다. 이에 비해 문법형태소는 상대적으로 그 개수가 많지 않고 미등록어가 거의 없다. 또한, 어절의 구조는 어휘형태소 하나와 문법형태소 0개 이상으로 구성되고, 특히 문법형태소가 2개 이상으로 구성되는 경우가 많으므로 형태소를 분리하고 인식할 때 문법형태소가 중요한 역할을 한다[13].

미등록어는 미등록 형태소가 하나 이상 포함된 것으로 어휘형태소가 사전에 수록되지 않은 경우가 대부분이다. 따라서 미등록어의 인식은 조사나 어미 등 문법형태소를 중심으로 가능한 모든 형태소들을 인식한 후에 인식되지 않은 어휘형태소를 추정하는 과정이다. 미등록 어휘형태소를 추정하려면 문법형태소를 인식하여 어절의 구조를 파악해야 하므로 문법형태소의 정확한 인식이 미등록어 인식 정확도에 많은 영향을 미치고 있다.

3. 조사/어미의 출현빈도

3.1 조사/어미 빈도조사

2) 옛 한글에서 사용되는 조사와 어미는 제외한다.

말뭉치에 나타난 조사와 어말어미의 단순빈도와 상대빈도를 조사하기 위하여 말뭉치를 수집하였다. 말뭉치 분야에 따라 문법형태소의 출현빈도가 달라질 수 있으므로 말뭉치를 두 가지 유형으로 구성하였다. 첫 번째 유형은 전문 분야에 대한 기술적인 문서로 문장의 유형에 약간의 제약이 있는 문서이다. 이러한 유형의 문서 집합으로 전산학 및 문헌정보학 분야 논문 집합인 KT Test Set과 신문기사에 대한 말뭉치이다. 두번째 유형은 다양한 유형의 단어와 문장 형태가 사용되는 문서로서 초등학교 교과서와 문학 작품에 대한 말뭉치이다. 각 말뭉치의 크기는 표 1과 같다.

〈표 1〉 조사/어미 추출 실험에 사용된 말뭉치

논문요약	신문기사	교과서	문학작품	전 체
14만 어절	66만 어절	49만 어절	34만 어절	163만 어절

말뭉치에서 조사와 어미를 추출하고 출현 빈도를 구하는 과정은 다음과 같다.

① 한국어 형태소 분석기를 이용하여 말뭉치의 각 단어에 대하여 형태소 분석을 한다.

② 형태소 분석된 단어 중에서 추정된 결과 및 분석실패 어절을 제외하고, 분석 성공 및 복합명사로 추정된 단어에 대하여 조사와 어미를 추출한다.

③ 추출된 조사를 출현 빈도에 따라 정렬(sorting)하여 각 조사의 출현 횟수(단순 빈도) 및 모든 조사의 출현 횟수 합계에 대한 각 조사의 상대적 출현 빈도(상대 빈도)를 계산한다. 어미에 대해서도 동일한 방법으로 단순빈도와 상대빈도를 계산한다.

빈도수 계산을 위해 사용된 형태소 분석기는 HAM 라이브러리이다.[3] 이 형태소 분석기는 체언과 용언의 형태소 분석 결과로 품사 정보를 단순화시켜 명사, 대명사, 의존명사, 수사 등 체언은 'N', 동사와 형용사 등 용언은 모두 'V'로 출력한다. 따라서 어휘형태소의 품사 중의성이 많지 않으나 분석 실패한 단어와 형태론적 중의성이 있는 단어가 포함되어 있다. 조사와 어미를 추출할 때 형태론적 중의성에 의하여 조사와 어미로 2가지 이상의 분석 결과가 존재할 때는 각 분석 결과를 조사, 어미의 출현 횟수로 중복하여 계산하였다.[4]

3) 한성대학교 한글공학연구소 홈페이지 참조. http://ham.hansung.ac.kr/

조사의 상대빈도 계산 방법은 다음과 같다. 말뭉치에 출현한 모든 조사에 대해 출현빈도가 높은 조사들을 순서대로 $j_1, j_2, \ldots, j_n$이라 할 때 조사 j_i의 단순빈도는 출현 횟수 $n(j_i)$이고 상대빈도 $p(j_i)$는 아래 식과 같이 계산된다. 동일한 방법으로 단순빈도가 높은 어미들을 순서대로 $O_1, O_2, \ldots, O_m$이라 할 때 어미 O_j의 단순빈도는 $n(O_j)$이고 상대 빈도는 $p(O_j)$이다.

$$p(j_i) = \frac{n(j_i)}{\sum_k n(j_k)} \times 100$$

3.2 출현빈도 분석

네 가지 유형의 말뭉치에서 출현하고 있는 조사와 어미에 대한 빈도 조사에서 자주 출현하는 조사와 어미의 수는 각각 표 2, 표 3과 같다[14]. 표 2는 조사 실험 결과에서 누적빈도가 70%, 90%, 95%, 99%를 차지하는 고빈도 조사의 개수이다. 표 2에서 평균 9개의 조사가 전체 출현 빈도의 70%, 평균 69개의 조사가 99%를 차지하고 있다. 즉, 고빈도 조사 9개만으로 형태소 분석을 하더라도 실제 문서에서 약 70%의 조사들을 분석할 수 있으며, 상위 69개의 조사만으로 전체 조사들의 99%를 분석할 수 있다.

표 3은 어말어미의 빈도수 조사에서 누적빈도가 각각 70%, 90%, 95%, 99%를 차지하는 고빈도 어말어미의 개수이다.[5] 표 3에서 평균 10개의 어말어미가 전체 출현 빈도의 70%를 차지하고 있으며, 평균 117개의 어말어미가 99%를 차지하고 있다. 즉, 고빈도

〈표 2〉 출현 빈도가 높은 상위 n개의 조사

말뭉치 / %	논문요약	신문기사	초등학교 교과서	문학작품	평 균
70%	8	9	9	9	9
90%	16	20	20	22	20
95%	25	31	32	39	32
99%	49	65	68	93	69

4) 한 어절의 조사(또는 어미) 분석결과가 2개 이상인 경우는 1로 계산한다.

5) 복합어미에서 명사형 전성어미 '-음/기'는 분리된 상태로 빈도수를 조사하였다. 즉, '-기가'는 '-가'로, '-음은'의 경우에는 '-은'으로 계산하였다.

〈표 3〉 출현 빈도가 높은 상위 n개의 어미

말뭉치 / %	논문요약	신문기사	초등학교 교과서	문학작품	평 균
70%	6	8	14	14	10
90%	16	28	40	47	33
95%	27	43	66	80	54
99%	60	99	139	170	117

어말어미 10개로 형태소 분석을 한다고 할 때 실제 문서에서 약 70%의 어말어미를 분석할 수 있으며, 117개의 어말어미만으로 99%의 어말어미를 분석할 수 있다.

실험 결과에 의하면 논문 요약집에 사용되는 조사의 수는 107개, 누적 빈도 99%에 속하는 조사의 수 49개로 다른 말뭉치에 비해 적게 나타난다. 이에 비해 소설 등 문학 작품에서 사용되는 조사의 수는 245개, 누적 빈도 99%에 속하는 조사의 수 93개로 다른 말뭉치에 비해 많이 나타난다. 이로부터 논문과 같이 특정 유형의 단어와 특정 유형의 문장이 많이 사용되는 문서와 문학 작품처럼 다양한 유형이 많은 문서는 조사와 어말어미의 수 및 출현 빈도에 차이가 있음을 확인할 수 있다. 이러한 현상은 어말어미의 경우에도 유사하다.

3.3 태깅된 말뭉치의 출현빈도 조사

약 75만 어절 크기의 태깅된 말뭉치인 ETRI 말뭉치[6]에서 조사와 어말어미의 출현빈도를 조사하였다. 빈도수에 따른 조사의 개수는 각각 8개(70%), 15개(90%), 25개(95%), 72개(99%)였고, 어말어미는 12개(70%), 40개(90%), 65개(95%), 148개(99%)였다. 태깅된 말뭉치의 조사 결과는 표 2, 표 3의 결과와 유사함을 알 수 있다.

4. 문법형태소 사전의 구성

형태소 분석을 위한 조사 사전과 어미 사전을 구성하는 방법으로는 단일조사와 단일어미만을 사전의 항목으로 구성하고 복합조사와 복합어미는 결합 정보를 이용하여 처리하는 방법이 있다[13]. 이 방법은 복합조사와 복합어미 사전이 구축되기 이전에 시도되었으며, 형태소 분석시에 조사와 어미 분해가 가능한 장점이 있으나 결합정보표의 크기가 커지는 단점이 있다. 차정원(1998)은 기본사전과 멀티 형태소 사전의 2중 구조를 사용하여 어휘형태소와 문법형태소 사이의 결합제약을 형태소 유형에 따라 분류하여 검사하는 방법을 취하고 있다[15]. 형태소 분석 사전을 구축할 때는 주로 복합조사와 복합어미를 모두 사전에 수록하는 방법이 사용된다[4,16,17]. 이 방법은 결합형을 모두 사전에 수록하므로 상대적으로 사전의 크기가 크지만 문법형태소간에 결합 제약 검사 문제가 해소되는 장점이 있다.

일반적으로 사전의 크기가 커지면 사전 탐색 부담이 커지게 되지만 160만 어절에서 추출된 조사는 250여 개이고, 어미는 440여 개이다. 그리고 실험 말뭉치에는 사용되지 않았지만 형태소 분석기에서 사용되고 있는 조사가 1,250여 개이고 어미가 1,370여 개임을 감안하더라도 어휘형태소 사전에 비해 매우 작다.[7] 따라서 복합조사와 복합어미를 사전에 모두 수록하는 방식을 취하는 것이 바람직하다. 다만, 구문분석 등에서 문장 성분을 파악할 때 조사, 어미의 대표형을 인식해야 하므로 복합조사와 복합어미를 단위형태소로 분해할 필요가 있다. 그런데 대부분의 복합조사와 복합어미는 분해 중의성이 없으므로 조사/어미 사전에 분해된 형태를 수록한다.[8]

조사/어미 사전은 출현 빈도에 따라 고빈도어와 저빈도어로 나누어 계층적으로 구성함으로써 사전 탐색 효율을 높일 수 있으며, 미등록어를 추정하거나 분석 결과에 대한 중의성을 해결할 때 우선 순위를 결정하는 중요한 정보로 활용될 수 있다. 고빈도어와 저빈도어를 구별하는 기준점을 발견하기 위하여 표 2의 평균에 대한 누적빈도는 (그림1)과 같다.

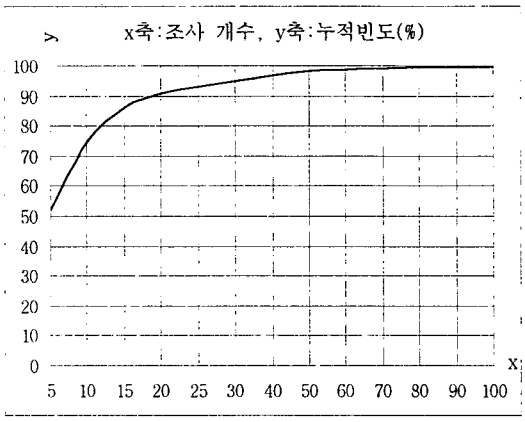
(그림1)에서 처리 증가율이 둔화되는 점과 처리율을 고려하여 조사 사전은 초고빈도어 15개, 고빈도어 35개, 기타 조사로 나누어 사전을 구축한다. 이 때 초고빈도어가 약 86%, 고빈도어가 12%, 기타 조사가 나머지 2%를 차지하게 된다.

표 3의 어말어미의 평균에 대한 누적빈도는 (그림 2)와 같다. 어미 사전은 (그림 2)에서 처리 증가율이

6) ETRI 자연어처리 연구부의 홈페이지 참조. http://aladin.etri.re.kr/~nlu/STANDARD

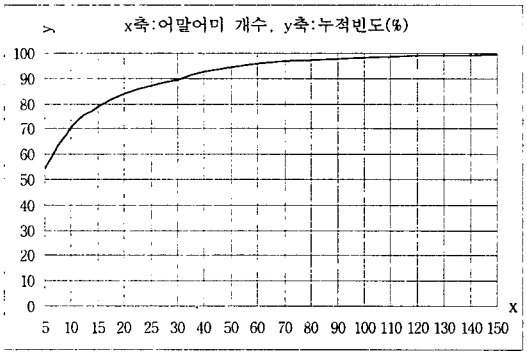
7) '-에게'와 '-께', '-게', '-한테'로 시작되는 조사들은 '-에게'로 시작되는 것만 수록했으며, '-에게'로 시작되는 조사의 수가 200여 개이므로 '-에게'의 변이체를 모두 수록하면 600여 개의 조사가 추가된다. 또한, 대화체 조사/어미가 누락되어 있으므로 이를 포함하면 더 많아진다.

8) 복합조사와 복합어미의 분해 중의성은 없으나 분해된 조사/어미의 기능은 2가지 이상이 가능한 것도 있다.



(그림 1) 조사의 누적 빈도

둔화되는 점과 처리율을 고려하여 초고빈도어 10개, 고빈도어 50개, 기타 어미 세 개의 그룹으로 나누어 사전을 구축한다. 초고빈도어가 약 71%, 고빈도어가 약 25%, 기타 어미가 나머지 4%를 처리하게 된다.



(그림 2) 어말어미의 누적 빈도

5. 결론

조사와 어미의 결합형은 매우 다양하며 일정한 결합규칙을 발견하기가 어려우므로 복합조사와 복합어미를 사전에 수록하는 것이 바람직하다. 한국어 정보처리에서 미등록어 인식이나 중의성 해결 등 여러 가지 응용 분야에서 활용할 수 있도록 160만 어절의 말뭉치에 출현한 조사와 어말어미의 상대적 출현 빈도를 조사하였다. 상대빈도에 대한 누적빈도를 살펴보면 소수의 조사/어미만이 자주 사용되고 대부분의 조사/어미는 드물게 사용됨을 확인하였다.

누적빈도의 증가율이 둔화되는 점과 처리 범위를 고려하여 조사/어미 사전을 각각 초고빈도 사전, 고빈도 사전, 기타로 구분하여 구축한다. 조사 사전을 초고빈도어 15개, 고빈도어 35개, 기타 조사로 구분했을 때 초고빈도 조사가 약 86%, 고빈도 어미가 12%, 기타 어미가 나머지 2%를 처리하게 된다. 어말어미는 초고빈도어 10개, 고빈도어 50개, 기타 어미로 구분했을 때 초고빈도 어미가 약 71%, 고빈도 어미가 25%, 기타 어미가 나머지 4%를 처리하게 된다. 조사와 어미의 상대적 출현 빈도는 형태소 분석 효율을 높이거나 미등록 어휘형태소를 추정할 때, 형태론적 중의성을 해결하는 데 유용한 정보로서 활용된다.

참고문헌

[1] 고영근, *국어 형태론 연구*, 서울대학교 출판부, 1989.

[2] 김진수, *국어 접속조사와 어미 연구*, 탑출판사, 1987.

[3] 서태룡, *국어 활용어미의 형태와 의미*, 탑출판사, 1988.

[4] 강승식, *음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석*, 서울대학교 컴퓨터공학과 박사학위 논문, 1993년 2월.

[5] 김덕봉, 최기선, 강재우, "한국어 형태소 처리와 사전 - 접속정보를 이용한 한글 철자 및 띄어쓰기 검사기 -", 어학연구, 26권, 1호, pp.87-113, 1990.

[6] 이은철, 이종혁, "계층적 기호 접속정보를 이용한 한국어 형태소 분석기의 구현", 제4회 한글 및 한국어 정보처리 학술발표 논문집, pp.95-104, 1992.

[7] 강승식, "한국어의 형태론적 모호성 유형 및 해결 방안", 제9회 한글 및 한국어 정보처리 학술발표 논문집, pp.83-87, 1997.

[8] 강승식, "한글 문장의 자동 띄어쓰기", 제10회 한글 및 한국어 정보처리 학술발표 논문집, pp.137-142, 1998.

[9] 김성혁 외 5인, "자동색인기 성능시험을 위한 Test Set 개발", 정보관리학회지, 11권 1호, pp.81-100, 1994.

[10] 금성출판사, *뉴에이스 국어 사전*, 금성출판사, 1989.

[11] 부산대학교, *조사의 유형*, Technical Report 90-1, 부산대학교 전자계산학과 인공지능 연구실, 1990.

[12] P. Denning, J. Dennis, J. Qualitz, *Machines, Languages, and Computation*, Prentice-Hall, 1978.

[13] 김성용, 최기선, 김길창, "Tabular Parsing 방법과 접속 정보를 이용한 한국어 형태소 분석기", 한국정보과학회 인공지능연구회 춘계 인공지능 학술발표회 논문집, pp.133-147, 1987.

[14] 강승식, "상대적 출현빈도를 이용한 조사/어미 사전의 구성", 제7회 한글 및 한국어 정보처리 학술발표 논문집, pp.188-194, 1995.

[15] J. W. Cha, G. B. Lee, and J. H. Lee, "Generalized unknown morpheme guessing for hybrid POS tagging of Korean," Proceedings of Sixth Workshop on Very Large Corpora in Colling-Acl 98, pp.85-93, 1998.

[16] 최재혁, 이상조, "양방향 최장일치법에 의한 한국어 형태소 분석기에서의 사전 검색 횟수 감소 방안", 정보과학회논문지, 20권, 10호, pp.1497-1507, 1993.

[17] H. C. Kwon, Y. S. Chae, and G. O. Jeong, "A Dictionary-based Morphological Analysis," Proceedings of Natural Language Processing Pacific Rim Symposium, pp.87-91, 1991.