상대적 출현 빈도를 이용한 조사/어미 사전의 구성

강 승 식
한성대학교 정보전산학부

A Construction of Josa/Eomi Dictionary using Relative Frequency

Kang, Seung-Shik
School of Information and Computer Engineering, Hansung University

요 약

한글 문서에서는 일부 조사와 일부 어미가 자주 출현하며 그 외의 조사/어미는 출현 빈도가 낮을 것으로 추측되고 있다. 본 연구에서는 실험에 통해서 이러한 사실을 확인하고 자주 출현하는 통합형 조사와 어미의 빈도를 구하기 위하여 한국어 말뭉치에서 통합형 조사와 통합형 어미의 상대적 출현 빈도를 조사하였다.

통합형 조사의 상대적 출현 빈도를 조사한 결과 말뭉치의 분야에 따라 약간의 차이가 있으나 평균 상위 9개의 통합형 조사가 전체 조사의 70%를 차지하고 있으며 상위 20개, 32개, 69개의 통합형 조사가 각각 90%, 95%, 99%를 차지하고 있음을 확인하였다. 통합형 어말어미의 경우에는 상위 10개의 통합형 어말어미가 전체 어말어미의 70%를 차지하고 상위 33개, 54개, 117개의 통합형 어미가 각각 90%, 95%, 99%를 차지하고 있다.

본 논문에서는 조사, 어미의 상대적 출현 빈도에 따라 계층적으로 조사/어미 사전을 구성함으로써 형태소 분석 효율을 높이고 형태소 분석기가 다양한 응용 분야에 쉽게 적응할 수 있도록 하는 방법을 제안한다. 또한 통합형 조사, 어미의 상대적 출현 빈도는 미등록어 추정을 용이하게 하거나 형태론적 모호성을 해결할 때에도 유용하게 활용될 수 있음을 보인다.

1. 서 론

한국어 단어의 구조는 '어휘형태소+문법형태소'로 이루어지는데 어휘형태소의 수에 비해 문법형태소의 수는 매우 적다[1,2,3]. 따라서 어휘형태소에는 복합어와 신조어, 외래어 등 사전 미등록어가 많지만 문법형태소에는 미등록어가 거의 없는 편이다. 형태소 분석 과정에서 문법형태소는 형태소를 분리할 때 중요한 역할을 하는데, 문법형태소의 수가 어휘형태소에 비해 매우 적다는 사실과 미등록 문법형태소는 거의 없다는 사실을 이용하면 문법형태소는 미등록 어휘형태소를 인식하는데 핵심적인 요소로 활용할 수 있다[4,5,6].

문법형태소는 한국어 분석시에 형태소 분석뿐만 아니라 구문 분석이나 의미 분석에도 중요한 역할을 하고 있다[7]. 그러나 문법형태소에 관한 연구는 그 중요성에 비추어 볼 때 미미한 실정이다. 현대 국어에서 사용되는 문법형태소의 수가 몇 개인지에 대해서도 검증된 바가 없고 단지 추정치만 알려져 있을 뿐 정확한 숫자가 알려져 있지 않다.

본 논문에서는 우선 현대 국어의 문어체에서 사용되는 문법형태소 중에서 자주 출현하는 통합형 조사와 통합형 어말어미를 발견하고 상대적인 출현빈도를 구한다. 통합형 조사와 통합형 어말어미의 상대적 출현 빈도를 구함으로써 자주 출현하는 문법형태소에 가중치를 부여할 수 있고 그 결과를 형태소 분석을 비롯한 한국어 분석 시스템에서 체계적으로 활용할 수 있다.

통합형 조사와 통합형 어말어미의 출현 빈도를 구하기 위하여 분야별로 말뭉치를 구성하였다. 실험에 사용된 말뭉치는 약 160만 어절로 전산학 및 문헌정보학 논문 요약집 14만 어절(KT Test Set), 신문 기사에서 추출한 어휘 66만 어절, 국민학교 교과서에서 추출한 어휘 49만 어절, PC 통신에 공개된 소설 등 문학 작품에서 추출한 어휘 34만 어절이다[8].

2. 문법형태소 집합의 특성

한국어 문법형태소 중에서 국어 사전에 수록된 단위 조사의 수는 140여개이고, 어말어미의 경우에도 국어 사전

에 수록된 것은 약 400여개이다[1][4,9]. 그러나 단위 조사와 2개 이상의 조사가 결합된 형태를 포함한 통합형 조사의 수는 약 1,000~3,000여개로 추정되고 있으며, 통합형 어미의 경우에도 어미와 어미 혹은 어미뒤에 조사가 결합된 형태를 포함하더라도 약 1,000~3,000여개로 추정된다(선어말어미가 결합된 것은 제외한다)[10].

[가정-1] 문법형태소 집합은 유한집합이며 어휘형태소에 비해 그 수가 매우 적다.

문법형태소의 집합은 조사와 어말어미, 선어말어미, 접두사, 접미사를 포함한 허사(虛辭)들로 구성되는데, 국어사전에 모든 문법형태소가 나열되었다고 볼 수 없으며, 또한 언어는 시대에 따라 조금씩 변하므로 정확한 숫자를 파악하기가 어렵다. 그러나 문법형태소는 어휘형태소에 비해 상대적으로 변화 속도가 느리며, 국어사전에 수록된 어휘의 수를 비교해 볼 때 문법형태소의 수는 매우 적다. 또한 현재 표준어로 채택되어 있는 어휘들을 각 품사별로 분류할 때 각 품사들의 집합은 유한집합이라 할 수 있고 문법형태소의 집합도 유한집합이라 할 수 있다.

[가정-2] 통합형 조사와 통합형 어말어미의 집합은 유한집합이다.

일반적으로 유한집합인 알파벳 T와 문법 G에 의해 생성되는 언어 L(G)는 무한집합이다[11]. 그러나 L(G)를 구성하는 string의 최대 길이를 제한하면 T^*와 T^*의 부분집합인 언어 L(G)는 모두 유한집합이 된다. 따라서 1개 이상의 조사로 이루어진 통합형 조사의 집합과 1개 이상의 어말어미와 0개 이상의 조사로 이루어진 통합형 어말어미 집합의 경우에 최대 스트링의 길이가 유한하다고 가정할 때 유한집합이 된다. 즉, 임의의 통합형 조사 α와 임의의 통합형 어말어미 β, 그리고 스트링의 최대길이 c에 대하여 스트링 α, β의 길이 |α|, |β|는 $|\alpha| \le c$, $|\beta| \le c$이다.

[정리-1] 문법형태소에는 미등록어가 거의 없다.

가정-1에 의하여 문법형태소의 집합은 유한집합이며 그 수가 많지 않다. 또한 문법형태소는 언어의 변화 속도가 매우 느리므로 거의 모든 문법형태소를 문법형태소 사전에 나열할 수 있다. 가정-2에 의하여 통합형 조사와 통합형 어말어미를 문법형태소 사전에 수록하더라도 그 집합은 역시 유한집합이며 대부분의 문법형태소를 사전에 수록할 수 있다. 따라서 문법형태소에는 미등록어가 거의 없다.

1) 현대 국어에 사용된 것만 계산하고 고어는 제외한다.

[정리-2] 통합형 조사와 통합형 어미는 한글 문서에서 일부 조사와 일부 어미는 자주 사용되지만 그 외의 것은 드물게 사용된다.

한글 문서에 나타나는 통합형 조사는 수천 개이다. 그 중에서 '은/는/이/가/을/를/의/에/로' 등은 자주 사용되는 반면 대부분의 통합형 조사는 문서에서 드물게 나타난다. 통합형 어말어미의 경우에는 조사보다 집중도가 약간 떨어지지만 '아/어/는/다/ㄴ/ㄹ' 등이 자주 사용되고 대부분의 통합형 어말어미는 드물게 사용된다. 이러한 현상은 말뭉치에 대한 실험 결과에 의하여 확인된다.

[정리-3] 문법형태소는 형태소 분리 과정과 미등록 어휘 형태소의 인식 과정에서 중요한 역할을 한다.

국어 단어는 어휘형태소와 문법형태소로 이루어져 있고 형태소 분석은 각 어휘형태소와 문법형태소들을 분리-인식하는 것이다. 그런데 어휘형태소는 그 수가 많고 모든 어휘를 사전에 수록하는 것이 불가능하므로 일반적으로 미등록어가 다수 발생한다. 이에 비해 문법형태소의 수는 많지 않으며, 거의 모든 어휘를 사전에 수록할 수 있으므로 미등록어가 거의 없다. 또한 국어 단어의 구조는 어휘형태소 하나에 문법형태소 0개 이상으로 구성되는데 문법형태소가 2개 이상으로 구성되는 유형이 많다[12]. 따라서 형태소를 분리하고 인식하는 과정에서 문법형태소는 중요한 역할을 하게 된다.

미등록 어휘형태소의 인식은 단어의 구조를 파악하여 단어를 구성하고 있는 가능한 모든 형태소들을 인식한 후에 인식되지 않은 어휘형태소 부분을 미등록어로 추정하는 과정이다. 이 때 단어의 구조에 따라 미등록 어휘형태소를 제외한 다른 형태소(문법형태소)를 인식하기 위하여 문법형태소는 매우 중요한 역할을 한다. 이 과정에서 미등록 어휘형태소와 이웃한 문법형태소를 잘못 인식하면 미등록어 인식 오류가 발생하게 되므로 문법형태소의 인식 결과는 미등록 어휘형태소의 인식률을 좌우하는 핵심적인 요소이다.

3. 문법형태소의 출현빈도

3.1 말뭉치의 구성 및 실험 환경

문법형태소 중에서 통합형 조사와 통합형 어말어미의 출현빈도를 구하고 전체 조사/어미에 대하여 각각의 상대적 출현빈도를 구하기 위하여 한글 문어체 말뭉치에 대하여 실험하였다. 말뭉치의 성격에 따라 문법형태소의 출현빈도가 달라질 수 있으므로 말뭉치를 두 가지 유형으로 구성하였다. 첫번째 유형은 전문 분야에 대한 기술적인 문

시로 문장 유형에 제약이 있는 문서이다. 이러한 유형의 문서 집합으로 전산학과 문헌정보학 분야의 논문 요약집(KT Test Set), 그리고 신문 기사에 대한 말뭉치를 구성하였다. 두번째 유형은 다양한 유형의 단어와 문장 형태가 사용되는 문서이다. 이 유형의 문서 집합으로는 국민학교 교과서와 소설 등 문학 작품에 대한 말뭉치이다. 각 말뭉치의 크기는 표1과 같다.

표1. 조사/어미 추출 실험에 사용된 말뭉치

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

표1의 각 말뭉치에서 통합형 조사와 통합형 어말어미를 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 복합명사 추정된 단어에 대하여 통합형 조사와 통합형 어미를 추출한다.
③ 추출된 통합형 조사를 출현 빈도에 따라 sorting하여 각 조사의 출현 횟수(단순 빈도) 및 모든 조사의 출현 횟수 합계에 대한 각 조사의 상대적 출현 빈도(상대 빈도)를 계산한다. 통합형 어말어미에 대해서도 같은 방법으로 단순 빈도와 상대 빈도를 계산한다.

단계①에서 한국어 형태소 분석기는 PC 통신망에 실험 및 연구용으로 공개된 'HAM version 1.51'을 사용하였다. 본 실험에서는 형태소 분석 결과만을 이용하여 통합형 조사와 통합형 어말어미를 추출하였기 때문에 분석 실패한 단어, 형태론적 모호성이 있는 단어 등 문장의 구조 및 의미에 적합하지 않는 분석 결과가 존재할 수 있다.

단계②에서 통합형 조사와 통합형 어미를 추출할 때 형태론적 모호성에 의하여 조사와 어미로 2가지 이상의 분석 결과가 존재하는 단어에 대해 각 분석 결과를 조사, 어미의 출현 횟수로 중복하여 계산한다(한 단어에 대한 통합형 조사 혹은 통합형 어말어미가 2개 이상인 경우에는 1개로 계산한다). 실험에 사용된 HAM version 1.51은 체언과 용언의 형태소 분석 결과로 품사 정보를 단순화시켜 명사, 대명사, 의존명사, 수사 등 체언은 'noun', 용언은 모두 'verb'로 출력하므로 어휘형태소의 품사 모호성이 거의 없다.

단계③에서 통합형 조사에 대한 상대적 출현 빈도의 계산 방법은 다음과 같다. 말뭉치에 출현한 통합형 조사의 수가 n개이고 각 통합형 조사를 단순 빈도(term frequency)가 높은 순으로 $j_1, j_2, ..., j_n$이라 할 때 조사 j_i에 대한 단순 빈도는 j_i의 출현 횟수 $n(j_i)$이다. 조사 j_i의 상대 빈도(relative frequency)를 $p(j_i)$라 할 때 $p(j_i)$는 다음과 같이 계산한다.

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

통합형 어말어미를 단순 빈도가 높은 순으로 $o_1, o_2, ..., o_m$이라 할 때 임의의 통합형 어말어미 o_j에 대한 단순 빈도는 o_j의 출현 횟수 $n(o_j)$이고 상대 빈도는 $p(o_j)$이다.

3.2 실험 결과

4가지 분야의 말뭉치에 대한 통합형 조사와 통합형 어미의 실험 결과는 부록 1, 부록 2와 같다. 부록에서 각 말뭉치에 대한 기술 내용은 순서대로 조사(또는 어미), 단순 빈도, 상대 빈도, 그리고 상대 빈도에 대한 누적 빈도이다.

표2는 통합형 조사에 대한 실험 결과(부록 1)에서 누적 빈도가 70%, 90%, 95%, 99%인 고빈도 통합형 조사의 수이다.

표2. 출현 빈도가 높은 상위 n개의 통합형 조사

말뭉치 / %	논문요약	신문기사	국민학교 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

표2에서 평균 9개의 통합형 조사가 전체 출현 빈도의 70%를 차지하고 있으며 평균 69개의 조사가 99%를 차지하고 있음을 알 수 있다. 즉, 고빈도 통합형 조사 9개로 형태소 분석을 한다고 할 때 실제 문서에서 약 70%의 조사를 분석해 낼 수 있으며, 상위 69개의 조사만으로 전체 조사의 99%를 분석할 수 있음을 알 수 있다.

표3. 출현 빈도가 높은 상위 n개의 통합형 어미

말뭉치 / %	논문요약	신문기사	국민학교 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

표3은 통합형 어말어미에 대한 실험 결과(부록2)에서 누적 빈도가 70%, 90%, 95%, 99%인 고빈도 통합형 어말어미의 수이다. 표3에서 평균 11개의 통합형 어말어미가 전체 출현 빈도의 70%를 차지하고 있으며 평균 117개의 어말어미가 99%를 차지하고 있음을 알 수 있다. 즉, 고빈도 통합형 어말어미 11개로 형태소 분석을 한다고 할 때 실제 문서에서 약 70%의 어말어미를 분석해 낼 수 있으며, 69개의 어말어미만으로 99%의 어말어미를 분석할 수 있음을 알 수 있다.

표2와 부록1로부터 논문 요약집에 사용되는 통합형 조사의 수는 107개, 누적 빈도 99%에 속하는 통합형 조사의 수 49개로 다른 말뭉치에 비해 적게 나타난다. 이에 비해 소설 등 문학 작품에서 사용되는 통합형 조사의 수는 245개, 누적 빈도 99%에 속하는 통합형 조사의 수 93개로 다른 말뭉치에 비해 많이 나타난다. 즉, 논문과 같이 특정 유형의 단어와 특정 유형의 문장이 많이 사용되는 문서와 문학 작품처럼 다양한 유형이 많은 문서는 통합형 조사와 어말어미의 수 및 출현 빈도에 차이가 있음을 확인할 수 있다. 이러한 현상은 통합형 어말어미의 경우에도 비슷하다(표3, 부록2 참조).

4. 조사/어미 사전의 구성

형태소 분석을 위한 조사 사전과 어미 사전을 구성하는 방법으로는 단위 조사와 어미를 사전의 항목으로 구성하고 조사간의 결합, 어미간의 결합은 접속정보표를 참조하여 검사하는 방법과 통합형 조사, 통합형 어말어미를 사전의 항목으로 구성하는 방법이 있다[13,14,15,16]. 첫번째 방법은 사전의 크기가 작은 대신 접속정보표를 사용해야 하기 때문에 새로운 결합 유형을 등록하려면 접속정보표를 갱신해야 하므로 추가 및 수정이 용이하지 않다.

두번째 방법은 통합형을 모두 사전에 수록하므로 사전의 크기가 커지는 단점이 있지만 수정 및 삭제가 용이하다. 일반적으로 사전의 크기가 커지면 사전 탐색 부담이 커지게 된다. 그러나 본 논문의 실험에 사용된 160만 어절에 출현한 통합형 조사와 통합형 어말어미는 각각 250여개(부록1), 440여개(부록2)로 단위 형태소를 수록한 경우 140여개, 400여개와 비교할 때 통합형을 모두 수록하더라도 사전의 크기가 커지는 부담은 거의 없다. 따라서 조사/어미 사전은 통합형을 사전에 수록하는 방식을 취하는 것이 바람직하다.

통합형 조사/어미 사전을 구성할 때는 통합형 조사와 통합 어말어미의 출현 빈도에 근거하여 고빈도어와 저빈도어로 나누어 계층적으로 구성하면 사전 탐색의 효율을 높일 수 있다. 이 때 고빈도어와 저빈도어를 구별하는 기준점을 발견하기 위하여 부록1, 부록2의 누적 빈도를 도표로 나타내면 각각 그림1, 그림2와 같다.

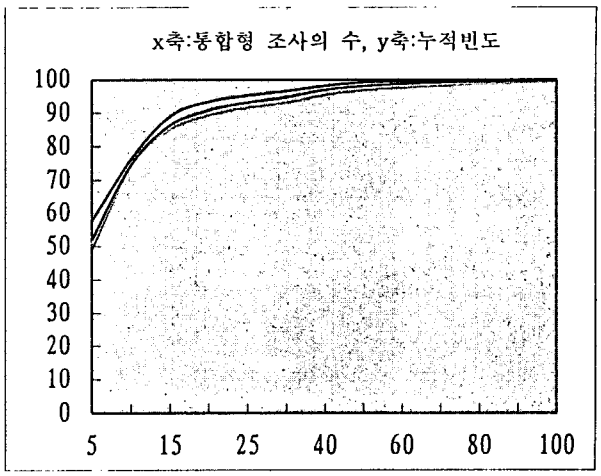

그림 1. 통합형 조사의 누적 빈도

통합형 조사 사전은 그림1에서 처리 증가율이 둔화되는 점과 처리율을 고려하여 초고빈도어 15개(조사그룹1), 고빈도어 35개(조사그룹2), 그 외의 조사(조사그룹3)로 세 개의 그룹으로 나누어 계층 구조로 사전을 구성하면, 조사그룹1이 약 86%, 조사그룹2가 12%, 조사그룹3이 나머지 2%를 처리하게 된다.

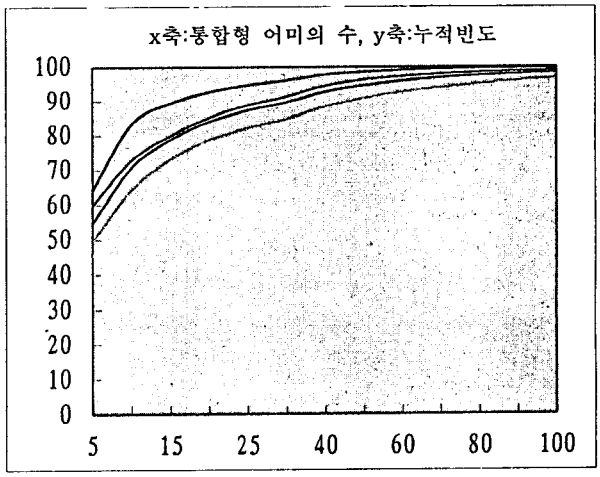

그림 2. 통합형 어말어미의 누적 빈도

통합형 어말어미 사전은 그림2에서 처리 증가율이 둔화되는 점과 처리율을 고려하여 초고빈도어 10개(어미그룹1), 고빈도어 50개(어미그룹2), 그 외의 어미(어미그룹3)로 세 개의 그룹으로 나누어 계층 구조로 사전을 구성하면, 어미그룹1이 약 71%, 어미그룹2가 25%, 어미그룹3이 나머지 4%를 처리하게 된다.

5. 결 론

국어의 조사와 어미는 서로 다양하게 결합된 통합형이 존재한다. 형태소 분석을 위해서는 다양한 통합형 조사와 통합형 어말어미를 문법형태소 사전으로 구성해야 하는데 이 때 사전의 효율적인 구축을 위해 분야별로 160만 어절의 말뭉치에서 출현한 조사와 어말어미의 상대적 출현 빈도를 조사하였다.

상대적 출현 빈도에 대한 누적 빈도를 조사한 결과 한 글 문서에서는 소수의 조사와 어말어미만이 자주 출현할 뿐이고 대부분의 조사와 어말어미는 드물게 출현함을 확인하였다. 누적 빈도의 증가율이 둔화되는 점과 처리 범위를 고려하여 조사/어미 사전을 구축할 때 각각 초고빈도 사전(Ultra-High Frequency Dictionary: UHFD), 고빈도 사전(High Frequency Dictionary: HFD), 기타로 나누어 계층적으로 구성하였다.

통합형 조사 사전을 초고빈도어 15개, 고빈도어 35개, 기타 조사로 나누어 계층 구조로 구성하면, 초고빈도 조사가 약 86%, 고빈도 어미가 12%, 기타 어미가 나머지 2%를 처리하게 된다. 통합형 어미 사전은 초고빈도어 10개, 고빈도어 50개, 기타 어미로 구성하면, 초고빈도 어미가 약 71%, 고빈도 어미가 25%, 기타 어미가 나머지 4%를 처리하게 된다.

조사와 어미의 상대적 출현 빈도는 형태소 분석시에 미등록 어휘형태소를 추정하거나 한국어 tagging 시스템에서 형태론적 모호성을 해결할 때도 유용한 정보로서 활용할 수 있을 것이다.

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] 최기선, "한국어 해석을 위한 격이동 패턴의 고찰", 인지과학, 민음사, pp.364-388, 1989.

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

[9] 금성출판사, *뉴에이스 국어 사전*, 금성출판사, 1989.

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

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

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

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

[14] 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.

[15] A. Copestake, "An Approach to Building the Hierarchical Element of a Lexical Knowledge Base from a Machine-Readable Dictionary," Proceedings of a Workshop on Inheritance in Natural Language Processing, 1990.

[16] Boguraev, Btanimir and Mary Neff, *Dictionary Structure and Lexical Relations: A Study in Applied Computational Lexicography*, IBM Technical Report, IBM T.J. Watson Research Center, Yorktown Heights, New York, 1990.

부록 1. 상위 70개 통합 조사의 상대적 출현빈도

KT Data Set				신문 기사				국민학교 교과서				문학 작품			
을	8944	16.97	16.97	을	25941	12.98	12.98	1	을	30266	15.50	15.50	을	16498	13.65	13.65
의	8708	16.52	33.48	의	22161	11.09	24.06	2	의	21511	11.02	26.51	의	16023	13.26	26.92
를	5350	10.15	43.63	에	20200	10.10	34.17	3	를	16400	8.40	34.91	이	10296	8.52	35.44
에	4950	9.39	53.02	이	15021	7.51	41.68	4	이	16185	8.29	43.20	에	9074	7.51	42.95
으로	2378	4.51	57.53	를	14937	7.47	49.15	5	에	14945	7.65	50.85	를	8297	6.87	49.81
은	2366	4.49	62.02	으로	11973	5.99	55.14	6	는	12061	6.18	57.03	는	8189	6.78	56.59
이	2202	4.18	66.20	은	11615	5.81	60.95	7	은	11419	5.85	62.88	가	6787	5.62	62.21
는	1916	3.63	69.83	는	11426	5.72	66.67	8	가	9898	5.07	67.95	은	6109	5.06	67.27
에서	1672	3.17	73.00	가	8348	4.18	70.84	9	로	5839	2.99	70.94	로	4239	3.51	70.77
과	1630	3.09	76.10	로	6547	3.28	74.12	10	고	5349	2.74	73.68	으로	4107	3.40	74.17
에서는	1598	3.03	79.13	하고	5639	2.82	76.94	11	과	5056	2.59	76.26	도	3368	2.79	76.96
로	1486	2.82	81.95	과	5244	2.62	79.56	12	도	4923	2.52	78.79	에서	3250	2.69	79.65
하고	1340	2.54	84.49	에서	4984	2.49	82.06	13	에서	4783	2.45	81.24	고	2868	2.37	82.02
가	1332	2.53	87.01	도	3795	1.90	83.95	14	으로	4495	2.30	83.54	과	1822	1.51	83.53
와	1014	1.92	88.94	와	3169	1.59	85.54	15	와	3796	1.94	85.48	하고	1627	1.35	84.88
고	610	1.16	90.09	이다	2905	1.45	86.99	16	에는	2605	1.33	86.81	이다	1403	1.16	86.04
이다	444	0.84	90.94	고	2576	1.29	88.28	17	여	1954	1.00	87.82	와	1137	0.94	86.98
여	444	0.84	91.78	부터	1624	0.81	89.09	18	께	1581	0.81	88.62	만	1030	0.85	87.83
으로써	426	0.81	92.59	까지	1506	0.75	89.85	19	하고	1463	0.75	89.37	에게	949	0.79	88.62
도	404	0.77	93.35	께	1317	0.66	90.51	20	나	1225	0.63	90.00	께	762	0.63	89.25
다	230	0.44	93.79	에는	1254	0.63	91.13	21	에게	1205	0.62	90.62	나	751	0.62	89.87
에서의	226	0.43	94.22	이라고	1043	0.52	91.66	22	이다	1076	0.55	91.17	에는	615	0.51	90.38
든	192	0.36	94.58	만	1011	0.51	92.16	23	만	1052	0.54	91.71	며	454	0.38	90.76
과를	190	0.36	94.94	라고	957	0.48	92.64	24	이나	1008	0.52	92.22	든	439	0.36	91.12
보다	140	0.27	95.21	보다	901	0.45	93.09	25	라	866	0.44	92.67	라고	427	0.35	91.47
라	136	0.26	95.47	에도	775	0.39	93.48	26	께서	828	0.42	93.09	대로	418	0.35	91.82
나	134	0.25	95.72	다	746	0.37	93.85	27	야	816	0.42	93.51	처럼	411	0.34	92.16
에는	132	0.25	95.97	토록	690	0.35	94.20	28	아	666	0.34	93.85	라	377	0.31	92.47
으로서	120	0.23	96.20	에게	641	0.32	94.52	29	다	637	0.33	94.18	이나	340	0.28	92.75
만	114	0.22	96.41	나	582	0.29	94.81	30	에서는	621	0.32	94.50	이라고	336	0.28	93.03
이나	102	0.19	96.61	대로	535	0.27	95.08	31	며	572	0.29	94.79	아	333	0.28	93.31
며	102	0.19	96.80	에서는	532	0.27	95.34	32	보다	561	0.29	95.08	에도	325	0.27	93.57
로부터	100	0.19	96.99	이나	487	0.24	95.59	33	께서는	561	0.29	95.36	다	310	0.26	93.83
로서	98	0.19	97.18	이며	404	0.20	95.79	34	까지	545	0.28	95.64	까지	309	0.26	94.09
에게	92	0.17	97.35	요	380	0.19	95.98	35	니	510	0.26	95.90	야	281	0.23	94.32
까지	86	0.16	97.52	든	374	0.19	96.17	36	대로	465	0.24	96.14	니	278	0.23	94.55
과는	84	0.16	97.67	으로써	344	0.17	96.34	37	라고	464	0.24	96.38	부터	268	0.22	94.77
께	80	0.15	97.83	같이	315	0.16	96.49	38	처럼	459	0.24	96.61	에서는	246	0.20	94.97
만을	76	0.14	97.97	로는	313	0.16	96.65	39	이라고	433	0.22	96.84	이며	232	0.19	95.17
으로는	68	0.13	98.10	밖에	296	0.15	96.80	40	부터	418	0.21	97.05	이라는	228	0.19	95.36
대로	66	0.13	98.22	과의	290	0.15	96.94	41	에도	400	0.20	97.25	이라	215	0.18	95.53
이고	60	0.11	98.34	며	283	0.14	97.09	42	마다	311	0.16	97.41	엔	210	0.17	95.71
에도	60	0.11	98.45	로부터	262	0.13	97.22	43	요	275	0.14	97.55	보다	179	0.15	95.86
과의	60	0.11	98.57	처럼	262	0.13	97.35	44	든	260	0.13	97.69	여	177	0.15	96.00
로의	56	0.11	98.67	아	254	0.13	97.48	45	이고	242	0.12	97.81	이고	161	0.13	96.13
으로부터	54	0.10	98.77	라	242	0.12	97.60	46	같이	198	0.10	97.91	에서의	138	0.11	96.25
이며	50	0.09	98.87	여	227	0.11	97.71	47	에서도	163	0.08	98.00	에서도	118	0.1	96.35
에서도	46	0.09	98.96	으로는	227	0.11	97.82	48	이라	157	0.08	98.08	만이	115	0.1	96.44
로써	42	0.08	99.04	이고	227	0.11	97.94	49	까지의	129	0.07	98.14	으로써	112	0.09	96.53
로는	40	0.08	99.11	에서의	206	0.10	98.04	50	로부터	116	0.06	98.20	라도	111	0.09	96.63
까지의	38	0.07	99.18	이라는	186	0.09	98.13	51	로가	114	0.06	98.26	으로는	109	0.09	96.72
요	38	0.07	99.26	만에	164	0.08	98.22	52	으로는	108	0.06	98.32	로부터	108	0.09	96.81
이라는	28	0.05	99.31	으로부터	164	0.08	98.30	53	로는	107	0.05	98.37	만을	107	0.09	96.89
으로의	26	0.05	99.36	에서도	163	0.08	98.38	54	이야	102	0.05	98.42	토록	102	0.08	96.98
이라고	24	0.05	99.40	와의	147	0.07	98.45	55	로의	93	0.05	98.47	요	102	0.08	97.06
만이	18	0.03	99.44	엔	139	0.07	98.52	56	치고	91	0.05	98.52	에선	102	0.08	97.15
와의	18	0.03	99.47	만을	124	0.06	98.58	57	만큼	91	0.05	98.56	로는	101	0.08	97.23
마다	18	0.03	99.51	부터는	122	0.06	98.64	58	으로써	90	0.05	98.61	과는	96	0.08	97.31
와는	18	0.03	99.54	만의	115	0.06	98.70	59	이며	88	0.05	98.65	로써	96	0.08	97.39
이라	16	0.03	99.57	야	110	0.06	98.76	60	만을	86	0.04	98.70	보다도	88	0.07	97.46
에만	16	0.03	99.60	까지의	107	0.05	98.81	61	에게는	85	0.04	98.74	마다	82	0.07	97.53
라고	16	0.03	99.63	과는	103	0.05	98.86	62	이라는	83	0.04	98.78	이라도	82	0.07	97.60
보다는	16	0.03	99.66	치고	103	0.05	98.91	63	라도	79	0.04	98.83	치고	78	0.06	97.66
처럼	12	0.02	99.69	과를	100	0.05	98.96	64	께서도	73	0.04	98.86	로의	77	0.06	97.73
부터	12	0.02	99.71	으로의	98	0.05	99.01	65	랑	69	0.04	98.90	과의	76	0.06	97.79
로도	10	0.02	99.73	까지는	84	0.04	99.06	66	으로부터	67	0.03	98.93	에게는	73	0.06	97.85
으로서의	10	0.02	99.75	보다는	76	0.04	99.09	67	이라도	67	0.03	98.97	보고	72	0.06	97.91
로서의	10	0.02	99.76	만이	72	0.04	99.13	68	과의	64	0.03	99.00	으로부터	72	0.06	97.97
만으로	8	0.02	99.78	에만	70	0.04	99.16	69	으로도	64	0.03	99.03	로서	69	0.06	98.03
토록	8	0.02	99.80	로의	69	0.03	99.20	70	만이	60	0.03	99.06	조차	67	0.06	98.08
총 107가지 52,718회 출현				총 197가지 199,906회 출현					총 219가지 195,279회 출현				총 245가지 120,825회 출현			

부록 2. 상위 100개 통합 어미의 상대적 출현빈도

	KT Data Set			신문 기사				국민학교 교과서				문학 작품		
ㄴ	12502	26.53	26.53	ㄴ	30037	17.79	17.79	1	어	30784	13.77	13.77	ㄴ	20300	16.06	16.06
어	5440	11.54	38.07	다	23353	13.43	31.22	2	ㄴ	24874	11.13	24.90	는	11866	9.39	25.44
는	4964	10.53	48.60	어	21818	12.55	43.77	3	는	17821	7.97	32.87	다	11258	8.90	34.35
다	4372	9.28	57.88	고	15254	8.77	52.54	4	고	13939	6.24	39.11	고	10065	7.96	42.31
ㄹ	2902	6.35	64.23	는	12599	7.25	59.79	5	ㄹ	13802	6.17	45.29	어	9036	7.15	49.45
ㄴ다	2942	6.24	70.47	ㄹ	11369	6.54	66.33	6	습니다	11098	4.97	50.25	ㄹ	6412	5.07	54.53
고	2900	6.15	76.63	은	2980	1.71	68.04	7	게	8342	3.73	53.98	게	3288	2.60	57.13
기	1462	3.10	79.73	기로	2911	1.67	69.71	8	다	7440	3.33	57.31	ㅂ니다	3245	2.57	59.69
게	830	1.76	81.49	게	2769	1.59	71.31	9	은	6398	2.86	60.17	ㄴ다	2933	2.32	62.01
은	794	1.68	83.17	다고	2691	1.55	72.85	10	지	5906	2.64	62.82	지	2927	2.32	64.33
어서	784	1.66	84.84	지	2690	1.55	74.40	11	ㅂ시다	5301	2.37	65.19	은	2788	2.21	66.53
지	632	1.34	86.18	기	2494	1.43	75.83	12	어서	4448	1.99	67.18	던	2149	1.70	68.23
며	514	1.09	87.27	ㄴ다	2411	1.39	77.22	13	ㅂ니다	4427	1.98	69.16	을	2118	1.68	69.91
으며	510	1.08	88.35	면	2389	1.37	78.59	14	지	3891	1.74	70.90	어서	2063	1.63	71.54
도록	472	1.00	89.35	으며	2106	1.21	79.81	15	기	3698	1.65	72.55	기	1971	1.56	73.10
러	402	0.85	90.21	라고	2076	1.19	81.00	16	을	3622	1.62	74.17	며	1930	1.53	74.63
나	388	0.82	91.03	나	1961	1.13	82.13	17	어라	3171	1.42	75.59	나	1787	1.41	76.04
는데	266	0.56	91.59	러	1925	1.11	83.23	18	ㄴ다	3054	1.37	76.96	면	1254	0.99	77.03
어라	250	0.53	92.12	으나	1704	0.98	84.21	19	니	3014	1.35	78.31	지만	1079	0.85	77.88
을	236	0.50	92.62	어라	1568	0.90	85.12	20	면	2342	1.05	79.36	는데	970	0.77	78.65
어야	228	0.48	93.11	며	1437	0.83	85.94	21	며	1991	0.89	80.25	다고	966	0.76	79.42
든	192	0.41	93.52	도록	1301	0.75	86.69	22	러	1840	0.82	81.07	야	877	0.69	80.11
진	186	0.39	93.91	던	1290	0.74	87.43	23	야	1732	0.77	81.84	으며	832	0.66	80.77
어서는	184	0.39	94.30	을	1195	0.69	88.12	24	어야	1690	0.76	82.60	이	794	0.63	81.40
므로	176	0.37	94.67	다는	1069	0.61	88.73	25	는지	1630	0.73	83.33	러	787	0.62	82.02
ㅁ	132	0.28	94.95	면서	979	0.56	89.30	26	오	1452	0.65	83.98	라고	739	0.58	82.60
이	122	0.26	95.21	ㅁ	956	0.55	89.85	27	던	1438	0.64	84.62	다는	733	0.58	83.18
면서	122	0.26	95.47	어야	930	0.54	90.39	28	ㅁ	1311	0.59	85.21	라는	666	0.53	83.71
면	122	0.26	95.73	는데	864	0.50	90.88	29	ㄴ가	1296	0.58	85.79	ㅁ	658	0.52	84.23
니라	110	0.23	95.96	어서	788	0.45	91.34	30	는가	1250	0.56	86.35	어야	614	0.49	84.71
고자	106	0.22	96.19	어서는	775	0.45	91.78	31	이	1084	0.48	86.83	습니다	599	0.44	85.16
지만	102	0.22	96.41	이	698	0.40	92.18	32	아라	1079	0.48	87.32	어도	499	0.39	85.55
다는	96	0.20	96.61	아	676	0.39	92.57	33	러	1051	0.47	87.79	든	498	0.39	85.95
지는	84	0.18	96.79	ㄴ다는	517	0.30	92.87	34	면서	968	0.43	88.22	니	480	0.38	86.33
소	76	0.16	96.95	러	495	0.28	93.16	35	라고	896	0.40	88.62	니지	432	0.34	86.67
거나	68	0.14	97.09	ㄴ데	484	0.28	93.43	36	ㄴ지	848	0.38	88.99	으나	422	0.33	87.00
는다	66	0.14	97.23	자	476	0.27	93.71	37	도록	747	0.33	89.33	ㄴ다는	420	0.33	87.33
기에	62	0.13	97.36	오	428	0.25	93.95	38	니	703	0.31	89.64	면서	411	0.33	87.66
으니	60	0.13	97.49	ㄴ다고	417	0.24	94.19	39	다고	697	0.31	89.95	습니까	396	0.31	87.97
으므로	60	0.13	97.62	거나	400	0.23	94.42	40	지만	685	0.31	90.26	러	394	0.31	88.28
려는	54	0.11	97.73	든	381	0.22	94.64	41	ㅂ니까	649	0.29	90.55	ㄴ다고	371	0.29	88.58
던	50	0.11	97.84	요	354	0.20	94.85	42	으면	606	0.27	90.82	는다	354	0.28	88.86
ㄴ다는	46	0.1	97.94	라는	338	0.19	95.04	43	어도	602	0.27	91.09	지	346	0.27	89.13
지를	42	0.09	98.03	소	336	0.19	95.23	44	습니까	591	0.26	91.35	ㄴ데	346	0.27	89.40
요	38	0.08	98.11	진	312	0.18	95.41	45	거나	526	0.24	91.59	란	329	0.26	89.66
질	38	0.08	98.19	련	302	0.17	95.59	46	ㄴ데	518	0.23	91.82	지는	313	0.25	89.91
어서도	34	0.07	98.26	지만	274	0.16	95.74	47	으며	512	0.23	92.05	진	306	0.24	90.15
미	32	0.07	98.33	어도	226	0.13	95.88	48	는데	507	0.23	92.28	고자	305	0.24	90.39
이	30	0.06	98.39	군	226	0.13	96.01	49	라고	454	0.20	92.48	기에	281	0.22	90.62
ㄴ지	30	0.06	98.46	지를	202	0.12	96.12	50	라는	442	0.20	92.68	어서는	273	0.22	90.83
라고	30	0.06	98.52	어서도	195	0.11	96.23	51	는다	439	0.20	92.87	으면	271	0.21	91.05
음	30	0.06	98.58	디	191	0.11	96.34	52	어요	406	0.18	93.05	ㄴ가	265	0.21	91.26
므로써	30	0.06	98.65	기에	189	0.11	96.45	53	래	406	0.18	93.24	다가	254	0.20	91.46
ㄴ데	30	0.06	98.71	나라	187	0.11	96.56	54	다는	387	0.17	93.41	긴	254	0.20	91.66
란	28	0.06	98.77	더라도	169	0.1	96.66	55	ㄹ라	357	0.16	93.57	도록	242	0.19	91.85
려	26	0.06	98.82	래	166	0.1	96.75	56	ㄹ까요	351	0.16	93.73	고는	236	0.19	92.04
라는	24	0.05	98.88	지는	162	0.09	96.85	57	나요	339	0.15	93.88	어라	201	0.16	92.19
어선	22	0.05	98.92	될	161	0.09	96.94	58	이서	331	0.15	94.03	다면	199	0.16	92.35
니이다	22	0.05	98.97	러	153	0.09	97.03	59	다가	330	0.15	94.17	듯	191	0.15	92.50
ㄹ수록	22	0.05	99.02	ㄹ라	152	0.09	97.11	60	러	324	0.14	94.32	ㄴ다면	187	0.15	92.65
어도	22	0.05	99.06	ㅂ니다	140	0.08	97.19	61	네	322	0.14	94.46	어요	169	0.13	92.78
다고	22	0.05	99.11	려는	140	0.08	97.27	62	돈	300	0.13	94.60	ㄹ지	165	0.13	92.92
기로	18	0.04	99.15	란	140	0.08	97.36	63	ㄴ다고	298	0.13	94.73	므로	165	0.13	93.05
건	16	0.03	99.18	ㄹ수록	138	0.08	97.43	64	려면	298	0.13	94.86	래	164	0.13	93.18
오	16	0.03	99.21	ㄴ데다	137	0.08	97.51	65	을까요	294	0.13	94.99	는지	163	0.13	93.30
어야만	16	0.03	99.25	ㄴ지	132	0.08	97.59	66	기로	291	0.13	95.12	군	160	0.13	93.43
ㄴ가	16	0.03	99.28	ㄴ가	131	0.08	97.66	67	진	289	0.13	95.25	거나	159	0.13	93.56
더라도	16	0.03	99.32	는데다	122	0.07	97.73	68	어서는	288	0.13	95.38	니라	156	0.12	93.68
으면서	14	0.03	99.35	매	121	0.07	97.80	69	ㄹ지	279	0.12	95.51	라고	151	0.12	93.80
는가를	14	0.03	99.38	음	110	0.06	97.87	70	요	276	0.12	95.63	어야만	150	0.12	93.92
는지를	14	0.03	99.41	는다	105	0.06	97.93	71	기에	272	0.12	95.75	으므로	145	0.11	94.03
면서도	14	0.03	99.44	어선	103	0.06	97.99	72	란	266	0.12	95.87	지요	145	0.11	94.15
ㄴ다면	12	0.03	99.46	다가	103	0.06	98.05	73	ㄴ지를	264	0.12	95.99	ㄹ까	144	0.11	94.26
고서	12	0.03	99.49	러며	97	0.06	98.10	74	구나	259	0.12	96.11	지도	142	0.11	94.37
이서	12	0.03	99.51	ㄴ다면	95	0.05	98.16	75	므로	233	0.10	96.21	길	141	0.11	94.49
려면	10	0.02	99.53	질	90	0.05	98.21	76	니라	232	0.10	96.31	기로	141	0.11	94.60
으면	10	0.02	99.55	지는	86	0.05	98.26	77	ㄴ다면	210	0.09	96.41	지를	141	0.11	94.71
려고	10	0.02	99.58	라면서	84	0.05	98.31	78	ㄹ까	209	0.09	96.50	러	139	0.11	94.82
이서만	10	0.02	99.60	건	80	0.05	98.35	79	디면	203	0.09	96.59	이서	135	0.11	94.93
래	8	0.02	99.61	ㄹ지	70	0.04	98.39	80	미	195	0.09	96.68	ㄹ러	133	0.11	95.03
고도	6	0.01	99.63	나	70	0.04	98.43	81	ㄴ다는	182	0.08	96.76	듯이	122	0.1	95.13
ㄴ다고	6	0.01	99.64	습니다	68	0.04	98.47	82	이야	173	0.08	96.84	음	121	0.1	95.22
ㄴ가를	6	0.01	99.65	으면	67	0.04	98.51	83	더니	165	0.07	96.91	건	120	0.09	95.32
러	6	0.01	99.66	이야	63	0.04	98.55	84	으나	162	0.07	96.98	랍니다	117	0.09	95.41
는가에	6	0.01	99.68	어다	62	0.04	98.58	85	어다	145	0.06	97.05	어다	115	0.09	95.50
는다는	6	0.01	99.69	다며	62	0.04	98.62	86	어서도	145	0.06	97.11	는가	109	0.09	95.59
라도	6	0.01	99.70	므로	61	0.04	98.65	87	지를	138	0.06	97.18	요	104	0.08	95.67
되	6	0.01	99.72	는데로	60	0.03	98.69	88	지는	130	0.06	97.23	ㅂ시오	104	0.08	95.75
지는	4	0.01	99.72	고는	58	0.03	98.72	89	으니	130	0.06	97.29	죠	101	0.08	95.83
나	4	0.01	99.73	는가	58	0.03	98.75	90	긴	127	0.06	97.35	미	98	0.08	95.91
디면	4	0.01	99.74	고서는	56	0.03	98.79	91	음까	127	0.06	97.41	리면	98	0.08	95.99
ㄹ만한	4	0.01	99.75	세	54	0.03	98.82	92	단다	124	0.06	97.46	더라도	97	0.08	96.06
니	4	0.01	99.76	는데도	53	0.03	98.85	93	세요	121	0.05	97.52	으니	96	0.08	96.14
지도	4	0.01	99.77	다면서	53	0.03	98.88	94	고는	118	0.05	97.57	리도	95	0.08	96.21
네	4	0.01	99.78	음에도	52	0.03	98.91	95	을	118	0.05	97.62	아야	94	0.07	96.29
ㄴ지를	4	0.01	99.78	면서도	51	0.03	98.94	96	니까	118	0.05	97.67	지면	93	0.07	96.36
으리라	4	0.01	99.79	는지	44	0.03	98.96	97	려는	115	0.05	97.73	음까	93	0.07	96.44
을수록	4	0.01	99.80	지고	44	0.03	98.99	98	질	114	0.05	97.78	려는	90	0.07	96.51
구나	4	0.01	99.81	ㄴ가를	44	0.03	99.01	99	지요	109	0.05	97.83	고서	85	0.07	96.57
음에도	4	0.01	99.82	다면	41	0.02	99.04	100	으면서	97	0.04	97.87	네	85	0.07	96.64
총 134가지 47,128회 출현				총 299가지 173,886회 출현					총 400가지 223,517회 출현				총 432가지 126,430회 출현			