자연언어처리용 전자사전을 위한 한국어 기본어휘 선정

배희숙, 이주호, 시정곤, 최기선*
한국과학기술원

Hee-Sook Bae, Juho Lee, Chung-Kon Shi and Key-Sun Choi. 2003. Selection of Korean General Vocabulary for Machine Readable Dictionaries. *Language and Information 7.1*, 41–54. According to Jeong Ho-Seong (1999), Koreans use an average of only 20% of the 508,771 entries of the Korean standard unabridged dictionary. To establish MRD for natural language processing, it is necessary to select Korean lexical units that are used frequently and are considered as basic words. In this study, this selection process is done semi-automatically using the KAIST large corpus. Among about 220,000 morphemes extracted from the corpus of 40,000,000 eojeols, 50,637 morphemes (54,797 senses) are selected. In addition, the coverage of these morphemes in various texts is examined with two sub-corpora of different styles. The total coverage is 91.21 % in formal style and 93.24 % in informal style. The coverage of 6,130 first degree morphemes is 73.64% and 81.45%, respectively. (**Korea Advanced Institute of Science and Technology**)

Key words: 기본 어휘(general vocabulary), 반자동 선정(semi-automatic selection), 기계 가독형 전자사전(MRD)

1. 서론

자연언어처리에는 명사사전, 동사사전, 전문용어사전, 고유명사사전, 복합명사사전, 개념사전, 의미관계사전 등 다양한 전자사전이 필요하다. 이러한 전자사전을 구축할 때, 제일 먼저 수행하는 것이 사전의 항목 선정이다. 언어사전[1]에서 다루어진 모든 어휘[2]를 기계용 전자사전 항목으로 선정하여 처리하는 것은 사전의 구축 시간과 활용성을 고려할 때 모두 현실적이지 않다. 또한 언어사전이 가능한 한 많은 어휘를 수록하고 처리하려는 데에 반해, 전자사전에서는 실제 사

* 305-701 대전광역시 유성구 구성동 373-1 한국과학기술원 전문용언언어공학연구센터, E-mail: elle@world.kaist.ac.kr

[1] 여기서 언어사전이란 백과사전에 대립되는 의미로 사용하였다.

[2] 본 연구에서 기본어휘라 함은 기본형태소를 말한다. 카이스트 대용량 코퍼스를 카이스트 형태소분석기(http://morph.kaist.ac.kr/~morph)로 자동분석하고, 그 결과로 빈도를 갖춘 서로 다른 형태소 목록에 대하여 한국어에서 기본어로 간주되는 형태소를 선정한 작업이므로 단위는 형태소이다. 그러나 국어학적 기준을 엄격하게 적용한 형태소가 아니라 어휘의 뜻을 살리기 위하여 결합 형태까지 포괄하는 넓은 의미의 형태소이다. 또한 본 연구에서 선정한 기본 형태소는 보완을 통하여 기본어휘로 정리될 것이다.

용되지도 않은 어휘들을 매번 함께 다루는 것이 시스템의 과부하를 낳는 중요한 요소가 될 수 있다. 구체적인 연구 결과를 예로 들어 보자. 김광해(1999)에 의하면, "한국어 사전에서 실제 사용되는 어휘는 대략 10-15 %에 불과하다". 정호성(1999)[3]의 조사에 따르면, 『표준국어대사전』의 경우 총 508,771 개 항목 중에서 실제 사용되는 어휘는 약 20% 정도라고 한다. 따라서 기본적으로 사용되는 어휘를 선정하여 전자사전의 항목으로 사용하고 점차적으로 항목의 범위를 확장하는 것이 효과적이다.

지금까지 한국어 교육용 기초어휘 혹은 기본어휘 조사로 약 10여 건의 연구가 있었고(표 2 참조), 현재 국립국어연구원의 한국어 기본어휘 선정 작업이 진행 중에 있다. 국립국어연구원의 기본어휘 선정은 이제 코퍼스의 빈도순 목록이 완성된 단계여서 앞으로도 4년 정도의 시간이 더 필요할 것으로 예상된다. 지금까지의 연구 결과들은 모두 한국어 교육용 어휘 선정을 목적으로 한 것으로 전자사전 항목으로 삼기에는 지나치게 적거나 지나치게 많다.

본 연구에서는 대용량 코퍼스를 기반으로 어휘의 빈도를 고려하고, 아울러 기존의 기초어휘 조사 결과와 비교하면서 50,000[4] 내외의 기계 가독형 전자사전 항목을 위한 기본어휘를 선정하고자 한다. 그러나 본 연구는 아직 많은 보완 작업을 필요로 하는 한국어 기본어휘 선정의 첫 발자국에 불과하다고 하겠다.

2. 기본어휘의 정의

한 국가의 언어를 말하고 이해하는 데에 필수적인 기초어휘와 자주 사용되는 고빈도 어휘를 통틀어 기본어휘라고 할 수 있다. 기본어휘는 'fundamental', 'basic' 혹은 'general'과 같은 다양한 형용사로 표현되었고, 이들 형용사의 의미 구분에 대한 의견도 일치되지 않았다. 그런데 최근 ELDA(www.fr.elda)[5]를 통해 제시되는 언어자원 목록에 따르면 기본어휘라고 부르는 어휘군이 일괄적으로 'general vocabulary'[6]로 표현되는 것을 확인할 수 있다.

심리언어학적 관점[7]에서 기본어휘란, 어린아이가 언어를 습득할 때 여러 어

[3] 김광해(2001)에서 재인용.

[4] 약 50,000이라는 수를 상정하는 것은 7,000 만 어절 카이스트 대용량 코퍼스에 대한 간단한 조사 결과에 근거한다. 2001년 코퍼스 재정비가 이루어지기 이전의 7,000 만 어절 카이스트 코퍼스로부터 세 개의 부분코퍼스를 구성하고, 이를 기계적으로 분석하여 각 부분코퍼스의 95% 정도를 점유하는 형태소의 수를 계산하였다. 그 결과로 약 50,000 개 정도의 서로 다른 형태소가 코퍼스를 구성하는 총 단어수의 95-98% 정도를 커버한다는 사실을 파악하게 되었다. 이는 엄밀하게 통계언어학적 기준을 적용한 조사라기보다는 코퍼스의 95% 정도를 몇 개의 서로 다른 형태소가 점유하는지에 대한 아이디어를 얻기 위하여 기계적으로 행해진 조사였음을 밝힌다.

[5] ELRA는 유럽언어자원협회(European Language Resources Association)이고, ELDA (European Language resources Distribution Agency)는 ELRA의 전략과 계획을 발전 실행하기 위한 센터이다. 21세기 중요한 시장을 형성할 언어자료 제공자와 이용자 간의 관계를 단순화 하는 것을 목적으로 ELDA는 언어자원 제공자와 사용자를 연결해 주는 역할을 한다. 현재 유럽 언어자원 센터의 구심점이라 할 수 있다.

[6] Ch. Muller(1987)에 따르면, "vocabulary"는 "lexicon"에 대립된다. 즉, 파롤과 랑그의 대립과 같이 둘 다 어휘 혹은 어휘집이란 뜻이지만 후자가 랑그와 마찬가지로 잠재적이고 추상적 개념이라면 전자는 실현된 어휘집으로 구체적이다. 예를 들어 Proust의 어휘 연구라고 할 때 '어휘'는 lexicon이라기 보다는 vocabulary인 것이다. 물론 이 vocabulary가 좀더 총체적 의미를 부여 받으며 lexicon이 될 수 있다. Lexicon의 원소가 lexeme 혹은 lemma라면 vocabulary의 원소는 type이며 word form이다. 이러한 Muller의 용어 정리는 계량언어학계에서의 기본지식이며, 본 논문에서 이를 소개하는 것은 vocabulary와 lexicon의 차이에 대하여 짚어 봄으로써 기본어휘의 vocabulary가 lexicon에 비하여 갖는 특성을 밝히기 위함이다.

[7] 이민행, 문미선, 신효식 옮김 (1996) 『새로운 의미론』. 한국문화사. [원저: Scharz, M/ Chur, J. (1993) Semantik: Ein Arbeitsbuch. Tubingen: Narr]

휘 중에서도 우선적으로 습득하는 어휘들을 가리킨다. 외국어 교수법에서는 외국어를 가르치는 데 있어서 기본적으로 알아야 하는 어휘를 기본어휘라 한다. 계량언어학적으로 한국어 텍스트를 분석하면 많은 경우 텍스트를 구성하는 서로 다른 형태소의 약 20%가 코퍼스 전체를 이루는 형태소 수의 95-98%를 커버한다.[8] 이와 같이 다양한 관점에서의 기본어휘 정의를 종합하면 기초적 어휘와 고빈도 어휘의 합집합이라 할 수 있다. 이 점에서 자연언어처리용 기본어휘와 한국어 교육용 기본어휘는 분명히 유사하다. 그러나 자연언어처리용 기본어휘란 일상적이고 기초적인 생활 어휘를 포괄하면서도 전문적 문서에 자주 나타나는 표현을 처리할 수 있도록 어느 정도 높은 수준의 어휘들까지 고려되어야 한다. 따라서 본 글에서의 기본어휘란 한국어를 말하고 이해하는 데 있어서 가장 기초가 되면서 자주 쓰이는 어휘의 모음이며, 동시에 한국어 습득용 어휘보다는 각 전문분야의 가장 기초적인 어휘까지 포괄하는 수준의 어휘 모음이라고 정의하고자 한다.

3. 어휘 선정의 문제점 및 전략

어떤 어휘를 어떤 과정을 통하여 선정할 것인가? 코퍼스를 기반으로 고빈도어를 추출하여 얻은 결과를 기본어휘로 간주할 수는 없다. 그 이유는 첫째 계량언어학적 측면에서 완벽하게 균형 잡힌 코퍼스를 구성한다는 것이 어렵고, 고빈도로 나타나지만 기본어휘로 보기 어렵거나 저빈도로 쓰이더라도 한국어를 사용하는 데 있어서 기본적인 어휘로 보아야 할 것들도 있기 때문이다.[9] 둘째, 사회언어학적 측면에서 어휘는 끊임없이 변화한다. 특히 인터넷을 통해 각국의 정보가 활발히 교류되는 최근에는 어휘가 생기고 유행하고 다시 사라지는 현상이 매우 급격하게 이루어지는 만큼 기본적 어휘에 포함될 단어들이 빨리 변하여 어휘 선정이 더욱 어렵다. 사실, 어휘집이란 이상적 목록일 뿐이다. 어휘의 역동성을 생각하면, 기본어휘 목록 또한 지속적으로 보완되어야 할 과정상의 목록이 된다. 어찌 되었건 이 과정적 목록을 선정하기 위하여 사전에 들어 있는 어휘 목록이나 실제 어휘들이 사용된 코퍼스를 이용해야 한다. 그렇다면, 기본어휘 선정의 문제는 적절한 균형 코퍼스 구성과 어휘 선정 과정의 객관성 확보로 좁혀진다.

균형 코퍼스를 위한 절대적인 비율은 없다. 연구의 목적에 따라 유형별 가중치가 달라진다. 본 연구에서는 자연언어처리용 기본어휘 선정이 목적이므로 다양한 전문분야 텍스트가 포함된 코퍼스를 기반으로 하였다. 또한 기초어휘를 제대로 선정할 수 있도록 텍스트의 다양성을 많이 고려하였으며 고빈도 어휘의 정당성을 높이기 위하여 4,000만 어절 대용량 코퍼스를 이용하였다. 어휘 선정 과정의 객관성 확보를 위해서는 동일한 어휘를 난이도에 따라 여러 사람이 분류하고, 그 결과에 대한 조정이 필요하다.

3.1 고빈도 어휘

고빈도 어휘는 기본어휘 선정을 위한 중요한 정보로 고정성이 높은 어휘에 대한 정보를 제공한다. 그러나 코퍼스의 완벽한 균형성이 보장될 수 없으므로 고빈도어 추출에 더하여 저빈도이지만 기초적인 어휘들을 보완하여야 한다. 예를 들면,

[8] 텍스트 조사에 대한 과정과 결과는 배희숙 외(2001)를 참조하라.

[9] 비슷한 빈도로 출현한 어휘라도 아래의 예와 같이 기본적인 어휘로 간주되는 정도가 같지 않다. 저빈도이면서도 중요한 어휘로 선정되는 경우가 있었다. 구체적인 예로 '마당놀이', '쌀농사', '벼농사', '책상보', '저녁별', '저울판' 등이 있으며, 이러한 예 중에서 의외로 여겨지는 어휘로는 '전나무', '점검', '향상', '적녹색', '처분', '골뱅이', '해석', '채색화', '꺾쇠', '기본' 등이 있었다. 여기서 '향상', '처분', '해석'이 낮은 빈도를 보인 것은 코퍼스에서 동사로 쓰이지 않고 명사로 쓰인 경우에 해당하기 때문일 것이다.

4,000만 어절 카이스트 코퍼스에 비서술성 명사로 쓰인 '점검', '향상', '처분', '해석', '기본' 등은 단 세 번밖에 나타나지 않는다. 또한 '골뱅이', '꺽쇠' 등도 단 세 번밖에 나타나지 않는다. 그러나 이들 모두 한국어를 이해하기 위해 중요한 기본어휘로 간주된다.

3.2 동형어와 다의어

수천만 어절의 의미 분석된 코퍼스를 갖추기는 현실적으로 매우 어렵다. 본 기본어휘 선정을 위하여 사용한 코퍼스는 의미 분석된 코퍼스가 아니므로 동형어와 다의어 처리의 문제를 그대로 안고 있다. 자동 처리 과정에서 품사가 다른 경우 동형어가 자동으로 분리되지만 품사까지 같은 동형어의 경우는 『우리말 큰사전』에 의거하여 수동으로 조정하였다. 물론 동형어 중에서 기본어휘로 간주될 수 없다고 판단하면 처리 대상에서 제외하였다. 다의어 또한 위 사전에 입각하여 구분하였다. 그러나 이 구분은 코퍼스에서 실제 사용된 의미별 처리가 아니라 사전에 의거한 의미 구분이었다.

3.3 등급분류

기본어휘 선정을 위하여 본 연구에서는 각 어휘의 의미를 난이도에 따라 1등급부터 5등급까지 분류하였다.[10] 이 과정이 엄격하게 객관적이어야 한다. 그러나 주관성을 완전히 배제할 수는 없어 두 명의 작업자가 동일한 형태소에 대하여 등급을 매긴 후 서로 결과를 바꾸어 조정하는 방식을 취하고 후처리를 통하여 애매한 경우들을 모아 토론과 조정을 거쳤다.

4. 기본어휘 선정 과정

7,000만 어절 카이스트 코퍼스(www.kibs.kaist.ac.kr)는 2001년 KDC 도서 분류 번호를 갖춘 서지정보와 텍스트 내부의 오류 등을 전면적으로 수정 보완하고 균형 코퍼스를 자유롭게 구성할 수 있도록 분야별로 재구성하여 4,000만 어절로 정리되었다.[11] 이를 카이스트 형태소분석기에 의해 기계 분석한 후 빈도순 형태소 목록[12]을 얻었다. 이 목록에서 적어도 세 번 이상 나타나는 형태소만 고르면 228,574에 이른다. 자료 검토 결과 괄호를 비롯한 구두점이 포함되거나 잘못 분할된 자동분석의 오류 대부분이 빈도 1과 2에 몰려 있어 빈도 3부터의 형태소를 대상으로 작업을 수행하고 후처리에서 빈도 1과 2의 형태소 목록 중에서 올바른

[10] 인간이 사물을 판정하는 어림수는 일곱 가지를 넘지 않는다는 것이 심리학자들의 실험결과이다(http://www.g-matrix.pe.kr/feature/ai/fuzzy4.htm). 이것을 <매직세븐>이라 하며 그 어림수를 NEGATIVE LARGE, NEGATIVE MEDIUM, NEGATIVE SMALL, ZERO, POSITIVE SMALL, POSITIVE MEDIUM, POSITIVE LARGE로 분류한다. 어떤 대상을 인지할 때 위에서 이 일곱 가지 수준 중의 하나로 나타낼 수 있으며 대상이 단순하다면 일곱 가지를 다 사용하지 않고 다섯 개나 세 개를 사용하게 된다. 이러한 이론으로부터 어휘의 난이도 수준을 5등급으로 나누어 분류하였다.

[11] 재정비 이후 카이스트 코퍼스는 크게 [학술], [교양], [문학], [전문], [언론], [종교], [교육], [실용]으로 대분류되었다. 각각의 대분류 파일들은 다시 분야별로 세분류되었다. 한편, 카이스트 코퍼스는 양적으로 문학 분야에 집중되어 있는 면이 있어, 본 코퍼스의 정당성에 대한 확인을 위하여 프랑스어 기본어휘 선정에 참여하였던 INRIA의 연구원인 Ch. Bernet로부터 서신 조언을 받았다. 프랑스어 기초어휘 선정을 위하여 "20명의 전문가가 코퍼스 검수에 참여하였으며, 전체 코퍼스의 80%가 문학 작품이었다"고 한다.

[12] 카이스트 형태소분석 사이트 http://morph.kaist.ac.kr/~morph/에서 누구나 원하는 텍스트를 올리고 자동 분석 서비스를 받을 수 있다. 본 형태소 분석기에 대한 성능평가나 자세한 설명은 강영수(2002)와 안효은(2002)을 참조하라. 태그셋은 본 논문의 #부록으로 첨부하였다.

형태의 형태소만 기계적으로 찾아 보완하는 것이 효율적이라고 판단하였다.

4.1 웹 기반 작업 환경[13]

웹 기반 작업 환경을 설정하고, 빈도 3 이상의 228,574 개의 서로 다른 형태소에 대해 기존의 자료와 비교하면서 난이도에 따라 전문가[14]들이 수동으로 형태소를 등급별로 분류하였다. 이는 여러 명의 작업자가 컴퓨터로 같은 자료를 정확하고 신속하게 처리할 수 있도록 하기 위한 것이다. 작업자들은 각자 형태소에 대하여 등급을 부여하고 복합어 여부를 표시하였는데, 이는 숫자를 써 넣는 방식이 아니라 선택만 할 수 있도록 하였다. 기타 동형어나 다의어에 관하여 기록해야 할 사항은 비고란을 이용하였다. 이러한 환경은 속도를 빠르게 할 뿐 아니라 작업상의 오류도 줄일 수 있게 해준다.

일련번호	형태소	품사	출현횟수	상대빈도(%)	등급
524944	못벌	ncn	3	0.000003832001625484	오류
615888	불표	ncn	3	0.000003832001625484	오류
926954	저울판	ncn	3	0.000003832001625484	2
1085209	평판	ncpa	311	0.000397250835175174	3

[표 1] 테이블형 웹기반 작업환경

표 1에서 출현 횟수는 코퍼스에 해당 형태소가 나타난 경우의 수이며 상대빈도란 각 형태소의 출현 횟수를 전체 형태소 수(78,288,067)로 나눈 값이다. 등급 열에는 해당 등급이 표시되는데, '못벌'이나 '블표'와 같은 오류는 이 과정에서 오류로 표시되었기 때문에 제외시켰다.

4.1.1 비교 목록. 기본어휘 선정을 위해 외부 어휘 목록과 코텀 어휘 목록을 비교하였다. 표 2의 목록을 어휘 선정 과정에서 각 형태소와 비교한 것이다. 예를 들어 모든 비교 목록에서 기초어휘나 기본어휘로 선정되어 있으면, 작업자는 해당 형태소를 기본어휘로 선정할 가능성이 높아진다.

4.2 어휘 등급

기본어휘를 선정하기 위하여 각 형태소를 1에서 5까지 등급을 부여하며 구분하였다. 작업자는 정확하게 해당 형태소가 어느 등급에 속하는지 알 수 없다. 때문에 몇 가지 기준을 적용하면서 해당 형태소가 어느 등급으로 기우는지 다섯 단계로 나누어 판단을 하게 되고, 이를 여러 사람이 기록하였다.

4.2.1 어휘 등급 기준. 어휘의 난이도를 다섯 단계로 분류하기 위하여 카이스트 코퍼스의 빈도와 다른 기초어휘 조사 결과를 비교하여 어휘의 등급을 부여하였다. 등급 기준으로는 초등학교 저학년까지 1등급, 중학교 보통 수준까지 2등급, 고등학교 보통 수준까지 3등급, 고등학교 상위 수준과 교양 있는 성인 수준까지를 4등급으로 정하였다. 5등급은 사전에 존재하나 매우 전문적 수준의 어휘에 해

[13] 본 연구를 위한 웹기반 작업환경과 각종 언어자료의 데이터베이스화는 은광희에 의하여 이루어졌으며 이 자리를 빌려 감사를 표하는 바이다.

[14] 여기서 전문가라 함은 카이스트 전문용어언어공학연구센터의 전자사전 팀 연구원을 말하며, 이 연구에 참여한 연구원은 최소 3년 최대 7년 동안 한국어 전자사전에 대한 연구 경력을 갖고 있었다.

연구자	연구기관	자료	어휘 수	연도
이응백 외	국어연구소	국민학교 교육용 어휘	1,556	1987
홍재성 외	서울대학교	21세기 세종계획 전자사전 개발 자료	63,919	1999
김홍규	고려대학교	국어사전 표제어 자료	82,940	2000
서상규 외	연세대학교	한국어 교육 기초어휘 의미빈도 사전의 개발	1,087	2000
김광해	서울대학교	한국어 등급별 총어휘 선정	1,000 (견본)	2001
김홍규/강범모	고려대학교	한국어 형태소 및 어휘 사용 빈도의 분석	150만 어절의 구성 형태소	2001
최기선 외	한국과학기술원	코텀 고빈도 어휘 목록	50,000[15]	2001

[표 2] 자료 목록

당한다. 이러한 등급 기준은 전적으로 임의적이며, 주관적 작업이지만 작업자[16] 교육을 통하여 기본 지침 적용에 일관성을 부여하도록 하였다.

4.2.2 이해도와 사용도. 어휘의 이해도와 사용도가 항상 일치하는 것은 아니다. 어린이가 해당 어휘의 의미를 이해하면서도 실제 사용은 못하는 경우가 있고, 실제로 많이 사용하지만 의미를 명확히 알지는 못하는 경우도 있다. 이러한 경우, 무엇을 우선적으로 하여 등급을 결정할 것인가? 대부분의 경우 어휘의 뜻을 아는 것이 우선이고, 익숙해진 다음에 사용할 줄 안다. 외국어를 배울 때, 단어의 뜻을 알고 있으나, 실제 주어진 환경에서 사용하지 않는 어휘들이 상당수이다. 의미와 활용에 많이 익숙해진 다음에서야 그 어휘를 제대로 활용할 줄 알게 되는 것이다. 하지만 본 작업을 수행할 때에는 어휘의 의미를 정확히 알고 있으면 해당 수준에 맞는 것으로 처리하였다. 이러한 결정을 위하여 한편으로는 각 등급의 기준이 되는 연령별 모국어 화자에게 사전 지식 없이 어휘를 제시하고 의미를 이해하는지 조사하고, 또 한편으로는 해당 어휘를 포함하여 적절한 문장을 구사하는지 보아 어휘의 등급을 판단하였다. 원칙적으로는 적합한 피험자 수를 확보하여 통계학적으로 올바른 테스트를 함으로써 등급상의 객관성을 확보해야 하나 등급에 대한 판단이 애매한 경우에 한하여 위와 같은 간단한 절차를 거쳤다.[17]

4.3 선정 지침[18]

4.3.1 오표기. 코퍼스를 카이스트 형태소 분석기로 자동 분석한 후 처리하였다. 코퍼스를 구성하는 텍스트가 원래 가지고 있는 오표기를 비롯하여 형태소분석기 오류나 구두점으로 인한 오류 등 상당수의 오류가 있다. 그러나 오표기가 나타나는 경우, 올바른 표기로 같은 형태소가 있는지 확인한 후 있으면 제외시키고, 없으면 옳은 표기의 형태소를 추가하였다.

[16] 형태소에 등급을 부여한 작업자는 2001년부터 2002년 초반까지 카이스트 코텀의 한국어 전자사전 팀 연구원들로 전공은 언어학이고 학력은 평균 석사학위 소유자이다.

[17] 전문가의 직관에 의존하였음을 인정하며 이 부분은 보완되어야 할 사항이다.

[18] 어휘 선정 지침에 대하여 본 논문에서는 간략한 설명만을 하고 지나가고자 한다. 이에 대한 좀 더 체계적인 연구는 가칭 「기본어휘 선정지침에 대한 연구」로 따로 준비하고 있음을 밝힌다.

4.3.2 개체명. 코퍼스에 쓰인 수많은 인명, 지명, 조직명, 제품명, 단위명 등의 개체명을 기본어휘로 간주할 수는 없다. 원칙적으로 개체명을 기본어휘에서 제외하였다. 다만 '김치'와 같이 자주 사용되는 음식 이름이나 조직명 등은 예외로 하였다.

4.3.3 의성어, 의태어, 외래어. 다양한 문학작품 및 구어체 텍스트, 신문이나 매뉴얼 등을 포함하는 실제 코퍼스를 기반으로 이루어진 작업으므로 상당수의 의성어, 의태어, 외래어 등이 나타났다. 원칙적으로 이들 의성어, 의태어, 외래어를 기본어휘로 간주하지 않았다. 그러나 고빈도이고 뜻풀이가 용이하면 등급을 부여하였다.

4.3.4 복합어와 파생어. 완전히 자율적 단어의 결합이 전혀 새로운 의미를 만들지 않을 때, 다른 형태소의 삽입이 자유로울 때에는 일시적 결합으로 간주하고 각각 처리였다. 그러나 고유한 의미를 만들어 내거나 (의미론적 관점), 다른 어휘의 삽입이 자유롭지 않거나 구성소의 위치 변환이 불가능할 때 (통사론적 관점), 또는 파생어가 많이 생산할 때 (형태론적 관점), 그 자체로 관용적 표현으로 활용될 때(usage)에는 어휘화된 복합어로 간주하였다.

그러나 이러한 일반적인 원칙으로 복합어와 파생어 문제를 해결할 수는 없다. 복합어와 파생어 문제는 기계번역이나 정보검색 등의 자연언어처리에서 핵심 문제 중 하나이다. 이 오랜 문제를 해결하기 위해 언어학적 원칙을 적용하고 기존 언어 사전들을 비교하고 코텀과 한국전자통신연구원(ETRI)의 접사 목록을 비교 분석하였으나 일관성 있는 원칙을 세우기에는 부족하였다. 가장 최근에 많은 언어학자들이 함께 구축한 국립국어연구원의 『표준국어대사전』에 전적으로 의거하여 복합어와 파생어를 처리하기로 한 것은 이러한 이유에서이다. 『표준국어대사전』에 의거하여 파생어와 복합어를 다룬다 하여도 이들의 등급 결정에는 여전히 어려움이 존재한다. 단일어로서의 어휘는 잘 쓰이는데 이 단일어들이 결합하여 복합어로 쓰이면 기초적인 어휘로 인식되지 않는 경우가 있다. 거꾸로 복합어로는 잘 쓰이는데 그 구성소인 단일어들은 잘 쓰이지 않는 경우가 있다. 이러한 경우, 사용 정도에 따라 등급을 부여하였다.

4.3.5 동형어와 다의어. 동형어와 다의어 구분은 한글학회 『우리말 큰 사전』을 기준으로 하였다. 현재 의미 분화가 이루어져 다의어로 사용되고 있지만 『우리말 큰 사전』에서 이를 구분하지 않은 경우, 국립국어연구원의 『표준국어대사전』을 참고하여 구분하였다. 또한 해당 형태소는 기본어휘의 구성소로 간주되나 그에 대한 동형어나 다의어는 난이도가 높은 경우, 혹은 드물게 사용되는 경우 이를 등급 부여 작업에서 제외시켰다.

동형어와 다의어 처리 과정에서 카이스트 태그셋과 사전의 품사 체계가 일치하지 않아 조정이 필요하였다. 실제 코퍼스에서 사용된 형태소의 의미와 사전의 의미 분류가 일치하지 않아 발생하는 문제도 있었다. 품사 체계의 경우, 54 개로 구성된 카이스트 태그셋에 28 개로 분류된 『우리말 큰 사전』의 품사 체계를 맞추었고[19], 의미 분류는 전적으로 『우리말 큰 사전』에 의거하여 이루어졌다. 『우

[19] 품사체계 통합 과정에서 지정사와 서술격조사처럼 명칭은 다르나 내용이 같은 경우 문제 없이 두 품사체계의 매칭이 이루어진다. 그러나 동일한 문법적 기능을 지시하면서 명칭이 다른 경우 카이스트 태그셋을 기준으로 하여 『우리말 큰 사전』을 맞추었다. 『우리말 큰 사전』이 카이스트 태그셋 보다 더 자세히 분류한 자동사, 타동사, 피동사, 사동사, 불완전 타동사, 불완전 자동사와 같은 구분은 존중할 수 없었다.

리말 큰 사전』의 의미는 표제어, 동형어, 다의어 대분류, 다의어 세 분류로 구분되어 기호화 되었다. 그러나 앞서 설명한 바와 같이 코퍼스에서 사용된 의미가 사전에서의 의미보다 세분화 되어야 할 때에는 『우리말 큰 사전』의 최하위 세분류에서 라인을 추가하여 의미 분화를 표시하였다. 반대로 어휘 사용면에서 의미가 통합되어야 하는 경우, 가장 상위의 의미 분류 번호로 통합하였다.

4.4 결과 정리

기본어휘 선정을 위한 작업 대상은 비고란에 기록된 형태소를 포함하여 등급이 부여된 모든 형태소이다. 원칙적으로 이 중에서 1등급부터 3등급까지의 형태소를 기본어휘로 간주한다. 비고란에는 제외되어야 할 어휘, 오태깅된 어휘, 국어사전에 등재되어 있지 않은 어휘 등의 정보와 의미 구분된 다의어의 등급 정보가 있다.

형태소	품사	빈도	상대빈도(%)	등급	비고	추가
게시문	ncn	3	0.00000383200162548	3		
검사지	ncn	3	0.00000383200162548	3		
잔영	ncn	39	0.00004981602113129	4		
부지런	ncps	338	0.00043173884980453	1		
기름지	paa	331	0.00042279751267840	3		

[표 3] 형태소에 대한 정보

5. 후처리

웹 환경에서 전문가들이 작업한 결과를 정리하는 과정에서 복합어 처리, 'ㄱ', 'ㄴ'과 같은 낱자 처리 등의 일관성 있는 처리를 위하여 문제시 되는 자료의 유형을 파악하고 각 유형에 해당하는 형태소 목록을 분류하였다. 예를 들어, 등급 부여 과정에서 추가 설명이 필요하거나 동형어 구분이 필요한 경우 비고란에 그 내용을 기록하였다. 주지하는 바와 같이 기계적 처리에서는 점 하나의 차이로도 오류가 발생하므로 기록 형식이 작업자에 따라 미세한 것이나마 차이가 있는 경우 어려움이 있었다. 비고란 정리 및 품사 수정은 기계적으로 가능하나 복합어 처리 및 동형어 정리는 섬세한 판단을 요하므로 정확성을 기하기 위하여 수동으로 정리하였다.

5.1 후처리 내용

비고란에서의 정보를 충실히 반영하기 위하여 비고란에 아무 것도 없이 등급만 부여된 형태소들, 비고란에 기록이 있으면서 동시에 등급이 부여된 항목들, 비고란에 기록이 있지만 등급 표시는 안 된 형태소들을 따로 분리하여 일관성 있는 형식으로 정리하였다. 표기가 잘못 되었거나 품사가 올바르지 않은 경우를 지적하는 경우 올바른 표기와 품사를 부여하고 이에 맞는 등급을 부여하였다.

복합어 처리를 위하여 3음절 이상의 복합형 형태소를 추출하였다. 복합어는 『표준국어대사전』에 의거하여 처리하였으므로, 이들 복합 형태소 역시 『표준국어대사전』에 등재되어 있으면 그대로의 형태에 대하여 등급을 부여하고, 그렇지 않으면 분할하여 각각의 단일 형태소에 등급을 부여하였다.

같은 의미의 형태소를 두 사람 이상이 처리한 경우 일괄적으로 정리하였다. 이는 같은 품사이면서 다른 의미를 지닌 형태소 처리 과정에서 발생할 수 있는 것으로, 등급이 똑같이 부여되어 있으면 하나만 남기고 나머지는 삭제하였다. 또한 등급이 달리 부여되어 있으면 판단이 애매한 경우로 간주하여 지침에 따라 이해도와 사용도를 조사하고 토론 과정을 거쳤다.

동형어의 경우 뜻풀이를 모두 기록하였다. 아울러 품사가 잘못 부여되어 있는 형태소도 바르게 수정하였다.

5.2 통계

이러한 재정리 과정을 통해 획득한 최종 결과에 대한 통계는 다음과 같다. 전체 의미수는 54,797개이다. 선정된 기본어휘에서 가장 빈도가 높은 30개 형태소를 나열하면 표 4[20]와 같다.

형태소	품사	등급	상대빈도 (%)	형태소	품사	등급	상대빈도 (%)
을	jco	1	2.32357889229785	가	jcs	1	0.991877855408028
의	jcm	1	2.13264046997099	것	nbn	1	0.927744198870053
다	ef	1	2.10534001305716	있	paa	1	0.782485790586706
ㄴ	etm	1	2.03200699795027	ㄹ	etm	1	0.737303936754499
이	jp	1	1.88070935510517	하	pvg	1	0.727517004603013
에	jca	1	1.82753777788382	들	xsn	1	0.726446598815628
었	ep	1	1.75837525787934	으로	jca	1	0.703628306469746
하	xsv	1	1.73840031073957	도	jxc	1	0.671640034234081
이	jcs	1	1.5324404420408	에서	jca	1	0.600368636001704
는	etm	1	1.4516720664466	하	xsm	1	0.563751305802454
는	jxc	1	1.40288813108644	적	xsn	1	0.530654307763149
를	jco	1	1.34681317396686	적	xsn	3	0.518602662651001
어	ecx	1	1.27006456807779	되	xsv	1	0.462153446705997
은	jxc	1	1.2660218063629	로	jca	1	0.456788644430319
고	ecx	1	1.24095157439511	고	ecc	1	0.451357420793133

[표 4] 고빈도 기본어휘

표 4는 기본어휘로 선정된 형태소를 빈도순으로 배열한 것이다. 왼쪽에서 오른쪽으로 위에서 아래의 순서로 본다. 최고빈도 기본어휘 목록에서 최상위 30 위 안에 있는 실질어는 '것', '있다', '하다' 단 세 개이다. 등급별로 형태소수를 분류하면 표 5와 같다.

등급	1	2	3	4	5
형태소수	6,130	11,449	17,737	11,378	8,103

[표 5] 등급별 분포

[20] 표 4에서 품사를 표시하는 기호는 카이스트 태그셋을 적용한 표시이다. 카이스트 태그셋은 본 글의 끝에 부록으로 첨부한다. 인터넷에서 보고자 하는 경우, http://morph.kaist.ac.kr/ ~morph를 참조하기 바란다.

여기서 4등급은 교양 있는 성인 수준의 형태소이고 5등급은 전문성이 높은 형태소로서 기본어휘에 포함되지는 않는다.

기본어휘로 선정된 형태소 전체에 대하여 품사별로 비율을 나열하면 표 6과 같다.

품사	분포		품사	분포		품사	분포	
	형태소	비율		형태소	비율		형태소	비율
ncn	39,513	72.11	npp	78	0.14	ecx	10	0.018
ncpa	6167	11.25	xp	76	0.14	jcs	7	0.013
pvg	2907	5.31	jca	52	0.099	jcm	6	0.011
mag	2143	3.91	npd	44	0.08	jcv	6	0.011
paa	1424	2.6	jxc	43	0.08	jct, jcr	6	0.011
ncps	927	1.45	nq	38	0.06	xsm	5	0.009
nbu	279	0.51	px	32	0.058	xsv	4	0.007
xsn	228	0.42	maj	30	0.055	xsa	4	0.007
ii	182	0.33	etm	20	0.036	jco	3	0.005
ecs	140	0.26	ecc	19	0.035	etn	3	0.005
ef	130	0.24	nnc	15	0.027	jcc	2	0.004
mma	127	0.23	jcj	11	0.02	jp	2	0.004
nbn	101	0.18	ep	11	0.02	jxf	1	0.002

[표 6] 품사별 비율

등급이 부여된 모든 형태소의 품사별 분포를 고빈도에서 저빈도로 나열하여 살펴보면 가장 다양한 형태소를 제시하면서 최대의 비율을 보이는 품사는 예상처럼 명사였고 전체 기본어휘의 72%에 달한다. 서술성 명사와 상태성 명사까지 포함하면 84.81%에 이르며 그 뒤를 동사, 부사, 형용사가 차지한다. 이를 다시 실질어와 기능어로 분류하여 그 비율을 보면 실질어[21]가 98.28%를 차지하고 기능어는 전체 기본어휘의 단 1.72%만을 차지할 뿐이다.

6. 평가

선정된 형태소는 한국어 기본어휘로서 얼마나 자격이 있을까? 목록을 주관적으로 평가하는 대신에 각종 텍스트 코퍼스에서 기본어휘가 얼마나 어휘를 걸러 내는지 실험할 필요가 있다.

6.1 실험 코퍼스

7,000만 어절 카이스트 코퍼스에서 분석 코퍼스에 포함되지 않은 50Kb 크기의 텍스트를 크기만 고려하여 고른 다음 형식이나 언어 수준에 구애 받지 않고 교양 수준의 텍스트를 골라 <비형식>으로 분류하고, 논문이나 법률 문서, 설명문 등의 다분히 학문적이고 정형적인 텍스트들을 골라 <형식>으로 분류하였다. <형식>과 <비형식>에는 각각 43개와 45개 파일이 선택되었다.

[21] 실질어 목록에는 모든 명사류와 감탄사, 관형사를 포함시켰다.

6.2 실험

두 가지 유형의 부분 코퍼스[22]를 형태소 분석한 뒤 기본어휘로 선정된 형태소들이 각 부분 코퍼스를 구성하는 형태소를 얼마나 커버하는지 조사하였다. 사실, 주제상의 일관성이 있는 경우 한정된 어휘가 커버할 수 있는 텍스트의 크기도 커진다. 그러나 본 실험 코퍼스는 적은 양의 서로 다른 주제의 텍스트들로 구성되어 점유율이 비교적 낮게 나올 수밖에 없다. 이러한 사실을 고려하면서 결과를 보자면, 기호(s), 외국어(f), 비매칭 경우를 제외하고 두 개의 부분 코퍼스에서 기본어휘 점유율은 다음과 같다.

형식	전체(460448/504824)	91.21%
	1등급(371765)	73.64%
	2등급(50241)	9.95%
	3등급(27612)	5.47%
	4등급(7224)	1.43%
	5등급(3606)	0.71%
비형식	전체(512624/549895)	93.24%
	1등급(447878)	81.45%
	2등급(42387)	7.71%
	3등급(16548)	3.01%
	4등급(4123)	0.75%
	5등급(1777)	0.32%

[표 7] 등급별 점유율

품사	카이스트 품사 분류 표시	형식	비형식
명사	ncn	65.36%	73.82%
동사, 형용사	pvg, paa	97.92%	97.88%
감탄사	ii	99.03%	77.53%
부사	mag	97.01%	98.03%,
접사	xsn, xsa, xsv, xsm	97.33%	97.59%
어미	ecc, ecs, ecx	98.70%	97.87%
조사	jcs, jco, jcc, jcm, jcv, jca, jct, jcr	98.37%	98.41%
접속사	maj	98.36%	99.64%
관형사	mma, mmd	98.95%	99.63%

[표 8] 품사별 점유율 (1등급)

명사와 감탄사만을 제외하고는 매우 높은 점유율을 보였다. 명사와 감탄사에서 점유율이 낮은 것은 형태소 분석기가 대부분의 오분석 결과를 명사로 처리하였기 때문이고, 감탄사는 기본어휘로 선정된 감탄사가 한정적이기 때문이다. 비매칭된 명사류를 관찰하면, 상당수가 오분석의 결과여서 좀더 섬세한 실험을

[22] 부분 코퍼스란 Ch. Muller의 용어로 코퍼스 내부에 구성되어 있는 또 다른 자료 집합을 말한다.

위해서는 이러한 오류를 수정하는 과정이 필요하다. 오분석 수정 과정이 들어간다면 점유율이 높아질 것으로 예상된다.

6.3 결과 분석

전체적으로 형식 코퍼스에서는 91.21 %, 비형식 코퍼스에서는 93.24 %의 점유율을 보였다. 이 결과가 다양한 주제의 텍스트를 모은 소규모 코퍼스에 대한 실험이라는 점을 감안하여도 약 50,000여 개 어휘면 많은 경우 텍스트의 95% 이상을 커버한다는 점에서 그다지 주의를 끌지는 않는다. 그러나 등급별 점유율과 품사별 점유율이 매우 흥미롭다. 단 6,130개의 1등급 형태소만으로 실험코퍼스 전체의 73.64%와 81.45%를 차지하였다. 그러나 품사별 점유율에서 명사와 감탄사 외에 다른 품사에서 1등급 형태소의 점유율이 모두 97% 이상이었던 점, 그리고 명사와 감탄사에 대한 결과가 형태소 오분석에 의한 것으로 추정되었던 만큼, 이를 수정한다면 1등급 형태소의 코퍼스 점유율은 더욱 높아질 것이다. 한편, 1등급 형태소의 점유율이 비형식 문서에서 보다 형식 문서에서 낮게 나타났는데, 이는 1등급이 난이도가 낮은 형태소이므로 고난이도 형태소를 다량 포함하고 있을 형식 문서에서 점유율이 내려가는 것으로 해석할 수 있겠다. 그러나 2등급 이하에서도 이러한 일관성을 보이지는 않았다. 이에 대한 좀더 면밀한 분석과 분석 결과를 바탕으로 한 보완이 필요하다.

7. 결론

기계용 전자사전 항목 결정을 위하여 선정된 국내 최초의 한국어 기본어휘인 만큼 한국어 자연언어처리 시스템에 중요한 역할을 하게 될 것이다. 그러나 본 기본어휘 선정은 아직 보완해야 할 점들이 많이 있다. 연구 자체에 대한 보완 사항을 살펴보면 다음과 같다. 첫째, 많은 부분이 전문가의 주관적 판단으로 이루어졌기 때문에, 각 형태소의 난이도 결정에 대한 객관성 확보가 요구된다. 둘째, 선정된 기본어휘에 대한 좀더 섬세한 평가가 요구된다. 연구 자체에 대한 보완과 아울러 실제 연구에 대해서도 보완이 이루어져야 한다. 첫째, 비고로 처리하였던 동형어와 다의어 관련 기록이 정리 과정에서 누락되었을 가능성이 있어 보완이 필요하다. 둘째, 개체명과 각 분야 전문용어에 대한 보완이 필요하다. 셋째, 한글학회『우리말 큰 사전』을 기준으로 다의어 구분이 이루어졌으나 다의어 구분자의 기호화가 필요하다. 넷째, 본 연구가 코퍼스를 중심으로 이루어져 코퍼스에 나타나지 않은 기본적 어휘들과 웹 언어에 대한 보완이 필요하다. 이와 아울러 형태소 분석 과정에서 오류로 나타난 항목들에 대한 재정리도 이루어져야 할 것이다.

<참고문헌>

21세기 세종계획. 2000. 전자사전 개발 자료. 문화관광부.

강영수. 2003. 한국어 형태소 사전의 품질향상을 위한 사전 관리 워크벤치의 개발. 한국과학기술원 전산학과 석사학위 논문.

김광해. 2001. 한국어 어휘 목록. 서울대학교 국어교육연구소.

김홍규. 강범모. 2000. 한국어 형태소 및 어휘 사용 빈도의 분석. 고려대학교 민족문화연구원.

배희숙 외. 2001. 한국어 형태소의 계량언어학적 연구-신문 사설을 중심으로-. 인간과 기계와 언어, 17-24.

서상규. 2000. 한국어 교육 기초어휘 의미 빈도 사전의 개발. 문화관광부.
안효은. 2002. 전문용어의 특성정보를 이용한 전문분야 형태소 분석기. 한국과학기술원 전산학과 석사학위 논문.
연대언어정보개발 연구원. 사전편찬학 1-8.
이민행 외 옮김. 1996. 새로운 의미론. 한국문화사.
이응백 외. 1987. 국민 학교 교육용 어휘 (1, 2, 3학년용). 국어연구소.
정호성. 1999. 표준국어대사전 수록 정보의 통계적 분석. 새국어생활 10.1.
최기선. 2001. KAIST 대용량 코퍼스. KAIST 언어자원 2001년도판. 과학기술부 핵심소프트웨어 과제 결과물 (http://kibs.kaist.ac.kr).
Charles Muller. 1977. *Principes et Méthodes de statistique lexicale.* Hachette.
http://128.134.207.22/hkn/Conclusion.htm
http://www.g-matrix.pe.kr/feature/ai/fuzzy4.htm

#부록 (카이스트 품사 분류표)

s(기호)		1. 쉼표(sp)	2. 마침표(sf)
		3. 여는 따옴표 및 묶음표(sl)	4. 닫는 따옴표 및 묶음표(sr)
		5. 이음표(sd)	6. 줄임표(se)
		7. 단위 기호(su)	8. 기타 기호(sy)
외국어(f)		9. 외국어(f)	
체언(n)	서술성 명사(ncp)	10. 동작성 명사(ncpa)	11. 상태성 명사(ncps)
	비서술성 명사(ncn)	12. 비서술성 명사(ncn)	
	고유명사(nq)	13. nq(고유명사고유명사)	
	의존명사(nb)	14. 단위성 의존 명사(nbu)	15. 비단위성 의존 명사(nbn)
	대명사(np)	16. 인칭 대명사(npp)	17. 지시 대명사(npd)
	수사(nn)	18. 양수사(nnc)	19. 서수사(nno)
용언(p)	동사(pv)	20. 일반 동사(pvg)	21. 지시 동사(pvd)
	형용사(pa)	22. 성상 형용사(paa)	23. 지시 형용사(pad)
	보조 용언(px)	24. 보조용언(px)	
수식언(m)	관형사(mm)	25. 성상 관형사(mma)	26. 지시 관형사(mmd)
	부사(ma)	27. 일반 부사(mag)	28. 지시 부사(mad)
		29. 접속 부사(maj)	
독립언(i)	감탄사(ii)	30. 감탄사(ii)	
관계언(j)	격조사(jc)	31. 주격 조사(jcs)	32. 목적격 조사(jco)
		33. 보격 조사(jcc)	34. 관형격 조사(jcm)
		35. 호격 조사(jcv)	36. 부사격 조사(jca)
		37. 접속격 조사(jcj)	38. 공동격 조사(jct)
		39. 인용격 조사(jcr)	
	서술격 조사(jp)	40. 서술격 조사(jp)	
	보조사(jx)	41. 통용 보조사(jxc)	42. 종결 보조사(jxf)
어미(e)	종결 어미(ef)	43. 종결 어미(ef)	
	선어말 어미(ep)	44. 선어말 어미(ep)	
	연결 어미(ec)	45. 대등적 연결 어미(ecc)	46.종속적 연결 어미(ecs)
		47. 보조적 연결 어미(ecx)	
	전성 어미(et)	48. 명사형 어미(etn)	49. 관형사형 어미(etm)
접사(x)	접두사(xp)	50. 접두사(xp)	
	접미사(xs)	51. 명사 파생 접미사(xsn)	52. 동사 파생 접미사(xsv)
		53. 형용사 파생 접미사(xsm)	54. 부사 파생 접미사(xsa)

접수일자: 2003년 4월 16일
게재결정: 2003년 6월 16일