21세기 세종계획 원시 말뭉치의 유니코드와 코드 변환

강승식
국민대학교 전자정보통신대학 컴퓨터공학부
sskang@kookmin.ac.kr

Unicode and Code Conversion for Sejong 21 Raw Corpus

Seung-Shik Kang
School of Computer Science, Kookmin University

요 약

21세기 세종계획은 국어정보화를 위한 범국가적 사업으로서 국어 기초 자원을 구축하는데 매우 큰 기여를 하였으며, 그 주요 결과물로 배포된 세종 말뭉치는 많은 연구자들에게 꼭 필요한 가치있는 결과물이다. 이처럼 소중한 국어 자원을 실제 연구자들이 활용하고자 할 때 불편함을 느끼는 경우가 있는데 그 이유는 균형 말뭉치의 구축이라는 말뭉치의 특성 및 원문 자료의 내용을 최대한 보존하기 위한 노력의 일환으로 사용자 정의 영역에 정의된 문자들이 다수 포함되어 있기 때문이다. 본 논문에서는 자연언어 처리, 정보검색 분야 연구자들이 세종계획 최종 결과물 중에서 원시 말뭉치를 활용하는데 있어서 말뭉치에 사용된 문자코드의 유형을 중심으로 코드 변환 문제점과 그 해결 방안을 모색하고자 한다.

키워드: 세종계획, 원시 말뭉치, 태깅 말뭉치, 국어 정보처리, 유니코드

1. 서론

"21세기 세종계획"은 문화관광부가 국립국어원 및 관련 학계와 더불어 국어 정보화를 위해 1998년부터 2007년까지 10년간 150억원을 투여한 국책 과제이다.[1] 이 과제의 주요 사업은 말뭉치 구축, 전자사전 편찬, 한민족 언어정보화, 인력 양성 등이다. 그 중에서 말뭉치 구축은 관련 연구자들에게 매우 파급효과가 큰 핵심적인 부분으로 국어 정보화를 위해 필수적인 기초자원과 국어 데이터베이스를 구축하고, 국어 정보처리 소프트웨어를 개발하여 그 결과물을 관련 연구자들이 활용할 수 있도록 배포하고 있다[1].

말뭉치는 자연언어 처리를 위한 기초 자료로서 많은 연구자들이 필요로 하는 활용 가치가 매우 높은 자원이다[2,3,4]. 대용량 국어 말뭉치 자료를 구하기가 어려웠던 시기에 국어 정보화를 위한 대규모 사업으로 시작되었던 세종계획 사업은 국어학, 언어학, 전산언어학, 그리고 자연언어처리와 정보검색 관련 분야 및 산업체에서 매우 많은 관심의 대상이 되었다. 최근에는 웹문서, 블로그, 신문기사 등이 데이터베이스로 관리되고 있기 때문에 대용량 자료를 비교적 쉽게 구할 수 있으나 이 자료들은 철자 및 띄어쓰기 오류 등 불필요한 요소(noise)들이 다수 섞여 있다. 또한, 정보화 사업의 일환으로 각종 정보자료 구축되어 특허, 학위 논문, 법률자료 등 대규모 데이터베이스가 구축되고 있으나 균형 말뭉치가 아닌 전문 분야별 말뭉치로서 가치가 있다.

세종 말뭉치는 균형 말뭉치의 특성과 태깅된 말뭉치를 제공한다는 장점 때문에 한국어의 다양한 언어 현상을 연구하는데 매우 유용하게 활용되고 있으며, 특히 균형 말뭉치 특성은 어휘 빈도 조사 및 용례 검색 등의 연구에서 중요하게 활용되고 있다[5,6,7]. 그런데 국어 및 언어학 관련 연구 분야에서 중요한 연구를 수행하는데 많은 기여를 하고 있는데 비해 상대적으로 자연언어 처리와 정보검색 연구 분야에서는 사업 초기에 예상했던 것보다 그 활용도가 기대했던 것보다 낮은 편이다. 그 이유는 세종 말뭉치의 구축 목적이 "국어 기초 자료 구축"이라는 특성과 국어학 연구에 필요한 균형 말뭉치의 구축 등 말뭉치를 구축하는 목적이 자연언어 처리나 정보검색 분야의 연구 목적과 부합하지 않은 부분들이 있기 때문이다.

세종 말뭉치를 활용하고자 할 때 발생하되는 문제점은 말뭉치의 특성이 해당 연구 목적에 부합되지 않은 점에 기인하기도 한다. 일례로, 현대 문어체 말뭉치는 현대 국어 문장들의 집합이라고 가정하고 통계 자료를 추출하고자 한다면 국어학 연구를 위해 국어학 분야 자료에 포함된 한글 고어와 가운데점, 인용부호 등으로 인해 세종 말뭉치에 대한 만족도가 낮아질 수 있다. 이로 인해, 관련 연구를 수행하는 연구자들은 세종계획 결과물 자료를 바로 사용할 수 있을 만큼 만족스럽지 않을 수 있다. 이 경우에 각 연구자들은 말뭉치에서 필요한 자료들을 추출하는 과정에서 부딪히는 여러 가지 문제점을 스스로 해결해야 한다. 아래 사례는 세종계획 결과물을 활용했던 연구자의 블로그에서 인용한 것으로 세종 말뭉치를 사용하는 과정에서 발생한 애로사항을 잘 표현하고 있다.[2]

> 그래서 인용부호의 오류를 수정하는 프로그램 및 각종 깨진 문자 등을 제거하기 위한 프로그램을 별도로 짜는데에만 거의 한 달 이상 작업을 진행중이다. 오늘도 [redacted], 이렇게 말뭉치 정재작업을 하고 있다. (계속 해야할지 자꾸 의문이 든다...)

1) 세종계획 홈페이지(http://www.sejong.or.kr/) 참조

2) 이 논문은 세종계획 최종 결과물인 2007년 12월 배포판을 활용하는 것에 관한 것으로, 차기 배포판들에서는 이 문제점들이 해결되어 더이상 문제가 되지 않을 것이다.

이와 같은 사례는 세종계획 결과물을 활용할 때 연구자들이 겪게 되는 것으로 이것은 세종 말뭉치를 구축할 때 최대한 원문 정보를 보존한다는 원칙에 기인한 것이다. 즉, 국어 자료를 입력할 때 문장부호 하나까지도 최대한 보존하는 원칙에 충실하게 되면 출판물 등에서 사용한 가운데점의 경우 점의 크기와 모양에 따라 여러 가지의 문자코드로 구축된다.[3] 세종 말뭉치의 이러한 특성을 알지 못하는 경우 말뭉치에서 필요한 정보를 추출할 때 가운데점의 일관성 문제 및 사용자 정의 영역에 정의된 인용부호로 인하여 말뭉치를 활용하는데 어려움을 있을 수 있다. 이러한 자료는 정보추출 혹은 통계 조사에서 제외시킬 수 있으며 이러한 자료들을 제외하기 위하여 EUC-KR(KS 완성형)에 정의되지 않은 문자가 포함된 문장들을 일괄적으로 제외시키는 방법을 사용하면 된다. 또한, 유니코드의 사용자 정의 영역(private use zone)에 정의된 문자 등 연구 목적과 무관한 문자들이 포함된 자료도 제외시킬 필요가 있기도 하다.

본 논문에서는 세종 말뭉치 2007년 배포판을 활용하는데 발생하는 장애 요인 및 그 원인을 파석하여 많은 사용자들이 쉽게 활용할 수 있는 방안을 모색하고자 한다. 그럼으로써 통계 정보를 추출한다든지 사용자가 원하는 목적으로 세종 말뭉치를 활용하고자 할 때 부딪히는 문제점을 최소화 하고 사용자가 세종 말뭉치를 가공할 때 최소한의 노력을 들일 수 있도록 한다.

2. 세종 말뭉치 2007년 배포판 개요

세종계획 최종 결과물로 2007년 12월에 배포된 말뭉치에는 병렬 말뭉치 등 여러 가지 유형의 말뭉치가 있다. 세종 말뭉치에서 현대 문어체 말뭉치는 균형 말뭉치(balanced corpus)이다. 21세기 세종계획 프로젝트의 최종 결과물로 2007년 12월에 배포된 말뭉치 중에서 현대 문어체 자료는 원시 말뭉치(raw corpus), 형태소 태깅(part-of-speech tagging), 형태-의미 태깅(sense tagging), 구문 태깅(syntactic tagging) 말뭉치로 구성되어 있다. 각 말뭉치의 크기는 표 1과 같다.[4]

표 1. 2007년 배포판 현대 문어 말뭉치의 크기

현대 문어 말뭉치	파일 개수	어절수
원시 말뭉치	2,109	6,700만
형태소 태깅 말뭉치	301	약1천만
형태-의미 태깅 말뭉치	339	약1천만
구문 말뭉치	31	

이 네 가지 유형의 말뭉치를 서로 구분하여 지칭하기 위하여 본 논문에서는 편의상 **원시/형태/의미/구문** 말뭉치라 칭한다. 원시 말뭉치는 현대 문어체 균형 말뭉치(balanced corpus)로 신문기사, 전문분야 문서, 문학작품 등으로 구성되어 있다. 각 말뭉치 파일은 XML 형식으로 구축되었으며 앞부분에는 말뭉치 내용에 대한 원문 정보, 말뭉치 구축 작업 관련 작성자, 작성일, 수정 내역 등 메타 정보가 기술되어 있다. 말뭉치 파일의 형식을 XML과 같이 구조적인 형태로 구축하는 것과 단순 텍스트 형식으로 구축하는 것은 논란의 여지가 있다.

원시 말뭉치는 국어학 연구 분야의 자료에서 나타난 고어 문자들이 포함되어 있으며, 이러한 문자들이 포함된 파일들은 활용 목적에 따라 제외하고 사용할 필요가 있다.[5] 원시 말뭉치 중에서 4BH20002.txt에는 "우ᄂᆞᆫ ᄋᆞ히 ᄯᅩᆼ 먹이기 ᄇᆡᆨ발로인 벗ᄒᆞ고 옹긔쟝ᄉᆞ 작ᄃᆡ 치기 ᄑᆡᄂᆞᆫ 곡식 이삭 ᄲᅦ기 잣친 밥에 모ᄅᆡ 넛키..."[6]라는 문장이 있다. 이 파일은 '홍보가'에 관한 국어학 서적에서 추출한 것으로 말뭉치 활용 목적에 따라 이와 같은 유형의 문장들을 제외시킬 필요가 있다.

형태 말뭉치는 형태소 분석 및 품사 태깅 결과물이고, 의미 말뭉치는 형태 말뭉치의 태깅 결과에 어휘형태소의 의미 정보를 추가한 것이다. 즉, 두 가지 이상의 의미로 사용되는 어휘형태소에 대해 문장내에서 어떤 의미로 사용되었는지를 결정해 주는 정보를 추가한 것이다.[7] 예를 들어, 아래 '정당구도에서는'의 예에서 '정당'은 7번째 의미로, '구도'는 10번째 의미로 결정되었음을 의미한다.

정당__07/NNG + 구도__10/NNG + 에서/JKB + 는/JX

따라서 의미 말뭉치는 어휘형태소의 의미 태깅 정보인 '어깨번호'를 제거하면 형태 말뭉치와 동일하게 된다. 다만, 현재 말뭉치 파일의 앞부분에는 파일 내용을 기술하는 메타 정보가 XML 형식으로 기술되어 있으며, 형태와 의미 말뭉치의 메타 정보 부분은 일치하지 않는다.

의미 말뭉치는 형태 말뭉치에 의미 정보를 추가한 것이므로 형태 말뭉치와 의미 말뭉치를 동시에 배포하기보다는 의미 말뭉치만 배포하고 이로부터 형태 말뭉치를 생성해 내는 도구를 지원하는 방법으로 말뭉치를 배포하는 것이 가능하다. 말뭉치 구축 작업은 원시말뭉치를 구축한 후에 형태 말뭉치를 구축하고, 다시 그 후속작업으로 의미 말뭉치와 구문 말뭉치를 구축하게 된다. 따라서 배포판에 수록된 의미 말뭉치와 구문 말뭉치는 형태 말

3) 인용부호와 관련된 이 문제점은 유니코드로 구축되어 있는 세종 말뭉치를 EUC-KR(KS 완성형)로 변환할 때 발견된다. 그런데 사용자는 해당 문자 모양이 표시되지 않거나 네모 형태의 점으로 표시되어 어떤 문자인지 확인이 곤란하다.

4) 세종 말뭉치는 여러 개의 파일로 구성되어 있어서 전체 어절수를 계산하기가 쉽지 않으므로 배포판에서 총 어절수 정보를 제공하는 것이 좋다.

5) 원시 말뭉치 파일 2BH9623.txt은 파일 뒷부분이 다양한 특수문자들로 채워져 있으며, 이 문자들은 코드변환 오류에서 제외하였다.

6) 이 문장은 윈도 환경에서 메모장이나 MS word 등 일반적인 소프트웨어에서는 폰트가 지원되지 않는 문자들이 포함되어 있으며, 훈글에서는 이 문자들의 폰트가 지원되고 있다.

7) 어휘형태소의 의미 구분 정보는 표준 국어 대사전을 기준으로 하며, 그 의미를 구분하기 위한 숫자 정보를 '어깨번호'라고 하는 위첨자(superscript) 형태로 표시한다.

뭉치 파일들과 부분-전체 관계로 구성되는 것이 자연스러울 것이다.[8] 구문 말뭉치는 31개 파일로 구성되어 있으며 형태 말뭉치를 기반으로 파스 트리 형태로 구문 태깅 작업을 수행한 것이다. 2007년도 배포판에서 구문 태깅 파일 중 형태/의미 말뭉치와 대응되는 파일 개수는 20여개이며 구문 태깅 파일에 대응하는 형태/의미 파일들이 배포판에 모두 포함된 것은 아니다.

3. 원시 말뭉치의 문자코드

세종 말뭉치 2007년도 배포판은 유니코드 UTF-16(Little Endian)으로 작성되었다. C/C++ 환경에서 말뭉치 활용 작업을 고려하면 말뭉치가 KS 완성형으로 구축되는 것이 편리할 것이다.[9] 그런데 EUC-KR(KS 완성형)은 아스키 확장 영역 0x8F-0xFF 에 정의되어 있는 Latin 계열 문자(독일어 움라우트, 발음기호 등)를 표기할 수 없는 등의 제약이 있다. 따라서 Latin 계열 문자가 포함되어야 하는 경우 및 국제 표준을 고려할 때 유니코드의 UTF-8 인코딩 방식을 따르는 것이 타당할 것이다.[10]

3.1 원시 말뭉치의 코드 변환 오류

원시 말뭉치 2,109개 파일을 유니코드에서 EUC-KR로 변환할 때 376개 파일에서 2,440개 문자에 대한 변환 오류가 발생한다.[11] 이 2,440개 문자 중 확장 아스키 영역의 184자를 제외하면 2,256자이다. 표 2의 문자들은 단순히 유니코드를 EUC-KR로 변환할 때 변환 오류가 발생하는 것으로 변환 오류가 발생하는 문자들 중에서 가운데점, 따옴표, 원 문자 등의 문자를 제외한 것이다. 이 문자들을 제외한 이유는 이 문자들의 사용 빈도가 높을 뿐만 아니라 원래 모양을 유지하는 것보다는 유사한 문자로 변환하는 것이 말뭉치의 활용도 측면에서 더 효율적이라고 판단되기 때문이다.

변환 오류가 발생하는 2,256자 중에서 출현빈도가 1인 것이 781자, 출현빈도 2인 것이 328자로 전체 문자의 49.2%를 차지한다. 출현빈도가 가장 높은 것은 빈도수 499인 한자 U+756B('획')이고, 두 번째로 빈도가 높은 것은 U+9ED9(한자음 '묵')이다.[12] 세 번째로 빈도수가 많은 것은 111회인 U+5D47(한자음 '혜')이다.

라틴어 확장 문자와 문장부호 등 U+00A0-U+33FF 영역에 정의된 문자들 중에서 많은 문자들이 EUC-KR에 정의되지 않거나 유니코드 환경으로 전환되면서 EUC-KR 문자표에 대응되지 않는다. 그 대표적인 예로 아스키 확장 영역에 정의되는 독일어 문자 움라우트와 발음 기호 등이 있다.

표 2. 원시 말뭉치의 코드 변환 오류 문자

코드 범위	문자수	설명
U+00A0 - U+33FF	184	Latin Extended 등 문자개수가 적은 언어, 문장 부호, 기호
U+3400 - U+4D00	1,772	CJK Unified Ideographs Extension A
U+4E00 - U+9FFF		CJK Unified Ideographs
U+DC00 - U+DFFF	2	Low-half zone of UTF-16 (U+DC50, U+DC51)
U+E000 - U+F8FF	434	Private Use Zone
U+F900 - U+FFFF	19	CJK Compatibility Ideographs, Arabic 등
U+20000-U+2A5F0	14	CJK Unified Ideographs Extension B
U+F0000-U+FFFFD	15	Supplementary Private Use Area-A

원시 말뭉치를 활용하는 과정에서 부딪히는 문제점은 운영체제가 해당 폰트를 지원하지 않아서 어떤 문자인지를 알 수가 없다는 점이다. 윈도의 메모장에서는 EUC-KR에 정의되지 않은 한자 1,786(=1,772+14)자를 '■'로 표시하게 된다. 특히, 사용자 영역에 정의된 문자 449(=434+15)자는 공백으로 표시해 주기 때문에 공백 문자와 혼동하게 된다. 원시 말뭉치에 사용된 문자코드 관련하여 아래와 같은 의문점들이 발견된다.

1) 원시 말뭉치에서 문장부호 등 특수기호의 원형을 보존할 필요성
2) U+F0000-U+FFFFD에 정의된 15자를 "일반 사용자 영역"이 아닌 "보조 사용자 영역"에 정의한 이유
3) UTF-16의 하위 영역(low-half zone of UTF-16)인 U+DC00-U+DFFF은 유니코드 보조셋 문자를 표기하기 위한 특수 목적으로 예약된 영역이다. 이 영역에 정의된 2개의 유니코드 문자 U+DC50, U+DC51이 사용된 이유[13]

3.2 원시 말뭉치의 가운데점, 원 문자, 인용부호

원시 말뭉치에는 가운데점에 관한 유니코드 문자가 여러 가지로 입력되어 있다. 가운데점의 크기에 따라 U+2218, U+2219, U+22C4, U+22C5, U+F85E(사용자 영역에 정의) 등이 있다. 모양이 조금씩 다른 가운데점을 입력하기 위해 사용된 이 문자들은 EUC-KR로 변환할

8) 형태 말뭉치와 의미 말뭉치가 서로 중복되지 않도록 구성할 수도 있지만 그 경우에 의미 말뭉치로부터 자동으로 형태 말뭉치를 생성해 낼 수 있기 때문에 굳이 그럴 필요가 없다.

9) Java 언어로 작업하는 환경에서는 자바의 기본 문자셋인 유니코드가 편리하다.

10) EUC-KR(KS 완성형)을 취할 경우 독일어 움라우트 등 아스키 확장영역 문자들뿐 아니라 KS 완성형에 정의되지 않은 한자를 사용할 수 없는 등의 제약이 있다.

11) 변환 오류가 가장 많이 포함된 파일은 4BH20002.txt이고 3,477개가 포함되어 있다. 두 번째로 많은 파일은 7BH02036.txt이고 1,997개이다.

12) 유니코드 문자는 http://www.unicode.org/charts 에서 유니코드 값을 입력하여 어떤 글자인지 확인할 수 있다. 또는, http://www.unicode.org/roadmaps/bmp/ 에서 각 영역별 유니코드 배치도를 알 수 있다.

13) 이 2개 문자의 출현빈도는 각각 1회씩이고, 이 문자들이 출현한 파일은 각각 2CC00118.txt, 3BA00B04.txt이다.

때 동일한 문자로 변환되는 것이 타당할 것이다.

가운데점 문자와 유사하게 유니코드를 EUC-KR로 변환할 때 대응문자 문제가 발생하는 유니코드 문자의 다른 예로는 원문자가 있다. 유니코드에는 원의 크기와 모양에 따라 여러 개의 원 문자 코드가 U+25C9-U+25CF로 부여되어 있다. 또한, 유니코드 문자 U+2776-U+2793은 1부터 10까지 숫자에 대한 원문자가 각각 바탕색이 흰색인 것과 검은색인 것이 구별되어 있다. 그런데 EUC-KR은 이러한 원문자에 대한 코드가 다양하지 않으므로 유사한 문자로 변환되어야 한다.

표 3. 원시 말뭉치의 가운데점, 원 문자, 인용부호

문자	유니코드
가운데점	U+2218, U+2219, U+22C4, U+22C5, U+F85E(사용자 영역)
원 문자	U+25C9 - U+25CF, U+2776 - U+2793
인용부호	U+FF62, U+FF63, U+F08DC, U+F09DC, U+F0ADC, U+F0BDC

한글 문서에서만 사용되는 문장부호 중에서 유니코드에 적절한 문자코드가 정의되어 있지 않은 큰따옴표와 작은따옴표는 원시 말뭉치에서 사용자 정의 문자로 정의하여 사용하는 경우가 있다. 그 예로는 U+FF62, U+FF63, U+F08DC, U+F09DC, U+F0ADC, U+F0BDC 등이다.[14] 가운데점과 원 문자, 그리고 사용자 영역에 정의된 인용부호 문자는 출현빈도가 높을 뿐만 아니라 원문에 표기된 모양을 보존하여 여러 가지로 표현하기 보다는 대표 문자 코드로 통일되는 것이 타당할 것이다.

4. 결론

2007년도에 배포된 세종계획 최종결과물을 활용하고자 할 때 사용자들이 불편함을 느끼는 경우가 있다. 본 연구에서는 현대 문어체의 원시 말뭉치를 중심으로 말뭉치를 활용할 때 발생하는 불편한 점을 원시 말뭉치에 사용된 문자코드의 유형을 중심으로 고찰하였다. 사용자 정의영역에 정의된 문자 등 코드변환과 관련된 문제는 형태소 분석 말뭉치와 형태의미 말뭉치에서는 원시 말뭉치에 비해 문제점이 매우 적으므로 원시 말뭉치만을 대상으로 하였다. 그 결과로, 원시 말뭉치를 구축할 때 원시 자료의 내용을 충실하게 보존하고자 하는 노력 때문에 EUC-KR에서 정의되지 않은 다양한 문자들이 포함되었기 때문이라는 것을 알 수 있었다.

본 논문에서는 원시 말뭉치에 포함되어 있는 문자들에 대하여 유니코드를 EUC-KR로 변환할 때 코드변환 오류를 발생시키는 문자들을 분석하여 유니코드 영역별로 분류해 보았다. 원시 말뭉치에서 통계 정보를 추출하는 등 연구 목적으로 활용하는 사용자들은 코드변환 오류가 발생하는 문자를 포함한 문장을 제외하여 필요한 정보를 추출하는 것이 신뢰도가 높은 자료를 얻는 방법이 될 수 있을 것이다. 또한, 코드변환 오류가 발생하는 파일들을 모두 제외시키더라도 2,109개 중에서 376개 파일만이 제외되므로 나머지 1,733개 파일만으로도 대용량 말뭉치로서 충분하며 전체 파일을 대상으로 정보를 추출하는 것보다 신뢰도가 높을 수 있을 것으로 예상된다.

세종 말뭉치는 다양한 사용자 계층의 요구사항을 반영하여 사용자들이 말뭉치를 활용하는데 불편한 점을 개선하기 위해 작업을 계속하고 있다. 2007년 배포판의 활용상의 문제점을 개선하여 새로운 버전을 배포하기 위해 준비하고 있으며, 새로운 배포판에서는 사용자 편의성과 활용성 부분이 개선되어 사용자들이 쉽게 활용될 수 있는 형태로 배포될 것이다. 그리고 새로운 배포판에서는 본 논문에서 제기된 활용상의 애로사항을 해소하여 사용자가 쉽고 편하게 말뭉치를 연구 목적에 활용할 수 있도록 할 것으로 예상된다. 또한, 세종 말뭉치를 활용하는데 애로사항이나 불편한 점은 http://www.sejong.or.kr/ 등 관련 홈페이지를 통하여 제언하거나 활용상의 문제점을 논의할 수 있으며, 그 해결 방안을 제시하거나 해결 방법을 찾을 수 있도록 준비하고 있다.

참고문헌

[1] 김홍규, 강범모 외, 21세기 세종계획 국어 기초 자료 구축 연구보고서, 문화관광부, 2006.

[2] J. Yoon, “Compound Noun Segmentation Based on Lexical Data Extraction from Corpus”, In *Proceedings of the 6th Applied Natural Language Processing*, pp.196-203, 2000.

[3] G. A. Miller and E. B. Newman, “Tests of a Statistical Explanation of the Rank-Frequency Relation for Words in Written English”, *American Journal of Psychology*, vol.71, pp.209-218, 1958.

[4] G. Leech, P. Rayson, and A. Wilson, “Word Frequencies in Written and Spoken English: based on the British National Corpus”, Longman, 2001.

[5] 강범모, 김홍규, 한국어 형태소 및 어휘 사용 빈도의 분석 2, 고려대학교 민족문화연구원, 2004.

[6] 강범모, 김홍규, 한국어 사용 빈도, 한국문화사, 2009.

[7] 임인빈, “세종 원시 말뭉치에서 살펴본 ‘상황’의 의미 운율”, KLing vol.3, pp.191-198, 2009.

14) 훈글 환경에서 구축된 말뭉치 파일을 유니코드로 저장할 때 아래한글 버전에 따라 훈글 내부에서만 사용되는 기호들을 유니코드로 변환한 코드 매핑 영역이 다르기 때문에 여러 가지 기호가 다양하게 변환되었을 가능성이 있다.