남북한 말뭉치 통합 방안에 대한 연구*

소 강 춘 (전주대)

< 목 차 >

국문초록

이 연구는 한국어 정보화의 초석이 되는 통합된 한국어 말뭉치 구축 방안을 모색하기 위한 것이다. 우리나라는 남북으로 분단되어 서로 다른 언어 규범을 사용하고 있을 뿐만 아니라, 중국의 조선족 동포들이 사용하고 있는 중국 조선어도 별도의 규범을 가지고 있다. 따라서 완전한 한국어 말뭉치를 구축하기 위해서는 이 세 지역의 언어규범을 포괄할 수 있는 방안이 마련되어야 한다.

그러나 이러한 외적 환경 조성은 용이한 일이 아니다. 그래서 이 연구에서는 그동안에 구축되었던 말뭉치 현황을 파악하고, 남북에서 구축된 말뭉치의 실례를 통해 문제점과 통합 가능성을 파악하여, 보다 효과적인 남북한 한국어 자료의 말뭉치 구축 방안을 제시하려 한다.

주제어: 정보화, 문자 말뭉치, 음성 말뭉치, 이미지 말뭉치, 분석표지

* 이 논문은 2015년 8월 14일 국립한글박물관에서 개최된 '광복 70주년 기념 겨레말 통합을 위한 국제학술회의'에서 '남북한 말뭉치 구축 방안에 대하여'라는 제목으로 발표한 자료를 수정한 것이다.

1. 서론

컴퓨터의 발달로 인하여 생생한 언어 자료를 연구 대상으로 삼으려는 언어학자들의 오랜 꿈이 이루어지고 있다. 대용량의 언어 자료를 컴퓨터로 처리할 수 있는 형태로 구축하여 활용하는 전산 언어학 또는 말뭉치 언어학의 발달은 자연언어를 전산 처리하는 기술을 발달시켰고, 나아가 인공지능이 가능한 컴퓨터의 출현이 가능하게 하고 있다.

말뭉치 언어학은 말뭉치를 근간으로 하는 학문이고, 말뭉치 언어학이 추구하는 목표가 언어의 정보화이기에 말뭉치는 언어 정보화의 가장 기초가 되는 자료이다. 따라서 언어를 정보화하기 위해서는 가장 우선적으로 효과적인 말뭉치 구축 방안이 마련되어야 하고, 구축된 말뭉치를 다양하게 활용할 수 있는 방안이 마련되어야 한다. 이것이 언어 정보화의 시작이다.

아래 <그림 1>과 <그림 2>는 언어 정보화 사업의 추진 과정과 결과를 그림으로 형상화한 것이다. 먼저 <그림 1>은 21세기 세종계획에서 제안한 내용이 성공적으로 완성되었을 때 얻어질 수 있는 다양한 결과물과 그 결과물을 활용해서 개발될 수 있는 열매를 그림으로 형상화한 것이다.[1] 이 그림에서는 국어정보베이스에 자연언어처리용 전자사전과 전문분야 대용량 시소러스를 개발해야 한다고 제안하고 있다. 그런데 전자사전과 시소러스를 개발하기 위해서는 더 기초가 되는 말뭉치 구축 작업이 전제되어야 하고, 실제로 21세기 세종계획에서는 다양한 종류의 대용량 말뭉치가 구축되었다.[2]

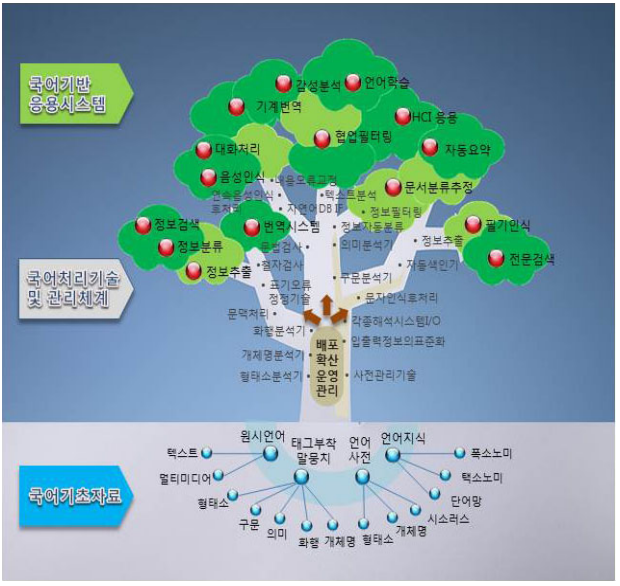
1) 이 국어 정보화 기술 나무는 특히 자연언어 처리의 기술 나무를 상정한 것으로, 한국과학기술정보연구원(KISTI)의 박동인 실장이 제시한 것이다.
2) 21세기 세종계획에서 구축된 말뭉치는 후술할 <표 1>에 정리되어 있다.

<그림 1> 세종계획 국어 정보화나무

이에 반해 <그림 2>는 포스트 세종계획을 위한 것이다.[3] 21세기 세종계획에서 추진됐던 결과물들의 부족한 점을 보완하고, 그동안 변화된 정보처리 환경을 반영한 것이다.

3) 이 그림은 강승식 교수가 2011년 문화체육관광부 과제로 수행한 '한국어 정보화 세부 실행 계획 수립'이라는 보고서에서 한국어 언어자원 역할의 중요성을 강조하기 위해 그린 국어 정보화 나무이다.

<그림 2> 강승식(2011)의 국어 정보화 나무



<그림 1>과 <그림 2>의 중요한 차이점은 <그림 1>에서 국어정보베이스가 되는 것은 자연어처리용 전자사전이나 대용량 시소러스를 구축하겠다는 내용인데, <그림 2>에서는 국어기초자료로 원시언어, 태그부착 말뭉치, 언어사전, 언어지식 등으로 구체화되고, 정밀화 되었다는 점이다.

그리고 <그림 1>에서는 통합관리시스템을 구축하려 했지만, <그림 2>에서는 국어처리기술 및 관리체계를 강조하고 있다. 즉 구축된 자료를 효과적으로 배포, 확산시킬 수 있는 국가차원의 통합 관리체계가 구축되어야 함을 강조하고 있는 것이다.[4] 이러한 차이는 세종계획 결과물이 축적되었고, 전산환경이 많이 변했기 때문이라고 할 수 있다.

<그림 1>과 <그림 2> 모두에서 공통적으로 중요하게 생각하고 있는 한국어 정보화 사업은 한국어 관련 자료들을 광범위하게 전산화하고, 이를 보

4) 이러한 체계 구축을 위해서 추진되고 있는 사업이 후술할 '개방형 한국어 지식대사전' 구축 사업이라고 할 수 있다.

다 정밀하게 분석, 분류, 체계화하여 범국가적인 관리 체계를 구축하여야 한다는 점이다.

이 연구의 목적은 한국어[5] 정보화의 초석이 되는 한국어 말뭉치 구축 방안을 모색하기 위한 것이다. 그런데 우리 한국어가 직면하고 있는 외적인 환경이 통일된 한국어 정보화에 많은 장애 요인을 가지고 있다. 우리나라는 남북으로 분단되어 서로 다른 언어 규범을 사용하고 있을 뿐만 아니라, 중국의 조선족 동포들이 사용하고 있는 중국 조선어도 별도의 규범을 가지고 있다. 따라서 완전한 한국어 말뭉치를[6] 구축하기 위해서는 이 세 지역의 언어규범을 포괄할 수 있는 방안이 마련되어야 한다.[7]

그러나 현실적으로 완전한 한국어 말뭉치 구축을 위한 외적 환경 조성은 용이한 일이 아니다. 그래서 이 연구에서는 그동안에 구축되었던 말뭉치 현황을 파악하고[8], 남북에서 구축된 말뭉치의 실례를 통해 문제점과 통합 가능성을 파악하여, 보다 효과적인 남북한 한국어 자료의 말뭉치 구축 방안을 제시하려 한다.

2. 남북한 원시 말뭉치 비교

원시 말뭉치를 구축하는데 있어서 중요한 것은 어떤 전자 문서 표준안을 따르느냐다.

---

5) '한국어'라는 용어는 남한에서 사용하고 있는 한국어, 북한에서 사용하고 있는 조선어, 중국에서 사용하고 있는 중국 조선어, 구소련 지역에서 사용되고 있는 고려어, 그리고 해외에서 사용되고 있는 한국어를 모두 통칭하는 의미로 사용한다.

6) 완전한 한국어 말뭉치는 규범이 다른 세 지역의 언어 자료를 하나의 규범이나 또는 새로운 통일 규범을 만들어 그 규범에 따라 구축된 말뭉치를 일컫는다.

7) 이러한 노력이 1994년부터 시작되었던 'Korean 컴퓨터 처리 국제 학술대회'이다. 이 학술대회에서는 자모순, 코드, 자판, 전산 용어의 통일을 위한 노력을 했고, 1996년에는 컴퓨터에서 사용한다는 제한을 달기는 했지만 자모순을 통일하고, 나머지 분야에서도 일정부분 통일안을 마련했다.

8) 말뭉치 구축 현황에 대해서는 자료의 현황을 소개한다는 측면에서 학술대회 발표 초록에는 상세하게 기술하였다. 그러나 이 논문에서는 생략하기로 한다.

2.1. 남한의 형태

남한의 원시 말뭉치 구축의 기본 원칙은 문헌의 경우 원문을 그대로 유지하고, SGML/TEI, 특히 TEI 라이트를 수용하여 문서 정보를 표기한다. 그리고 헤더는 자세히 표기하지만 본문 마크업을 최소한의 정보만 붙인다는 “21세기 세종 계획 말뭉치 구축 지침”(1998. 5.)을 준수하고 있다.

파일 이름은 도스 파일 명칭을 준수한다. 파일 이름은 8자리를 모두 사용하며, 3자리 확장자를 붙인다. 책의 경우, 한 권을 한 파일로 만드는 것을 원칙으로 한다.

<표 1> 파일명 부착 방법

자릿수	내 용	구 분	비고
첫 번째	년차 구분	1-9까지	
두 번째	나라 구분	N 북한 O 중국 R 구 소련	
세 번째	주요 분류	A : 신문 B : 잡지 C : 책-교과서 D : 책-사전 E : 책-소설 F : 책-시 G : 책-기타 상상적 텍스트(수필 등) H : 책-정보(인문, 사회, 자연, 예술), 기타 비상상적 텍스트	
네 번째 다섯 번째		신문 : 연도를 두 자리로 표시 잡지 : 연도를 두 자리로 표시 문학작품 : 자유롭게 두 자리 표시	
여섯 번째	부가정보	신문 : 종류 표시	

		Y : 연변일보 R : 레닌기치 잡지 : 종류 표시 A : 시사월간지 B : 시사주간지 C : 대중/연예지 D : 여성지 E : 경제지 F : 문화/예술지 G : 과학/기술지 H : 취미/실용지 I : 초등용 학생잡지 J : 중고등용 학생잡지 Z : 기타	
일곱 번째 여덟 번째	일련번호		

본문의 입력은 특별한 경우를 제외하고는 앞표지 부분, 목차 등과 부록, 색인, 뒷표지 등의 내용은 입력하지 않고, 본문만 입력하는 것을 원칙으로 한다.[9] 띄어쓰기는 원문의 형태를 유지하고, 상당한 오류가 발견될 경우에는 교정 태그를 사용한다. 문장 부호는 원문 그대로 유지하는 것을 원칙으로 하고, 가능하면 자판의 부호나 기호를 사용한다. 자판에 없는 경우는 일차적으로 전각기호를 다음에 반각기호를 사용한다. 가운뎃점은 전각기호(3404)을 말줄임표는 전각기호(3406)를 사용한다.

원문의 외국문자는 그대로 유지한다. 일차적으로는 자판에서 입력하되, 자판 입력이 불가능한 경우에는 전각 기호를 사용한다. 전각 기호가 없을 경우에는 반각 기호를 이용한다. 글자 겹침은 사용하지 않는다.

9) 이 부분은 북한과 차이가 있다. 이러한 원칙을 따른 것은 통계 처리에서 중복되는 어휘 숫자를 줄이기 위함이다.

‘흔글’에서 제공하는 틀(수식, 표, 그림) 등은 사용하지 않는다. 원문에 도표/그림이 있을 경우 생략하고, 그 사실을 명시적으로 밝혀 준다. 이때 <gap> 태그를 사용하고, reason 속성으로 그 이유를 밝힌다. 각주는 원칙적으로 생략하나, 보존할 필요가 있을 경우 <note> 태그를 사용하여 표기한다. 북한어, 연변어의 경우, 대화를 표시하는 각괄호는 전각 기호(《(3416), 》(3417))를 사용한다.

파일 전체 구조 및 SGML/TEI 태깅의 일반 원리는 아래와 같다.

(1) 남한 원시 말뭉치 견본

가. 문서의 구조

```
<teiHeader>
헤더 내용
</teiHeader>
<text>
본문 내용
</text>
```

나. 헤더의 구조

```
<!DOCTYPE tei.2 SYSTEM "c:\sgml\dtd\tei2.dtd" [
    <!ENTITY % TEI.corpus "INCLUDE">
    <!ENTITY % TEI.extensions.ent SYSTEM "sejong2.ent">
    <!ENTITY % TEI.extensions.dtd SYSTEM "sejong2.dtd">
]>
<tei.2>
<teiHeader>
<fileDesc>
<titleStmt>
<title>천리마, 전자파일</title>
<author>천리마사</author>
<sponsor>대한민국 문화관광부</sponsor>
<respStmt><resp>문헌 입력, 표준화, 헤더 붙임</resp>
```

```
<name>전주대학교</name>
</respStmt>
</titleStmt>
<extent>30704 어절</extent>
<publicationStmt>
<distributor>국립국어원</distributor>
<idno>NB96A030.hwp</idno>
<availability><p>2000-04 배포 불가</p></availability>
</publicationStmt>
<notesStmt>
<note><p></p></note>
</notesStmt>
<sourcedesc>
<bibl><author>천리마사</author>
<title>천리마</title>
<pubPlace>평양</pubPlace>
<publisher>천리마사</publisher>
<date>1996</date>
</bibl>
</sourceDesc>
</fileDesc>
<encodingDesc>
<projectDesc><p>21세기 세종계획 3차 년도 말뭉치 구축</p>
</projectDesc>
<samplingDecl><p>본문 전체 입력</p>
</samplingDecl>
</encodingDesc>
<profileDesc>
<creation><date>1996</date></creation>
<langUsage>
<language id=NK usage=99>북한어, 문화어</language>
</langUsage>
<textClass>
<catRef scheme='SJ21' target=N1260> 잡지: 사회, 일반</catRef>
</textClass>
</profileDesc>
<revisionDesc>
```

```
<change><date>(입력, 교정한 날짜)2000/06</date>
<respStmt><resp>입력자</resp><name>리현희
</name></respStmt>
<item>본문 전체 입력, 헤더 붙임</item>
</change>
<change><date>2000/06/12</date>
<respStmt><resp>제 1교 교정자</resp><name>한명숙
</name></respStmt>
<item>제 1교</item>
</change>
<change><date>2000/06/17</date>
<respStmt><resp>제 2교 교정자</resp><name>이성자
</name></respStmt>
<item>제 2교 및 헤더 수정</item>
</change>
<change><date>2000/06/23</date>
<respStmt><resp>최종 교정자</resp><name>소강춘
</name></respStmt>
<item>최종교정자</item>
</change>
</revisionDesc>
</teiHeader>
```

헤더 정보에 포함된 내용은 말뭉치가 완성되어 데이터베이스로 통합되어 사용될 경우에 필요한 정보와 21세기 세종계획을 추진하는 과정에 필요한 정보가 모두 포함되어 있다. 말뭉치가 구축되고 난 뒤에는 모든 말뭉치를 하나의 데이터베이스로 통합해서 운영할 경우 후자의 정보는 거의 사용할 필요가 없는 정보이다. 따라서 구축할 때도 헤더 정보를 간략화할 수 있을 것이다.

다. 본문 태그 구조

○ 일반적 형태

```
<text>
        <front>
                앞표지, 목차 등의 머리말 부분
        </front>
        <body>
                본문 내용
        </body>
        <back>
                부록, 색인, 뒷표지 등의 꼬리말 부분
        </back>
</text>
```

○ 합성적 형태

```
<text>
        <front>
                앞표지, 목차 등의 머리말 부분
        </front>
    <group>
                <text>
                <front>
                        책의 머리말 부분
                </front>
                <body>
                        <div>
                                본문 내용
                        </div>
                </body>
                </text>
        <back>
                부록, 색인, 뒷표지 등의 꼬리말 부분
        </back>
</text>
```

본문에서는 <front>와 <back> 태그가 설정되어 있다. 그러나 세종계획에서는 이들 태그는 거의 사용하지 않았다. 즉 세종계획 말뭉치가 국어의 언어

적 특성을 파악하려는 목적이 강하다. 모든 통계는 주로 텍스트의 본문만을 중심으로 처리하려는 입장을 취하고 있어서 빚어진 결과이다.

라. 기타 필요 태그

○ 반드시 표함되어야 할 태그
<head> : 제목, 부제목, 표 제목 등 모든 종류의 제목
<p> : 산문에서의 문단
<l>, <lg> : 운문에서의 시행과 시행군(연 등)

○ 희곡에 필요한 태그
<sp> 대사 부분, 대사자의 이름과 지문을 포함할 수 있다. 대사 자체는 <p>로 표시한다.
<speaker> 대사자의 이름 부분, 반드시 <sp>내에 있어야 한다.
<stage> 무대 지시 등 지문. <sp> 내 혹은 박에 있을 수 있다.

○ 판독 불가, 생략 부분 등에 대한 태그
<gap reason='판독 불가'>
<gap reason='그림 부분'>

○ 년도에 등에 대한 태그
<date value='1996-3'>1996. 3</date>

2.2. 북한의 형태

북한의 원시 말뭉치 구축 지침은 확인할 수 없다. 다만 남한에 전달된 북한의 말뭉치 견본에서 그 양상을 파악할 수 있다. 아래 (2)는 북한의 원시 말뭉치 견본이다.

(2) 북한 원시 말뭉치 견본

가. 헤더의 구조

```
<Header>
<PublStat>
<title>청년문학</title>
<number>주체92(2003)년 1호</number>
<date value=03-1-5></date>
<publisher>문학예술출판사</publisher></PublStat>
<EditStat>
<name>방정호</name>
<date>03-8-14</date>
</EditStat>
</Header>
```

북한의 헤더 정보에는 서명, 출판년도, 출판사, 저자, 말뭉치 구축 연도에 해당하는 정보만 담고 있다.

나. 머리말의 구조

```
<front>
<p>불세출의 령장 김정일장군님을 높이 모신 민족의 영광 만방에 빛내이자!</p>
<list>
<head>차례</head>
<item>달려 가자 강성대국의 새날을 향해(시)..... 송정우(4)</item>
<item>중지 당한 구내길 보수공사 (금수산기념궁전전설).....(5)</item>
<item>물음에 대한 생각(수필).....최국선(6)</item>
<item>신년사가 없는 사연(혁명전설).....(7)</item>
<item>한밤중에 온 전화(혁명일화).....(8)</item>
<item>동력(단편소설).....윤경찬(9)</item>

===중략===

<item>《전승 50돐기념창작경기》를 진행함에 대하여 ..... (64)</item>
</list>
</front>
```

<front> 태그는 머리말 부분에 해당하는 내용을 담고 있다. 일반적으로 앞표지, 서문, 목차 등을 포함한다. 남한에서는 이 태그는 별도로 기술하지 않았다.

다. 본문의 구조

```
<content>
<poem>
<head>달려 가자 강성대국의 새날을 향해</head>
<front><author>송정우</author></front>
<lg>
<l>밝아 온 새해의 언덕우에서</l>

===중략===

</lg>
</poem>
<text>
<front>금수산기념궁전전설</front>
<head>중지 당한 구내길 보수공사</head>
<body>
<p>주체80(1991)년이니까 금수산의사당을 지은지 10년하고도 여러해가 지났을 때였다.</p>
<p>우리 인민들이 온갖 성의를 다 기울여 정성껏 지은 집이였지만 거듭되는 비바람과 눈서리가 남겨 놓은 세월의 흔적이 여기저기 눈에 뜨이게 나타났다.</p>
<p>그중에서도 구내길에 그 흔적이 더 우심했다.</p>
<p>10여년세월에 로반이 어떤데는 심하게 내려 앉고 어떤데는 덜 내려 앉으면서 여기저기 균렬이 생기기 시작했던것이다.</p>
<p>균렬은 날을 따라 더 커졌고 그 짬에서 풀까지 돋아 나와 일군들을 당황하게 했다.</p>

===중략===

<p>그래서 우리 인민들의 가슴은 더더욱 쓰리고 아픈 것이다.
```

```
아, 수령님!</p>
  </body>
  </text>
  <text>
  <front>수필</front>
  <head>물음에 대한 생각</head>
  <front><author>최국선</author></front>
  <body>
  <p>이 세상엔 얼마나 많은 물음들이 있는가.</p>
  <p>존경과 진정이 담긴 부모와 자식간의 물음, 사랑과 우정이 비낀 친우들간의 물음, 증오와 조소가 어린 적아간의 물음...</p>
  <p>단순히 그 어떤 기대나 의문을 안고 사회생활에서 리용되던 물음이 얼마전에  나에게는 가장 뜨겁고도 숭고한 감정으로 받아 안았던 때가 있었다.</p>
  <p>내가 몇해만에 처음으로 고향에 갔을 때였다.</p>

  ===중략===

  <p>《우리 장군님 제일이시다. 위대한 사랑으로 천만의 심장을 움직이시고 주체혁명위업을 승리에로 이끄시는 우리 장군님 계시여 우리는 언제나 이긴다!》</p>
  </body>
  </text>
  </content>
```

본문 전체 내용을 표시하는 <content> 태그 안에 <poem>과 <text>를 구분하고 있으며, <front> 태그를 사용하여 장르를 구분하고 있다.

그러나 전체적으로 북한의 원시 말뭉치도 남한과 같은 TEI 라이트 방식을 취하고 있다.

2.3. 통합을 위한 방안

이상의 비교를 통해서 새롭게 구축할 원시 말뭉치의 구조는 헤더 파일에

말뭉치를 활용하는데 필요한 서지적인 정보만을 담았으면 한다. 즉 도서명, 저자, 출판년도, 출판사, 책의 성격을 규정하는 장르가 포함되면 된다. 머리말 <front>과 꼬리말 <back>에 포함되는 정보는 실제적으로 자료를 처리하는 과정에서는 제외시키는 것이 온당하다. 즉 머리말에 주로 포함되는 목차는 본문의 해당 내용 앞에 소제목 <title>으로 제공되기 때문에 통계 처리를 할 경우 이중적으로 계산이 된다.

다만, 종합지나 신문의 경우에는 <content> 안에 포함된 각 작품들에 대한 장르가 명시되어야 한다. 따라서 헤더 파일에서 단행본 전체의 장르를 규정한 것과는 달리 하위 장르를 규정할 수 있는 별도의 태그 <genre>가 필요하다. 그리고 언어학 연구자들을 위해서는 페이지 정보가 필요하다. 이 정보는 출처의 정확성을 담보하기 위해서 아주 중요한 요소이다. 따라서 앞으로 구축될 말뭉치에서는 <page=000> 형태의 태그를 정확히 페이지가 구분되는 위치에 삽입하는 것이 중요하다. 다만 데이터베이스로 불러들일 경우에는 어절의 중간에 있는 페이지 정보는 어절의 뒤로 옮기는 것이 바람직하다.

3. 남북한 주석 말뭉치 비교

주석 말뭉치는 <표 1>에 제시된 바와 같이 남한에서 167만 어절이 구축되어 있다. 그러나 북한의 경우는 정확히 알 수 없고, 국립국어원에 전달된 자료에서 그 구축 형식을 알 수 있다.

3.1. 분석표지

남한의 주석 말뭉치 구축 지침과 어절 분석표지는 아래 (3)과 같지만, 북한의 분석 지침은 확인할 수는 없다. 다만 제공된 자료를 통해 그 차이를 확인하기로 한다.

(3) 구축 지침의 원칙

○ 한 어절은 반드시 하나 이상의 주석 표지를 부여 받는 것을 원칙으로 한다. 다만, 아래의 지침에서 주석 표지를 부여하지 않는 것으로 규정한 것은 예외로 한다.

○ 띄어쓰기는 북한어 자료에 충실히 따른다.

○ 어절의 분석은 조선말대사전에 표제항으로 등재되어 있는 경우에는 더 분석하지 않고 표제항의 형태를 기본으로 따른다.

○ 어절의 분석은 표면형을 토대로 하며, 원칙적으로 태그를 제거했을 때 원어절로 복원 가능하도록 한다.

○ 축약형과 생략형은 원어절로 복원이 정확하게 가능하도록 한다.

그러나 주석을 위한 남북한의 분석표지는 <표 2, 3>과 같다.

<표 2> 남한 형태 주석 말뭉치 분석표지

대분류	소분류	세분류	주석 안	비고
(1) 체언	명사NN	일반명사NNG	NNG	
		고유명사NNP	NNP	
		의존명사NNB	일 NN 반 BB	
			단 NN 위 BC	
	대명사NP		NP	
	수사NR		NR	
(2) 용언	동사VV		VV	
	형용사VA		VA	
	보조용언VX		VX	
	지정사VC	긍정지정사VCP	VCP	
		부정지정사VCN	VCN	
(3) 수식언	관형사MM		MM	
	부사MA	일반부사MAG	MAG	
		접속부사MAJ	MAJ	

(4) 독립언	감탄사IC		IC	
(5) 관계언	격조사JK	주격조사JKS	JKS	
		보격조사JKC	JKC	
		관형격조사JKG	JKG	
		목적격조사JKO	JKO	
		부사격조사JKB	JKB	
		호격조사JKV	JKV	
		인용격조사JKQ	JKQ	
	보조사JX		JX	
	접속조사JC		JC	
(6) 의존 형태	어미E	선어말어미EP	시제 EPT	
			주체 존대 EPH	
		종결어미EF	EF	
		연결어미EC	EC	
		명사형전성어미ETN	ETN	
		관형형전성어미ETM	ETM	
	접두사XP		XP	
	접미사 XS	명사파생접미사XSN	XSN	
		동사파생접미사XSV	XSV	
		형용사파생접미사XSA	XSA	
	어근XR	어기	XR	
		어원어근	XB	
(7) 기호	마침표,물음표, 느낌표	SF	SF	
	쉼표, 가운뎃점, 콜론, 빗금	SP	SP	
	따옴표, 괄호표, 줄표	SS	SS	
	줄임표	SE	SE	
	붙임표(물결, 숨김, 빠짐)	SO	SO	
	외국어	SL	SL	

	한자	SH	SH	
	기타 기호(논리 수학기호, 화폐 기호) 등)	SW	SW	
	명사추정범주	NF	NF	
	용언추정범주	NV	NV	
	숫자	SN	SN	
	분석불능범주	NA	NA	

<표 3> 북한 형태 주석 말뭉치 분석표지

대분류	소분류	세분류	주석안	비고
체언	명사	일반	N	
		고유	Np	
	수사		U	
	대명사		D	
	불완		n	
용언	동사		V	
	형용사		X	
	보조동사		Vb	
	보조형용사		Xb	
수식언	관형		K	
	부사		F	
독립언	감동사		G	
토	체언토	격토	k	
		도움토	z	
		복수토	u	
	용언토	종결토	w	
		접속토	l	
		규정토	q	
		꾸밈토	g	
		시칭토	t	

		존칭토	j	
		상토	x	
	전성토	체언전성	bn	
		용언전성	bv	
접사	접두		Q	
	접미		H	
성어			C	
기호			S	
분석불능			BN	

<표 2>와 <표 3>을 비교해 보면, 먼저 남북한의 규범 문법의 차이로 인하여 기본적인 품사 분류 체계에서 차이가 있음을 알 수 있다. 이 문제는 규범의 통일이 이루어지기 전에는 어찌할 수 없는 부분이다. 다만 남한에서 북한 언어자료 말뭉치를 구축하면서 남한의 규범으로 분석할 것인가 아니면 북한의 규범으로 분석할 것인가 하는 문제가 제기될 수 있다. 남북 두 체제에서 적용하고 있는 언어 규범의 차이를 인정한다면 원칙적으로 북한의 규범대로 구축하는 것이 바람직할 것이다. 그러나 구축된 자료를 이용하는 측면에서 본다면 남한에서 구축된 자료는 남한에서 이용하기 때문에 남한의 다른 말뭉치 분석표지와 일치시키는 것이 활용에 편리할 것이다.[10] 21세기 세종계획에서는 후자의 입장을 취해 말뭉치 분석 표지를 설정했다. 다만 어휘별로는 북한의 조선말대사전에 올림말로 등재된 경우는 분석을 하지 않았다.

체언의 경우 의존명사의 분석에서 차이를 보인다. 남한에서는 일반의존명사와 단위성의존명사를 구분하고 있지만 북한에서는 불완전명사로만 분석하

10) 21세기 세종계획에서 구축된 모든 말뭉치에 가능하면 동일한 분석 표지를 사용하려 했다. 그러나 결과는 그렇지 못했다. 이 문제는 자료를 활용하는 과정에서 그 심각성이 노출되었다. 따라서 앞으로 구축되는 말뭉치는 대분류, 소분류, 세분류에서 각각 다른 단계를 택할 수는 있지만 다른 분석 표지를 사용하지 않도록 하는 노력이 필요하다.

고 있다. 이런 경우 남한 분석 자료를 북한 분석 자료로 변환할 경우에는 문제가 없다. 그러나 북한 자료를 남한 자료로 변환할 경우에는 이 부분을 보완해야 한다. 물론 의존명사로 통합해서 처리한다면 문제는 없다.

용언의 경우 남한에서는 보조용언을 설정하고 지정사를 분석했다. 반면에 북한에서는 보조용언을 보조동사, 보조형용사로 분석했다. 그러나 이 분석은 문제가 되지 않는다. 남한의 분석을 따를 경우 보조용언 앞에 오는 본용언이 동사이면 뒤에 오는 보조용언이 보조동사이고, 형용사이면 보조형용사이기 때문에 자동적으로 변환하는데 문제가 없다. 지정사의 경우도 '이다, 아니다'만을 대상으로 하기 때문에 자동으로 변환하더라도 큰 문제는 없을 것으로 생각된다.

수식언의 경우 남한에서 부사를 일반부사와 접속부사로 구분했다. 이 문제는 의존명사와 비슷한 문제로 남한의 접속부사를 축출해서 동일한 형태의 북한 부사를 자동으로 변환할 수 있을 것이다.

이상의 문제는 단순한 기계적인 처리로 쉽게 통합이 이루어질 수 있는 문제이다. 분석 표지의 분류 단계에서 차이가 있는 경우 상위 분류 단계를 채택한 자료를 하위 분류 단계를 채택한 체계로 전환하려고 하면 수작업이 필요한 부분이 있다. 즉 세부적으로 분석된 남한의 자료를 북한의 자료로 통합하는 데는 문제가 없지만 반대의 경우는 수작업이 필요하다. 그러나 이러한 분류 단계의 깊이 문제는 연구자들의 요구에 따라 차이가 있을 수 있음을 인정해야 한다. 그러나 관계언의 경우에는 이렇게 해결될 수 있는 문제가 아니다. 남한에서는 조사를 품사로 인정하고 있는 반면 북한에서는 조사를 의존형태인 토의 일종으로 처리하고 있기 때문에 분석 표지의 통합에서도 쉬운 문제가 아니다.

먼저 조사의 경우 남한에서는 관계언으로 보고 격조사, 보조사, 접속조사로 분석하지만, 북한에서는 체언토 중에 격토, 도움토, 복수토로 분석하고 있다. 여기서 일차적으로 문제가 되는 것은 남한의 접속조사와 북한의 복수토이다. 접속조사는 북한의 격토에 포함시키면 되기 때문에 문제가 없을 수 있다. 그러나 북한의 복수토는 남한의 복수접미사 '-들'이기에 범주의 문제가

제기될 수 있다. 다음은 분석 지침에서 남한의 격조사는 주격, 보격, 관형격, 목적격, 부사격, 호격, 인용격 조사로 하위 분류되어 분석되었지만, 북한은 격토로만 분석되어 있어 북한의 주격, 대격, 속격, 여격, 위격, 조격, 구격, 호격으로 하위 분류된 격조사는 분석되지 않았다. 따라서 북한에서 구축된 자료를 남한 자료로 변환할 경우 별도의 분석 작업이 이루어져야 한다.

어미도 차이가 많이 난다. 남한의 어미는 북한에서 용언토와 전성토로 분석되고 있다. 물론 한 단계 더 깊이 분석할 경우에 북한의 종결토, 접속토, 시칭토, 존칭토는 종결어미, 연결어미, 시제, 주체존대에 해당고, 규정토는 관형사형어미에 해당되고, 꾸밈토는 부사형어미에 해당한다. 상토는 남한에서는 분석하지 않고 있다. 이러한 차이는 단순히 용어의 차이만아 아니라 구체적인 분석에서 차이가 있을 수 있어 자동적으로 처리하려면 면밀한 검토가 요구된다.

접미사의 경우도 동일하다. 남한에서는 파생접미사를 명사, 동사, 형용사 파생접미사로 하위 분석하고 있지만 북한에서는 분석하지 않고 있다.

이상에서 살펴본 바와 같이 문법적인 규범의 차이가 없는 경우는 분석 단계의 깊이에 차이가 있어 세부적인 분석이 필요한 경우에는 추가적인 작업이 필요하지만 그렇지 않은 경우에는 상위 분석 표지로 통합하면 된다. 그러나 규범의 차이가 있는 경우는 쉽게 자동으로 변환이 어려울 것이다. 다만 분석된 각각의 자료를 정리하여 변환 테이블을 만들어서 변환한다면 많은 부분에서 변환의 일치점을 찾을 수 있을 것이다.

3.2. 분석 형태

남한과 북한의 형태 주석 말뭉치 결과물을 제시하면 아래 (4), (5)와 같다.

(4) 남한의 형태 주석 말뭉치 예

033118	우리	우리[NP]
033119	혁명은	혁명[NN]+은[J]

033120	천만자루의	천만[NR]+자루[NC]+의[J]
033121	창검을	창검[NN]+을[J]
033122	대신할만한	대신하[VV]+ㄹ [ETM]+만하[VX]+ㄴ [ETM]
033123	훌륭한	훌륭하[VA]+ㄴ [ETM]
033124	시를	시[NN]+를[J]
033125	더	더[MA]
033126	많이	많이[MA]
033127	창작해	창작하[VV]+어[EC]
033128	낼 것을	내[VV]+ㄹ [ETM]+것[NB]+을[J]
033129	절실히	절실히[MA]
033130	요구하고있다	요구하[VV]+고[EC]+있[VA]+다[EF]
033131	.	./SF

(4)에 제시된 바와 같이 남한의 형태 주석 말뭉치의 구조는 전체 어절 번호, 원 어절, 형태 주석으로 구성되어 있다. 원 어절은 북한의 조선말대사전 올림말을 기준으로 하고, 원문의 띄어쓰기 단위를 그대로 준수했다. 형태 주석은 분석된 형태소에 분석 표지를 부여하고 '+'로 형태소들을 연결시키고 있다.

(5) 북한의 형태 주석 말뭉치 예

위대한	[형(위대하다2)규(ㄴ)]	1
령도자	[명(령도자)]	2
김정일동지께서	[고(김정일)명(동지1)격(께서)]	3
사회와	[명(사회1)격(와)]	4
집단을	[명(집단)격(을)]	5
위하여	[동(위하다1)이(여)]	6
좋은	[형(좋다)규(ㄴ)]	7
일을	[명(일1)격(을)]	8
한	[동(하다)규(ㄴ)]	9
일군들과	[명(일군1)복(들)격(과)]	10
근로자들에게	[명(근로자)복(들)격(에게)]	11
감사를		12

보내시였다	[동(보내다)존(시)시(였)맺(다)]	13
===중략===		
온	[관(온1)]	95
나라	[명(나라)]	96
인민이	[명(인민)격(이)]	97
경애하는	[동(경애하다)규(는)]	98
장군님을	[명(장군님)격(을)]	99
높이	[부(높이2)]	100
모시고	[동(모시다)이(고)]	101
하나의	[수(하나1)격(의)]	102
대가정을 [	명(대가정)격(을)]	103
이루고	[동(이루다)이(고)]	104
화목하게	[형(화목하다)이(게)]	105
살아 가는	[동(살아가다)규(는)]	106
우리	[대(우리3)]	107
식	[명(식1)]	108
사회주의의	[명(사회주의)격(의)]	109
참모습을	[명(참모습)격(을)]	110
잘	[부(잘2)]	111
보여	[동(보이다)이(여)]	112
줄수 있게	[동(주다)토(ㄹ1)명(수1) 동(있다)이(게)]	113
하였다	[동(하다)시(였)맺(다)]	114

(5)에 제시된 북한의 형태 주석 말뭉치는 (4)에 제시된 남한의 형태 주석 말뭉치와 차이가 있다. 먼저 형식적인 면에서 보면 원 어절, 행태 주석, 전체 어절 번호 순서로 구성되어 남한과 차이를 보이고 있다. 그러나 이 차이는 전혀 문제가 되지 않는다. 반면에 원 어절에서는 중요한 차이가 있다. '살아 가는, 줄수 있게'와 같이 띄어쓰기가 된 어형들이 북한에서는 하나의 항목으로 제시되고 있다는 점이다. '살아 가는'의 경우는 '살아가다'의 어원적 어근을 분석하는 2000년도의 표기법 개정안에 따라 띄어 쓰게 된 예이다. 이를 위해서 남한에서는 어원적 어근 XB를 설정하고 있다.[11] '줄수 있게'의 경우

도 북한에서는 하나의 항목으로 제시되고 있다. 그러나 이 경우는 특별한 설명을 하기 어려운 예이다.

형태 주석의 경우는 '보내시였다 [동(보내다)존(시)시(였)맺(다)]'와 같이 제시하고 있다. 이 분석을 남한의 방법에 의해 재구성한다면 '보내[동]+시[존]+였[시]+다[맺]'처럼 쓸 수 있다. [동]은 동사의 원형을 제시하는 것이 아니라 동사의 자립형을, [시]는 시칭토인 시제 선어말어미, [존]은 존칭토인 존칭 선어말어미, [맺]는 종결토인 종결어미를 표시하고 있다. 이렇게 형식적 차이는 쉽게 변환이 가능하다. 그러나 분석 표지의 차이가 극복되지 않는다면 해결이 어려울 것이다.

3.3. 통일을 위한 방안

형태주석 말뭉치의 개발의 통일을 위해서는 우선적으로 남북한의 어문 규범이 통일되어야 한다. 그러나 이 작업은 두 체제의 통일이 이루어지기 전에는 달성될 수 있는 목표가 아니다. 따라서 제한적이지만 Korean 컴퓨터 처리 국제 학술 대회에서 이룬 성과처럼 컴퓨터 내부에서 활용되는 자료라는 단서를 달고, 통일 가능한 부분부터 통일시켜 나가는 일이 중요하리라 생각된다.

형태주석 말뭉치를 개발하면서 제시할 수 있는 통일 방안은 띄어쓰기된 어절은 하나의 어절로 분석한다. 주석안으로 사용되는 분석 표지는 남북 상호간에 소분류와 세분류의 분석 결과를 기계적으로 처리할 수 있도록 배려한다는 정도의 원칙을 정하는 것이 중요하리라 생각한다.

---

11) 21세기 세종계획에서는 북한의 정서법 규정이 바뀐 2000년 이후 자료를 말뭉치로 구축하면서는 어원적 어근을 위한 분석 표지를 설정했다.

4. 남북한 구문분석 말뭉치 비교

구문분석 말뭉치에 대한 남북한의 비교는 쉽지 않다. 남한의 주석 지침은 21세기 세종계획에서 상세하게 규정하고 있지만, 북한의 주석 지침을 확인할 수 없어 비교하기가 용이하지 않다. 그래서 이 장에서는 실제로 구축된 구문분석 말뭉치의 예를 제시하는 것으로 그 차이만을 드러내려 한다.

(6) 남한 구문분석 말뭉치 예

```
 <text>
 <body>
; 1991/01/05 19
(NP          (NP 1991/SN + //SP + 01/SN + //SP + 05/SN)
             (NP 19/SN))

; 새해 주제 '다시 뛰자' 기획시리즈 /
(NP          (NP        (VP        (NP        (NP 새해/NNG)
                                               (NP 주제/NNG))
                                    (VP        (S- +'/SS)
                                    (VP        (VP        (AP 다시/MAG)
                                                          (VP 뛰/VV + 자/EC))
                                               (S- +'/SS))))
                        (NP 기획/NNG + 시리즈/NNG))
             (S- +//SP))

; 일할 사람이 없다:2
(S           (NP_SBJ    (VP_MOD 일/NNG + 하/XSV + ㄹ/ETM)
                        (NP_SBJ 사람/NNG + 이/JKS))
             (VP 없/VA + 다/EF + :/SP + 2/SN))

; 젊은 일손 찾기 "별따기" /
(S           (S         (VP_SBJ    (NP_OBJ    (VP_MOD 젊/VA + 은/ETM)
                                               (NP_OBJ 일손/NNG))
                                    (VP_SBJ 찾/VV + 기/ETN))
                        (VP "/SS + 별/NNG + 따/VV + 기/ETN + "/SS))
             (S- +//SP))
```

; 건설회사마다 구인 아우성 /
(NP (NP (NP_AJT 건설/NNG + 회사/NNG + 마다/JX)
(NP (NP 구인/NNG)
(NP 아우성/NNG)))
(S- +//SP))

; 일당 7만원에도 못 구해 /
(S (VP (NP_AJT (NP 일당/NNG)
(NP_AJT 7/SN + 만/NR + 원/NNG + 에/JKB + 도/JX))
(VP (AP 못/MAG)
(VP 구하/VV + 아/EC)))
(S- +//SP))

; 공사장엔 40~50대 일색 /
(NP (NP_AJT 공사장/NNG + 에/JKB + ㄴ/JX)
(NP (NP 40/SN + ~/SO + 50/SN + 대/NNB)
(NP 일색/NNG)))
(S- +//SP))

; 임금과 상관없이 "궂은 일만은 안한다" /
(VP (VP (AP (NP_AJT 임금/NNG + 과/JKB)
(AP 상관/NNG + 없이/MAG))
(VP (S- +"/SS)
(VP (VP (NP_OBJ (VP_MOD 궂/VA + 은/ETM)
(NP_OBJ 일/NNG + 만/JX + 은/JX)
(VP 안/MAG + 하/VV + ㄴ다/EC))
(S- +"/SS))))
(S- +//SP))
</body>
</text>
</tei.2>

(7) 북한 구문분석 말뭉치 예

2,VP_LNK(NP_OBJ(BP_ATT(몸소/D)
NP_OBJ(시동/N+단추/N+를/k))
VP^LNK(VP_LNK(누르/V+어/e)
VP_LNK(주/V+시/j+며/e)))

3,SS(NP_ATT(NP_MOD(NP_MOD(주체67/U+[/S+1978/U+]/S+년/n)
NP(4/U+월/n))
NP(KP_MOD(어느/K)
NP(날/N)))
NP_SBJ(VP_MOD(경애하/V+는/cs)
NP_SBJ(장군님/N+께서/k+는/i))
NP_OBJ(NP_MOD(VP_MOD(NP_ATT(NP_MOD(KP_MOD(어느/K)
NP_MOD(한/U))
NP_ATT(신문사/N+에/k))
BP_ATT(새로/D)
VP_MOD(꾸리/V+ㄴ/cp))
NP_MOD(인쇄/N+공장/N+의/k))
NP_OBJ(륜전/N+직장/N+을/k))
VP_PRD(돌/V+어/e
보/V+시/j+었/q+다/m+./S))

구문분석 말뭉치의 통합을 위해서는 우선적으로 문장 분석 단위의 통일이 이루어져야 한다. 이는 어문규범의 통일이 전제되는 일이다. 따라서 이 장에서는 분석의 실제 예만을 제시했다.

5. 결론 및 제언

이 장에서는 구축된 말뭉치의 통합 운영에 관한 사항을 제안하려 한다. 그동안 남한에서 구축되었던 원시 말뭉치나 형태주석 말뭉치 나아가 구문분석 말뭉치는 하나로 통합되어 운영되지 못했다. 따라서 각각의 말뭉치가 가지고 있는 문제를 극복하기 위한 대안이 제시되지 못했다. 이 장에서는 원시 말뭉치와 형태주석 말뭉치를 통합해서 운영할 것을 전제로 몇 가지 제안을 하려 한다.

1. 헤더 파일 정보 테이블
   ○ 원시 말뭉치의 헤더 파일에는 도서명, 저자, 출판년도, 출판사, 장르 정보가 포함되어야 한다.
2. 원시 말뭉치 테이블
   ○ 문장 분리기를 개발하여 문장 단위로 구분하여 문장 고유번호를 부여하여야 한다. 이때 고려할 사항은 단순한 종결부호로 문장을 분리해서는 안 된다.
   ○ 문장을 분리하면서 운문을 표시하고 있는 <lg> <l> 태그는 별도의 식별 부호를 부여한다.
   ○ 구어나 준구어 자료의 <speaker> 태그도 별도의 식별 부호를 부여한다.
   ○ 부여된 페이지 번호 정보를 유지한다.
   ○ <title>, <p> 태그의 정보를 유지한다.
3. 형태주석 말뭉치 테이블
   ○ 일련번호, 원 어절, 풀어쓰기, 형태주석과 같이 4개의 필드로 구성한다.
   ○ 현재까지 구축된 말뭉치 모두에 대하여 어느 정도의 오류를 인정하면서 형태소 분석기로 선 처리한다.
   ○ 풀어쓰기 필드는 음소단위 검색을 위해서 필요하기에, 풀어쓰기 도구를 개발하여 선 처리한다.
4. 말뭉치 통합
   ○ 위 작업이 마무리되면, 모든 자료를 MS-SQL 서버나 기타 대용량 데이터를 처리할 수 있는 데이터베이스 프로그램에 탑재한다.
   ○ 탑재된 이후에 후 처리 과정을 통해 형태주석 필드를 수정한다.

위에 제안된 내용이 완벽한 말뭉치의 통합 방법은 아니다. 그러나 각각의 자료로 존재하고 있는 말뭉치를 우선 통합하여 활용함으로써 각각의 말뭉치가 가지고 있는 문제점을 파악할 수 있고, 그 개선책도 마련될 수 있을 것이다. 나아가 통합된 말뭉치를 통해 다양한 연구가 가능하고, 의미주석 말뭉치로 확장도 쉽게 이루어질 것으로 생각한다.

참고문헌

강승식(2011), 『한국어 정보화 세부 실행 계획 수립』, 국립국어원 보고서.

과학.백과사전출판사(1979), 『조선문화어문법』, 과학.백과사전출판사.

국립국어원(2007), 『21세기 세종계획 백서』, 국립국어원.

국립국어원(2015), 『2015년 시스템 고도화 제한 요청서』, 국립국어원.

권재일(2004), 「말뭉치 자료 정보 부착과 남북 문법 통합 방법」, 민족어 유산의 수집정리와 고유어 체계의 발전 풍부화에 관한 학술모임 발표 자료집.

권재일(2005), 「구어 말뭉치 구축과 형태·통사 정보」, 민족어 어휘구성의 변화와 통일적 발전에 관한 국제 학술회의 발표 자료집.

권종성(2004), 「균형 코퍼스에 대한 새로운 리해와 그 제작에서 나서는 몇가지 문제」, 민족어 유산의 수집정리와 고유어 체계의 발전 풍부화에 관한 학술모임 발표 자료집.

김성근(2004), 「민족어유산으로서의 조선어방언의 조사원칙과 방법」, 민족어 유산의 수집정리와 고유어 체계의 발전 풍부화에 관한 학술모임 발표 자료집.

김홍규 외(1998), 「현대국어 말뭉치 개발 및 세종 말뭉치 통합·표준화 연구」, 『21세기 세종계획 국어 기초자료 구축 보고서』, 문화관광부.

김홍규 외(1999), 「현대국어 기초 말뭉치 및 형태소 분석 말뭉치 개발」, 『21세기 세종계획 국어 기초자료 구축 보고서』, 문화관광부.

박경래(2005), 「현지 방언 조사 작업의 내용과 문제점」, 민족어 어휘구성의 변화와 통일적 발전에 관한 국제 학술회의 발표 자료집.

서상규 외(1999), 「현대국어 구어 말뭉치 개발」, 『21세기 세종계획 국어 기초자료 구축 보고서』, 문화관광부.

소강춘 외(1999), 「북한 및 해외 한국어 말뭉치 개발」, 『21세기 세종계획 국어 기초자료 구축 보고서』, 문화관광부.

소강춘 외(2012), 「『개방형 한국어 지식 대사전』 신어 분과의 표제어 선

정과 그 실제」, 『한국사전학』 20.
소강춘(2004), 「민족어 말뭉치 지도의 작성」, 민족어 유산의 수집정리와 고유어 체계의 발전 풍부화에 관한 학술모임 발표 자료집.
소강춘(2005), 「남북 음성 자료 정리를 위한 프로그램 개발」, 민족어 어휘구성의 변화와 통일적 발전에 관한 국제 학술회의 발표 자료집.
오철만(2004), 「방언조사 어휘선정에서 제기되는 문제」, 민족어 유산의 수집정리와 고유어 체계의 발전 풍부화에 관한 학술모임 발표 자료집.
이기갑(2005), 「방언 문법의 정밀 조사」, 민족어 어휘구성의 변화와 통일적 발전에 관한 국제 학술회의 발표 자료집.
이상규(2004), 「지역어 조사 항목의 설정 원칙과 방향」, 민족어 유산의 수집정리와 고유어 체계의 발전 풍부화에 관한 학술모임 발표 자료집.
이승재(2004), 「말뭉치 활용 도구의 개발」, 민족어 유산의 수집정리와 고유어 체계의 발전 풍부화에 관한 학술모임 발표 자료집.
이승재(2012), 21세기형 사전 『개방형 한국어 지식 대사전』, 『한국사전학』 20.
임홍빈 외(1998), 「한국어 정보처리를 위한 어절 분석 표지의 표준화 연구」, 『21세기 세종계획 국어 기초자료 구축 보고서』, 문화관광부.
임홍빈 외(1999), 「구문 분석 방법론 및 표지의 표준안 연구」, 『21세기 세종계획 국어 기초자료 구축 보고서』, 문화관광부.
진용옥(2006), 「한국어 정보학과 국어학」, 『세종학연구』 14, 세종대왕기념사업회.
최명옥(2004), 「지역어 조사의 원칙과 조사 방법」, 민족어 유산의 수집정리와 고유어 체계의 발전 풍부화에 관한 학술모임 발표 자료집.
홍석희(2005), 「컴퓨터에 의한 방언의 자료기지화에서 나서는 기술적 문제」, 민족어 어휘구성의 변화와 통일적 발전에 관한 국제 학술회의 발표 자료집.
홍윤표(2001), 「남북한 국어정보화 표준화를 위한 과제」, 한국정보과학회 언어공학연구회 학술발표 논문집 10, 한국정보과학회 언어공학연구회.

홍윤표(2004),「민족어 방언 검색 시스템 개발」, 민족어 유산의 수집정리와 고유어 체계의 발전 풍부화에 관한 학술모임 발표 자료집.
홍윤표(2012), 『국어 정보학』, 태학사.
홍윤표(2015),「통일을 대비하기 위한 북한어 연구 방향」, 상해 복단대학 국제학술대회 발표 자료집.

【Abstracts】

A method to integrate corpus in North and South Korea

So Kang-chun

In this study, I have a clear grasp of the present state of the electronic Korean corpus that was built by North and South Korea. Through that corpus, I am able to clarify the distinction between the North Korea corpus and its counterpart in the South. I am also able to find a way to close the gaps. Subsequently, I will search for a method to integrate corpus in North and South Korea.

Key words : computerization, letter corpus, phonecorpus, image corpus, tag set

이 논문은 2016년 9월 29일에 투고되었으며, 2016년 11월 4일에 심사 완료되어 2016년 11월 9일에 게재가 확정되었음.