한국구비문학대계 전자텍스트와 언어 코퍼스적 특성

김병선 (한국학중앙연구원)

1. 첫머리에

한국구비문학대계(이하 '구비대계')는 한국의 구비 전승 문학작품을 수집하여 정리한 한민족 근현대의 문화적 자원이다. 1980년대에 모두 85권으로 출판된 바 있는 1차 조사자료집은, 비록 한반도 남쪽의 삼분의 일 지역에 그치는 것이었지만, 그것만으로도 해방 이후 한민족의 대규모 문화적 업적으로 치부될 정도의 의미를 가지고 있었다. 그러다가 지난 10년간의 제2차 사업(한국구비문학대계 개정증보사업)을 통해 나머지 삼분의 이의 미조사지역에 대한 조사사업을 마무리하게 됨으로써 이제 적어도 한반도 남쪽의 구비 전승 자료는 총집성되었다고 할 수 있다. 이제는 이 자료를 어떻게 활용하여 한국의 문화적, 학문적 역량을 키울 것인가의 과제가 남아 있다. 이 발표는 이러한 과제의 하나로서 구비대계의 특성을 재점검하여 재구성함으로써 이를 학문적 연구의 대상으로 활용할 수 있는 방안에 관심을 두고 있다. 특히 이 구비대계가 언어적 자원이라는 점을 중시하여, 이를 언어적 코퍼스로 활용하는 방안을 논의해 보고자 한다.[1]

---

[1] 구비대계의 언어적 코퍼스로의 활용에 대해서는 "한국구비문학대계의 보급과 지식정보화의 확산을 위한 연구"(김병선 외, 한국학중앙연구원 정책과제 결과보고서, 2017)에서 중요 과제로 제시된 바 있다.

2. 한국학에서의 구비문학 대계의 의의

구비문학 작품은 일반 민중(Am-Haaretz)의 생활 감정이 진솔하게 담겨 있는 문화적 자원이다. 사대부들의 문학이 문자로 고착되어 전승된 반면 민중의 문학은 구술로 전승되고 있으며 시대적 환경과 개인적 처지 등에 따라 많은 변이형을 보인다. 구비대계는 말하자면 한민족에 유전되는 기본적 정서를 담고 있는 동시에 20세기 후반의(1차 조사), 그리고 21세기 초반(2차 조사)의 한민족의 생활과 경험을 생생하게 담고 있는 텍스트다.

이 시기의 한민족의 생활감정과 경험은 그 전승성에 있어서는 전시대로부터 유전된 것이면서 전통적인 문화적 담당층의 저작물이 포괄하지 못하는 일반 민중의 삶이 담겨 있다. 구비문학연구자들은 구비전승의 문학성에 관심을 두면서, 전승과 향유 계층의 삶의 의미를 천착한다든지, 비교문학의 관점에서 이를 들여다보기도 한다. 하지만 구비전승의 자원은 문학적 연구의 대상에 머무르지 않는다. 20세기 후반과 21세기 초반의 한국 민중의 삶과 사회, 정신과 태도라는 현상을 엿볼 수 있는 중요한 자원이 되기도 한다.[2]

한국문학에서 구비문학으로서 연구될 수 있다. 이제는 한반도 남쪽의 전지역을 포괄할 게 되었으므로 지역적, 지리적 특성을 반영한 연구 등에서 좋은 성과를 낼 수 있다고 본다. 다양한 변종들로부터 한국 설화의 원형을 탐구할 수도 있고, 이웃나라의 구비자료[3]

---

[2] 실제로 한국학중앙연구원에서는 공동연구과제로서 구비문학 자료에 나타나는 한국인의 심성 연구(연구책임자 이상훈 교수)가 2년간 수행된 바 있다.

[3] 향후에 구비문학 자료는 더블린코어(DC) 등의 포맷으로 인근 국가의 자료와 메타데이터를 공유

또는 전통적인 문헌설화들과의 비교 연구를 통해서 그 계통성을 탐색해 볼 수도 있을 것이다.

국어학 분야에서는 현재도 많은 연구자들이 이 자료에 관심을 가지고 있다. 특히 구비자료는 그 자체로서 지역어의 자원이기도 하므로, 방언연구나 언어의 사회성 연구에 활용될 수 있다.[4] 물론 지역 정보가 포함되어 있으므로 언어지도를 그리는 데에도 주요 자원이 된다. 음성 자료가 병렬되어 있으므로, 음성학과 음운학에서도 좋은 자료로 활용할 수 있다.

구비전승은 생활과 밀착되어 있으므로, 식문화, 주거문화, 의복문화 등에서도 많은 자원을 제공해 줄 수 있으며, 전통의학과 농경 관련 지식을 이 자료에서 얻으려고 하는 노력도 있다. 동물담과 그 상징성, 종교성과 신앙의 문제, 역사적 사건과 민중의 이해 등 수많은 분야의 다양한 연구적 관심을 만족시킬 수 있다고 본다.

본 발표자는 한국구비문학대계 1차 사업 결과물의 디지털화를 책임 맡았고(한국역사정보통합시스템 및 한국학전자도서관 사업), 현재는 2차 개정증보사업의 연구책임자 및 디지털편찬사업단의 책임을 맡고 있다. 이러한 입장에서 이 자료가 단지 아카이브로만 존재하거나, 데이터베이스의 자료로서 검색의 대상으로만 존재하거나, 특정 학문분야에 국한되지 않고, 보다 다양한 학문 분야(특히 한국학)에서 활용되기를 바란다.

할 예정이다.

[4] 몇몇 방언 연구자들이 구비대계의 텍스트로부터 방언 자료를 추출하고 있으며, 한국학중앙연구원 어문생활사연구소에서는 이러한 자료 요청을 꾸준히 받고 있다.

3. 언어적 코퍼스로서의 구비대계

여기서는 언어적 코퍼스로서의 구비문학 자료에 대해 논의해 보려 한다. 한국에서 대규모의 언어적 코퍼스는 국립국어원이 관리하고 있는 '21세기 세종계획'이 대표적이다. 이외에도 연세대학교의 코퍼스와 고려대학교의 말뭉치 등이 있으며, 한국어사전 편찬이나 언어학적 연구 및 인공지능을 위한 자연언어처리와 자동번역 등의 연구자원으로 활용되고 있다. 당초 구비대계의 일부 자료가 21세기 세종계획에서 구어 자료로 포함된 바 있다. 이 자료는 그러나 많이 활용되고 있지는 않는 것으로 알고 있다. 일부 자료가 시범적으로만 채택되었기 때문이다. 이제 구비문학대계 조사사업의 완성이 눈앞에 있으므로 이제는 전체 자료를 언어적 코퍼스로 활용하는 방안에 대해 고민할 때가 된 것이다. 특히 당초 종이책을 입력한 1차 구비대계의 자료와 조사 및 채록 단계에서부터 디지털 방식을 채택한 2차 구비대계의 자료 사이의 차이점도 중요한 고려사항이다. 여기서는 주로 1차 자료를 대상으로 논의하기로 한다. 1차 자료는 카세트테이프로 채록한 자료는 원고지에 전사하여 이를 책으로 출판하는 등 전과정이 아날로그 방식으로 처리되었으며, 1990년대 말부터 자료의 디지털 전환을 거쳐서 전자문서화되었다. 2차 자료는 조사의 첫 단계부터 100% 디지털 방식으로 이루어졌다. 비록 동일한 인터넷 검색 시스템으로 검색할 수 있지만 그 내부적 문서 구조에는 차이가 있다.

이러한 구비대계 1,2차 사업의 통합자료는 모두 5만 건이 넘는 설화와 민요, 무가 및

기타 자료로 구성되어 있다.[5] 일정한 조사원칙으로 수집된 자료가 음성 자료가 텍스트로 전사되었으며, 여기에 자료에 대한 메타데이터를 포함하고 있다. 이 자료는 일종의 XML 이라는 전자문서의 형식으로 데이터베이스화되어 있는 상태다. 애초에 언어적 코퍼스로 활용되는 것보다는 인터넷에서 구비자료의 검색과 읽기를 목표로 하였기 때문이다. 구비누리 사이트(gubi.aks.ac.kr)에서 이용자들은 제목과 본문을 대상으로 하는 검색을 통해서 구비문학의 단위 작품에 접근할 수 있다. and, or, not 등과 같은 연산자를 통해서 텍스트에 대한 불리언 검색까지 가능하다. 하지만 코퍼스 형식 자체로는 제공되지 않기 때문에 텍스트 자체에 대한 연구자의 접근은 원천적으로 차단되어 있다. 따라서 비록 전자문서의 형식으로 되어 있다고는 해도 현상태의 자료를 언어적 코퍼스로 바로 활용할 수는 없다.

이 발표에서는 비공개 상태로 있는 1차 자료의 구성 내용을 소개하고, 그 메타데이터와 문서의 구조를 소개한다.

4. 구비대계의 자료의 특징

먼저 한국 구비문학 집성으로서의 구비대계의 특성을 살펴본다.

(1) 조사지점의 포괄성과 균형성

[5] 2018년 8월 10일 현재 한국구비문학대계 스마트앱의 정보에 따르면 설화 24,534편, 민요 22,659편, 무가 3,977편 등으로 모두 51,170편에 이르고 있다. 조사 자료 중 아직 시스템에 탑재되지 않은 것들도 있어서 금년 10월 연구가 마무리될 때면 이보다 더 많은 편수가 될 것으로 예상된다.

1차, 2차 구비문학 조사사업은 한국의 기초자치단체(약 230개) 단위로 구획하고, 각 지자체의 읍면동 지역을 선별하여 조사하였다.[6] 기초자치단체를 기준으로 한다면 모든 자치단체가 다 조사 대상이 된 것이다. 아울러서 각 자치단체별로 책 한 권 분량의 구비자료를 수집함으로써 지역별 균형성도 고려하였다.

(2) 현장성과 지역성

제보자의 선정에 있어서는 가급적 조사 지역과 직접적으로 관련성을 연로한 분들을 대상으로 하였고, 가급적 제보자들이 인위적 조사가 아닌 평소의 '삶의 자리(Sitz im Leben)'에서의 연행이라는 느낌을 가지도록 하였다. 제보자 및 제보 상황 그리고 마을 정보를 수집하였고, 사진 자료도 첨부하였다. 또한 조사된 음성 자료를 전사할 때에도 표준어에 근거하여 채록하는 것이 아니라, 제보자의 발화를 존중하여 기본적인 음운화의 수준에서 전사하였다. 따라서 이 자료에는 각 지역의 지역어(방언)가 온전하게 보존되어 있으며, 문장으로 완성되지 못한 것도 그대로 수록하였다. 따라서 구비대계의 자료는 일정 시기의 구어체 담화를 집적한 언어적 자료라는 특성을 가지고 있다.

(4) 디지털 자료로서 전산 보존

아날로그적으로 수집되었던 1차 조사 자료도 2000년대 직전에 모두 디지털화를 하였고, 2차 조사에서는 수집 단계부터 디지털 방식으로 조사, 채록, 조직, 저장, 출판되었다. 여기에 포함된 자료의 형태는 다음과 같다.

---

[6] 1차 조사와 2차 조사 사이에는 행정구역 개편이 이루어진 곳도 적지 않다. 도농 통합이라든지 광역자치단체 편성에서 비롯된 현상이다. 2차 조사에서는 이에 유의하여 조사하였다.

1) 수집 현장의 발화를 채록한 음성 파일 또는 음성을 포함한 동영상 파일

2) 수집에 관련된 각종 정보 (시공간, 제보자 조사자 참여자 등의 정보)

3) 발화 내용을 문자로 전사한 텍스트 파일

4) 텍스트 파일과 음성 파일의 동기화 정보

5. 전사 텍스트 파일과 그 구조

이제 1차 조사 자료의 전사 텍스트 파일을 대상으로 그 구조와 특성을 살펴본다.

전사 텍스트 파일은 한국구비문학대계 출판본의 전자문서화 결과물이다. 이 텍스트 파일은 일반적인 XML 문서 양식을 취하고 있다. 이 문서는 종이책 출판본의 양식을 구현하는 것을 목표로 하여, 단위 구비 작품뿐 아니라, 책에 실린 대로 제목과 구연상황 및 어휘 주석 등을 그대로 포함하고 있다. 본문 중에는 표준어로의 표기가 병기된 곳도 있고, 한자 등을 ( )에 밝혀 추가하기도 하였다. 특히 이 문서에는 종이책의 페이지 구분(page break) 태그와 단락 구분(paragraph break) 태그가 삽입되어 있다.

이 텍스트 파일은 음성자료와의 연동을 염두에 둔 전자문서이다. 각 단위 작품 텍스트에는 이 텍스트와 관련이 있는 조사 카세트 테이프의 명칭과 순서 및 각 테이프 내에서의 텍스트의 위치 정보 등을 포함하고 있다. 이 파일과는 별도로 음성-텍스트 동기화 파일을 유지하고 있으며, 이는 검색 브라우징에서 음성과 텍스트의 동기화 표현이 가능하도록 만든 것이다.

(1) 1차 조사분 전자텍스트의 구조

<item id="1" 서지id="tsu_0001" 기사서지id="Q_0001_101_0032_0035">
** 단위작품 일련번호, 단위작품 기호, 채록 정보 기호
<refdesc> ** 지역 정보 ** 헤더 정보
<classInfo>서울특별시경기도/도봉구/미아동</classInfo> ** 조사지역
<indexInfo/> ** 향후 색인용 정보
</refdesc>
<classserial>[미아동 설화 1]</classserial> ** 지역, 장르, 일련번호
<tapenum>T. 도봉 1 앞</tapenum> ** 채록 테이프 명칭, 위치
<settingplace>미아1동 삼일노인정</settingplace> ** 채록 장소
<record_date>1979. 3. 11.</record_date> ** 채록 일자
<recorder>조희웅, 이영성, 양혜정 조사</recorder> ** 채록자 목록
<writer sex="남" year_old="78">공명수(孔命壽)</writer> ** 제보자
<title>오성과 한음(수박 농사, 감나무 임자)</title> ** 설화 제목(다른 제목)

<body> ** 본문 정보
<description> ** '구연 상황' 즉 조사 환경, 특기 사항의 서술
*조사자(조희웅)가 수집하려는 설화나 민요에 대해 간단히 설명하자, 역사 얘기도 되느냐고 묻고, 많이 알려진 얘기이므로 조사자들도 다 알고 있을 것이라면서, 오성과 한으멩 대한 단편적인 얘기들을 들려 주었다. 노인정에 모여 듣던 노인들은 감나무 임자에 대한 얘기는 들은 적이 있다면서 오성과 한음은 하늘이 낸 사람들이라면서 맞장구치기도 했다. *
</description>
<bodytext> ** 본문 텍스트
한음의 이름이 그전에 유벽(遺腹)이라 켔는데, 유벽이가 일곱 살 묵을(먹을)때 아 이거 장난이 어떻게 심하던지, 아 넘의(남의) 수박밭을 쫄딱 ** 본문 텍스트 단락
<page id="33"/> ** 페이지 구분자
(모두) 망해 삐렸어. 가서 수박밭을 갖다 전부 그만 못쓰게 헤리쳐비리놔 노은끼네(헤쳐 버려 놓으니까) 수박 그 숨은(심은) 주인이 와서 아 이거 수박 값을 물어 내라고 하거등. 그래 인저 수박 값을 물어주기로 하고 그 유벽이 어미니가 자기 인자 오빠한테 가서,
"아 수박 값을 물어돌라고 하는데 어떻게 해야 되겠느냐?"
그렁개(그러니까),
"아 수박 값을 내가 물어 줄꾸마."
그래 수박 값을 인자 물어 주었는데, 그래 그러고 난 뒤, 아 그 수박넝쿨을 수박밭을, 마 풋

```
        밭에 많이 나온개 갈아 뒤비 비렀어, 그놈을, 그래 갈아 뒤비 가(가지고) 그놈이 거름이 돼가지
        고, 아 그 이듬해 오늘은 수박이 어뜨큰 열어 놨는지, 딴 수박은 다 졌는데, 그 집 수박은 그마
        그밭, 밭에 많이 열었단 말이야. 이렇게 떡 해 놓은개 수박이 많이 열어 가지고 내다 판게네,
        아 이거 그전 농사진거하고 멧 곱쟁이가 이익을 봤단 말이야, 그래 그 수박 주인이, 수박밭 주
        인이,</p>
      <p>"아 이거 내가 이렇금 이득을 봐 안 되것다."</p>
      ..... (본문 이하 생략) .....
      <page id="35"/>
      <p>그래요, 이 오성이 권율선(선생이라 부르려다 취소하며) 대감님을 한 번 혼을 내준 적이 있더
        라요. (조사자 : 그 어린내 지혜가 대단하군요.)</p>
    </bodytext>
    <bodysound>Q_0001_101_0032_0035.asf</bodysound>  ** 음성 파일명
  </body>

  <profile-desc>  ** 작업 관련 정보
    <date>2003. 4. 25.</date>  ** 최종 처리일
    <work>교정, 태깅</work>  ** 작업 내용
    <maker>한국정신문화연구원 ○○○</maker>  ** 작업자(소속, 이름)
    <contents>구비문학대계 1-1 본문 XML파일</contents>  ** 출판도서명 설명
  </profile-desc>

  <copyright>  ** 저작권 정보
    <copywriter>한국정신문화연구원</copywriter>  ** 저작권 소유자
    <availibility>한국역사정보통합시스템을 통하여 무료 배포</availibility>  ** 저작권 유통 정보
  </copyright>
</item>
```

(2) 1차 조사분의 특징과 문제점

1) 데이터의 구조가 제대로 표현되지 못한 전자문서다.

각 단위 구비작품의 본문(body)은 정확히 표시되어 있지만, 헤더(header)와 푸터

(footer)의 구획이 불분명하다. 따라서 본문 앞부분을 헤더로, 본문 다음 부분의 푸터로 재조정하여 그 구조를 명확히 표현해야 하고, 태그 명칭도 조정해야 한다. 또한 이 전자텍스트는 종이책 출판의 단점을 상속한 전자문서이어서, 조사지역 개관 등이 단위작품과 분리되어 있거나 아예 표현되지 않았다. 또한 단위작품 관련 사진 자료 등도 표현되지 못한 상태이다. 또한 조사는 되었으나 출판되지 않은 내용은 이 코퍼스에 포함되어 있지 않다.[7] 따라서 이러한 문제가 되는 부분을 2차 조사분에 맞추어 확장할 필요가 있다.

2) 종이책의 구조를 표현한 전자문서다.

전자문서화를 착수할 때 당시에는 TEI Light라는 전자문서화의 기준을 적용한 "21세기 세종계획"의 지침을 좇았는데, 사실상 이 지침은 종이책의 전자적 구현이라고 하는 SGML의 하위 버전 또는 간략 버전이었다. 따라서 1차 조사분 중 일부분은 이와 같은 지침에 따라 작성되어 "21세기 세종계획"의 구어자료 말뭉치로 수록된 바 있다. 다음은 그 파일의 일부분이다.

```
<!DOCTYPE tei.2 SYSTEM "c:₩sgml₩dtd₩tei2.dtd" [
        <!ENTITY % TEI.spoken "INCLUDE">
        <!ENTITY % TEI.corpus "INCLUDE">
        <!ENTITY % TEI.extensions.ent SYSTEM "sejong1.ent">
        <!ENTITY % TEI.extensions.dtd SYSTEM "sejong1.dtd">
]>
<tei.2>
```

[7] 미전사 자료의 발굴과 전사에 대해서는 한국학중앙연구원의 공동연구과제로 수행된 바 있다. 종이책의 분량을 조절한다든가 또는 작품의 완성도를 고려한다든가 하는 것이 미전사의 이유였고, 몇몇 작품은 추가 전사를 통하여 구비대계에 포함시켰다.

<teiHeader>
<fileDesc>
<titleStmt>
<title>한국구비무학대계 1-7 32-35(토정 선생)</title>
<author>한국정신문화연구원</author>
<sponsor>대한민국 문화관광부</sponsor>
<respStmt><resp>문헌 입력, 표준화, 헤더 붙임</resp>
<name>한국정신문화연구원</name>
</respStmt>
</titleStmt>
<extent>397 어절</extent>
<publicationStmt>
<distributor>국립국어연구원</distributor>
<idno>QABG0101.HWP</idno>
<availability> 배포 불가 </availability>
</publicationStmt>
<notesStmt>
<note></note>
</notesStmt>
<sourceDesc>
<bibl><author>한국정신문화연구원</author>
<title>한국구비무학대계 1-7 32-35(토정선생)</title>
<pubPlace>서울</pubPlace>
<publisher>고려원</publisher>
<date>1982-12-27</date>
</bibl>
</sourceDesc>
</fileDesc>
<encodingDesc>
<projectDesc> 21세기 세종계획 1차연도 말뭉치 구축
</projectDesc>
<samplingDecl> 원문대로 입력하되, 틀린부분은 주석을 달아서 표시함
</samplingDecl>
<editorialDecl> 21세기 세종계획 문헌 입력 지침에 따름

</editorialDecl>
</encodingDesc>
<profileDesc>
<creation><date>1981-5-3</date></creation>
<langUsage>
<language id=KO usage=99>한국어, 방언</language>
</langUsage>
<particDesc>
<person id='P10' sex='?' age='?'> 성기열+최명동+김용범 </person>
<person id='P20' sex='남' age='77'> 박광서 </person>
<person id='P40' sex='?' age='?'> 일동 </person>
</particDesc>
<settingDesc>
노인회관에서 한 7~8명의 노인들이 둘러 앉아 있었으나 이야기를 꺼내기를 서로 주저하자, 자기가 먼저 주머니 끈을 끄르겠다고 자진하여 시작하였다. 몸짓과 표정이 매우 자유스러우면서도 꼿꼿이 앉은 자세로 진행하였다. 이에 분위기가 누그러져 다음 제보자가 계속 이야기를 꺼내게 되었다
</settingDesc>
<textClass>
<catRef scheme='SJ21' target='M2336'>구비문학자료</catRef>
</textClass>
</profileDesc>
<revisionDesc>
<change><date>1998-6-22</date>
<respStmt><resp>보조연구원</resp><name>윤승준,박옥주,신은경,김현진</name></respStmt>
<item>텍스트 입력</item>
</change>
<change><date>1998-8-10</date>
<respStmt><resp>보조연구원</resp><name>윤승준,박옥주,신은경,김현진</name></respStmt>
<item>교정 및 헤더 및 태그 부착</item>
</change>
<change><date>1998-11-2</date>
<respStmt><resp>보조연구원</resp><name>윤승준,박옥주,신은경,김현진</name></respStmt>
<item>교정</item>
</change>

</revisionDesc>

</teiHeader>

<text>

<body>

<head>토정선생</head>

옛날 옛적에요. 토정선생이란 이 있었대요. 토정이란 선생님이 있었대요. 근데 이 분이 아마 그 이인이나 마찬가지던 모양이지요? 해가 있기 전에 담뱃몰 하더래요. 담뱃모를 하니까, 그때 어떤 새파란 청년이 지나가다,</u>

<u who=p20> 여보, 해가 있기 전에 담뱃모를 하면 어떻게 됩니까? </u>

<u who=p20> 아니 얘 내일 열 한 시면 비가 올 테니까, 단발모를 한다. </u>

<u who=p20>그 분이 누구냐면, 그 신령이 누구냐면 그이여어. 비를 직접 내린대요.</u>

<u who=p20> 아, 정말 비가 오느냐? </u>

<u who=p20> 아, 열 한 시 비가 온다 고.</u>

<u who=p20> 아, 그러면 우리 내기를 합시다. </u>

<u who=p20> 그럼 무슨 내기요? </u>

<u who=p20>목줄내기를 하자구 그러드래요.</u>

<u who=p20> 그래 목줄내기를 하자. <stage>기침</stage></u>

…… (중략) ……

</body>

</text>

</tei.2>

그러나 한국역사정보통합시스템의 자료로서 구축이 될 때에 구비대계의 자료는 앞에서 보인 바와 같이 매우 간략한 수준에서 전자문서화가 이루어졌다. 그리고 2차 조사사업에서는 이보다 발전된 문서 구조를 가지다 보니 1차 조사분과는 적지 않은 차이를 보인다.(2차 조사분은 전자문서가 먼저 출판되었고, 후에 종이책이 출판되어서 이러한 문제는 없다.) 대표적인 것이 바로 쪽 구분 태그 및 주석 태그의 삽입이다. 특히 쪽 구분 태그가 문제가 된다. 본문 중에는 들어 있는 <page> 태그는는 종이책의 페이지를 표시한 것으

로서, 만일 본문이 한 페이지 내에 수록되어 있는 작품일 경우(민요 등)에는 이 태그가 표시되지 않음으로써 단위 작품별로 데이터의 구조가 다르게 되는 현상을 일으킨다. 그리고 이 <page> 태그는 본문 중 실제 발생 위치에 표시되는데, 그러다 보니 어절의 중간에 삽입되는 경우가 많다. 이러한 현상은 결국 장차 언어적 코퍼스로서 어휘 검색이나 정규식 검색 등을 방해하는 요소가 된다. 따라서 향후에는 본문에서 <page> 태그를 모두 삭제하고 이를 푸터(footer)의 출판 도서 설명에서 표현(예: p.123, pp.56~58 등)하는 것이 바람직하다.

3) 음성자료와의 연동성이 고려된 전자문서다.

이 자료는 구비문학 사이트(구비누리)에서 음성과 텍스트가 연동되어 제공되고 있다. 이 점은 한국의 구비문학 정보 서비스가 유사 서비스에 비해 매우 뛰어난 기능으로서 세계의 관련 학자와 기관으로부터 주목을 받고 있다. 1차 조사의 음성 파일은 2차 조사에서처럼 단위 작품별로 저장되어 있는 것이 아니라 테이프 단위로 저장되어 있어서 작품 단위의 활용에 장애가 되고 있다. 향후에 각 작품 단위별로 별도의 음성 파일로 구분하여 저장할 필요도 있고, 또한 현재의 파일 포맷(asf)을 통용성이 높은 포맷(mp3 등)으로 변환할 필요가 있다.

4) 데이터의 무결성을 재검토할 필요가 있다.

1차 조사 자료 중에는 출판본을 스캔하여 일차 텍스트를 입력한 것과 출판본을 직접 타이핑하여 입력한 것들이 있다. 그 어떤 경우든지 출판본의 오류가 고스란히 상속되는 문제가 있다. 아울러서 전산 입력 과정에서 타이핑의 잘못이나 OCR의 오류가 발생한다. 향후 이러한 오탈자는 일괄 교정하여 전자텍스트의 무결성을 높여야 한다.

6. 1차 조사 자료의 규모와 향후 과제

1차 조사 자료의 규모를 계량적으로 충분히 조사한 적은 없다. 이를 위해서는 데이터베이스 내부의 파일을 대상으로 전수 조사를 해야 하는데, 이것은 향후에 시도하기로 하고 여기서는 대략적인 규모를 샘플링을 통해서 산출해 보기로 한다.

조사 대상으로 삼은 것은 1차 조사분 중『한국구비문학대계 3-1』(충북 충주시, 충주군)이다. 여기에는 충주시의 설화와 무가, 그리고 도농 통합 전의 충주군의 노은면, 상모면, 소태면, 신니면, 주덕면의 설화, 민요, 무가가 수집되어 있다. 이 책은 구비작품 수록 면만으로 볼 때 대략 400면 정도이다. 그런데 1000면 이상이 자료집도 적지 않으므로 평균 500면으로 본다면 1차 조사분『한국구비문학대계』전 82권의 면수는 대략 4만 면을 넘는다. 그리고, 각 면의 규모를 어절로 본다면 설화 작품의 경우는 대략 200 어절 내외로 보인다. 민요나 무가는 면 당 어절수가 이에는 미치지 못한다. 따라서 면 당 어절수를 평균 170 어절로 본다면 모두 680만 어절로 계산된다. 만일 2차 조사분을 1차 조사분의 배가 되는 분량으로 계산한다면 구비대계의 총 규모는 어절로 2천만 어절 정도로 예상된다.

이것은 어디까지나 샘플링을 통한 예상치이므로, 먼저 텍스트의 무결성을 높이고, 구조를 완성한 이후 향후 보다 정밀한 조사가 필요하다. 이때에는 다음과 같은 계량 정보를 추가적으로 추출할 수 있을 것이다. 지역별, 장르별, 제보자 등에 따른 계량 정보, 어절 기준 혹은 음절 기준에 의한 단위 자료의 규모에 대한 계량 정보, 조사 테이프 및 음성

파일의 길이에 관한 계량 정보 등이다.

현재 구비대계의 자료는 현장성과 지역성을 존중하여 구어적으로 표기되었기 때문에, 현대의 일반적 독자들과는 유리된 상태이다. 더구나 해외의 한국학자들은 이 자료를 읽기 위하여 대단히 높은 수준의 한국어 공부를 해야만 한다. 따라서 이러한 언어적 장애를 제거하기 위한 노력이 필요하다. "현대화"라는 목표로 구비문학 자료를 현대어로 재서술하는 작업도 증편구비문학대계 개정증보 과제도 들어 있으나, 이는 대체로 시범적인 수준에 불과하고, 오히려 외부 출판사들에게 이러한 현대화 작업과 그 결과물의 출판을 독려하고 있는 상황이다.