군집분석 기법을 이용한 텍스트의 계통 분석*

- 수궁가 '고고천변' 대목을 대상으로 -

최 운 호** · 김 동 건***
(경희대학교 교양학부)

1. 들어가며

인문학은 이제 컴퓨터라는 새로운 매체 기술 위에서 새롭게 태어나고 있다 하겠다. 예를 들면, 고전문헌학의 이본 비교 연구나 언어학의 방언 자료 분석, 또는 언어계통론 연구에 전산언어학, 생물정보학, 텍스트 마이닝 분야의 방법론과 기법 등이 활용되고 있기 때문이다. 특히, 문헌학의 경우, 문헌학의 방법을 이용해서 원전을 복원하거나 문헌을 추적해서 필사본이 어떻

* 이 논문은 2007년 정부(교육과학기술부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2007-361-AL0016).
** 주저자
*** 교신저자

주 제 어: 텍스트 거리, 레벤시타인 거리, 최소편집거리, 군집, 분류, 계통분기, 텍스트 마이닝
text distance, Leveshtein distance, minimum edit distance, cluster, classification, phylogeny, text-mining

게 전승되어 왔는지를 밝히는 연구[1]들이 이루어져 왔는데, 다수의 문헌들 사이의 유사성과 차이성을 이용해서 문헌들 사이의 계통 관계를 밝히는 연구에 컴퓨터 분석 기법을 적용한 연구의 대표적 사례는 ≪캔터베리 이야기 프로젝트(The Canterbury Tales Project)≫이다. ≪캔터베리 이야기 프로젝트≫에서는 '캔터베리 이야기' 작품의 필사본 이본 58개를 디지털화해서 이 중 44개 이본을 컴퓨터를 활용한 계통분석 기법을 적용하여 필사본들의 계통관계를 분석해냈다.[2] 언어의 계통에 대한 연구로는 Kondrak(2002)가 있다. Kondrak(2002)에서는 동적 프로그래밍 기법을 적용해서 두 어휘의 유사한 소리를 대응시키고, 역사비교언어학 연구를 통해서 알려진 언어 변화 규칙을 적용하여서 동근어(cognate)를 추론해 내는 알고리듬을 고안해냈다. 방언 연구의 경우 지역별로 나타나는 어휘들의 차이를 측정할 수 있도록 동적 프로그래밍 기법을 적용한 연구가 이루어졌는데, Heeringa(2004)가 그 대표적인 예이다.[3] 이 연구들은 '전승과 변모'라는 과정을 통해서 확산되어 가는 필사본의 계통 관계를 드러내거나, 유사성과 차이성에 기초하여 차이를 측정하는 연구들이다.

고전문헌학, 방언학, 역사비교언어학 등에서 다양하게 적용되고 있는 알고리듬들과 기법들을 우리나라의 고전 자료에 적용해서 그러한 방법론의 타당성과 정밀성, 활용 가능성을 알아보는 것이 이 연구의 목적으로, 이 연구에서는 판소리 수궁가의 한 대목인 '고고천변'을 대상으로 이 대목이 전승 과정에서 어떻게 변모되면서 확산되어 왔는지를 확인하고 정밀하게 측정하는 것이 목표이다.

판소리는 구비 전승되는 장르이기 때문에 많은 텍스트가 존재할 뿐만 아니라 텍스트 간 차이도 나타난다. 특히 연행되는 음악이라는 장르적 특성으

1) 고전문헌학의 방법론에 대한 소개는 안재원(2008)을 참조.

2) 이에 대해서는 Barbrook et al.(1998)에 간략히 소개되어 있다.

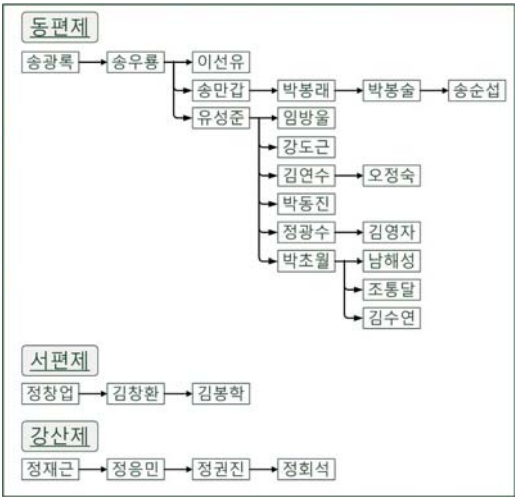
3) Heeringa(2004)에서 방언 어휘 자료를 대상으로 거리를 측정하는 방법을 적용한 반면, 이와 유사한 방법을 산스크리트 텍스트의 필사본 비교에 적용한 연구해서 필사본 이본들에 나타나는 어휘들의 차이를 측정하는 연구로는 Csernel & Patte(2007)가 있다.

로 인하여 판소리에는 많은 유파가 생겨났고 동일한 유파 내에서도 창자에 따라 사설, 음악어법 등에서 각기 다양한 모습을 보이고 있다. 수궁가의 '고고천변' 대목은 이러한 판소리의 전승과 변모의 과정을 잘 보여주는 대목으로, 별주부가 토끼를 찾아가는 여정을 그리고 있는 대목이며 수궁가의 대표적인 '눈대목'이다. '눈대목'은 판소리에서 가장 두드러지거나 흥미있는 대목으로 해당 곡의 핵심이 되는 대목이며, 가장 자주 불리는 대목이다. 따라서 창자들이 자주 불러왔기 때문에 음반 자료가 남아있어 다른 대목에 비해서 비교할 수 있는 연구 자료가 풍부하다. 또한, 핵심 대목이기 때문에 강한 전승력을 지니고 있는 것과 동시에 창자들이 자신의 특장으로 삼아 공력을 기울여 연마하는 대목이어서, 변이가 나타나는 대목이기도 하다. 이러한 '전승과 변모'라는 특성을 분석해서 텍스트 간의 거리를 측정해 보는 것이 이 연구의 목적이다. 즉, '전승'을 통해서 이어지는 유사성이 '변모'라는 과정을 통해서 어떤 차이로 나타나는지, 그 차이는 어느 정도나 되는지를 측정해 내는 것이 이 연구의 목적이다. 구체적으로는, 사설과 음악어법(정간보)을 각각 창자별로 비교를 하며, 사설의 경우에는 사설의 거리 측정 알고리듬을 적용해서 그 거리를 수치 자료로 계산해내고, 음악어법의 경우에는 장단을 단위로 그 차이를 비교한다. 그리고 각각의 결과는 군집분석(cluster analysis)을 통해서 분류하도록 한다.

2. 연구 대상

이 연구의 비교 대상은 판소리 수궁가의 '고고천변' 대목이다. 수궁가는 크게 동편제, 서편제, 강산제 수궁가로 나눌 수 있다. 동편제는 "송광록 → 송우룡 → [이선유, 송만갑, 유성준]"으로 이어지는 전승계보를 가지고 있는 소리제로 현재 가장 활발한 전승 양상을 보이고 있다. 동편제 수궁가는 다시 이선유제, 송만갑제, 유성준제 수궁가로 나누어진다. 이 중 이선유제는 현재 전승이 끊어졌으며, 이선유의 음반이 부분적으로 남아 있기는 하나

‘고고천변’ 대목은 찾아볼 수 없다. 송만갑제는 “박봉래 → 박봉술 → 송순섭”으로 이어지는 전승계보를 지니고 있는데, 박봉래의 음반은 남아 있지 않다. 하지만 송만갑, 박봉술, 송순섭의 음반이 남아 있다. 유성준제는 동편제 가운데에서도 가장 활발한 전승이 이루어지고 있는데, 유성준의 음반은 현재 남아 있는 것이 없다. 하지만 임방울, 강도근, 김연수, 박동진, 정광수, 박초월 등 많은 제자들에 의해 전승되고 있다. 한편 임방울, 강도근, 박동진의 소리는 현재 전승이 끊어진 상태이나 김연수의 소리는 오정숙에게, 정광수의 소리는 김영자에게, 박초월의 소리는 남해성, 조통달, 김수연 등에게 전승되고 있다. 서편제는 “정창업 → 김창환 → 김봉학”으로 이어지는 전승계보를 지니고 있는데, 정창업이나 김봉학의 소리는 음반이 남아 있지 않다. 하지만 김창환의 소리가 부분적으로 남아 있으며, 이 중에 ‘고고천변’이 포함되어 있다. 강산제는 “정재근 → 정응민 → 정권진 → 정회석”으로 이어지는 소리계보를 지니고 있다. 이 중 음반 자료가 있는 창자는 정회석뿐이다. 판소리 ‘수궁가’의 전승 계보는 <그림 1>과 같이 구성[4]할 수 있다.



<그림 1> ‘수궁가’ 전승 계보

4) 이 전승 계보는 인권환(1986, 1992)과 김미선(2001)을 바탕으로 구성하였다.

이상을 정리하면, 본 연구의 대상이 되는 창자는 송만갑, 박봉술, 송순섭, 임방울, 강도근, 김연수, 오정숙, 박동진, 정광수, 김영자, 박초월, 남해성, 조통달, 김수연, 김창환, 정회석 등 모두 16명이다. 이 연구에서 사용한 16명의 음반 목록은 <표 1>과 같다. <표 1>에서 제시된 것처럼, 자료의 구성에서 각 창자의 이름은 영문 약자로 표기하였다.[5)]

<표 1> '고고천변' 비교 연구에 사용한 음반 자료 목록

송만갑(SMK)	• 『동편제 판소리』 • 서울음반/1992년 (SRCD-1064) • 1935년경 녹음
박봉술(PBS)	• 브리태니커 판소리 『수궁가』 • 신나라/1982년 (REG.NO.117)
송순섭(SSS)	• 국악방송 새음원시리즈-새로운 천년의 약속 5. 『송순섭의 수궁가-1』 • 서울음반/2003년 (SRCD-1490)
임방울(IBU)	• 『한국의 전통음악 38. 수궁가 1』 • 오리엔탈 레코드/19XX (DYCD-1438) • 1959년 연세대 노천극장 실황
강도근(KTK)	• 신나라 판소리 명인 시리즈(002) 『수궁가』 • 신나라/1990 (SEL-RO664)
김연수(KYS)	• 동초 김연수 창 판소리 다섯마당 『수궁가』 • 신나라/2007년 (NSC-187-1)
오정숙(OJS)	• 오정숙 판소리 다섯마당 『수궁가』 • 신나라/2001년 (NSSRCD-047)
박동진(PTJ)	• 인간문화재 박동진 판소리 대전집 『수궁가』 • 예전미디어/1988년 (SKCD-K-0253)
정광수(JKS)	• 국악의 향연 vol. 43 정광수 박초월 창 『수궁가』 • 서울음반/1988년 (8810-G211)

5) 앞으로 제시되는 자료와 자료의 분석 결과에서 사용된 영문 약자는 이 표에서 제시된 창자를 가리킨다.

김영자(KYJ)	• 정광수제 수궁가『김영자 소리샘 I』 • MusicNet/1999년 (KACD-0003)
박초월(PCW)	• 『수궁가 1』 • 오아시스/1994년 (ORC-1448)
남해성(NHS)	• 『남해성 수궁가』 • CJ Music/2006년 (CMCC-0718)
조통달(JTD)	• 21세기를 위한 KBS-FM의 한국의 전통음악 시리즈 12『한국의 전통음악(판소리)』 • 해동물산/1994 (HAEDONG-112)
김수연(KSY)	• 『김수연의 수궁가』 • 김수연판소리연구소/2004년 (Z-C1-04-0132)
김창환(KCH)	• 한국의 위대한 명창들 (I)『판소리 5명창』 • 신나라/1996년 (SYNCD103) • 1920년대 중반 녹음
정회석(JHS)	• 소릿길 소리사랑『정회석 수궁가』 • 지구레코드/1999년 (TOPCD-026)

<표 1>에 제시된 음반에서 '고고천변' 대목을 대상으로 사설을 채록하고 소리를 채보하여서 그 자료를 비교의 대상으로 삼았다. 채보는 국악의 전통적인 음악어법을 표시할 수 있는 정간보로 작성하였다. 실전 판소리를 자료로 사용하였기 때문에, 사설의 채록은 음반을 청음하면서 우리말의 형태음소 표기를 존중하되 개인/지역 방언의 차이가 반영되도록 하여서, 형태음소 단위의 분철 표기로 이루어지도록 하였다. 또한 사설과 함께 음악적 어법이 표시되는 정간보의 채보에는 '박'을 단위로 사설이 대응되도록 하되 '리듬 패턴', '조', '붙임새'가 각 '박'마다 분석되어서 반영되었다.[6)]

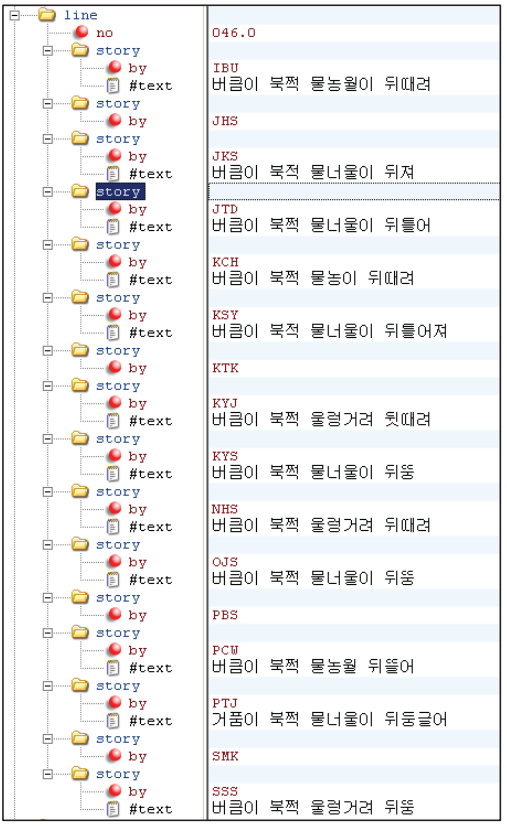
6) 사설의 채록과 정간보의 채보는 전문 국악인인 김미선 선생이 수고해 주셨다. 이 채록과 채보 방식은 판소리의 기록과 교육에서 전통적으로 활용되고 있는 정간보에서 사설과 음악적 요소를 채록하고 채보하는 방식을 사용하였다. 정간보의 예는 김진영 외(2008)을 참조하라.

3. 비교 방법과 결과

3.1. 사설의 비교

사설의 비교를 위해서 판소리 수궁가의 '고고천변' 대목을 채록하고 대응되는 사설 단위로 병렬로 구성하였다. 대응의 예는 <그림 2>와 같다. <그림 2>는 16명 창자의 사설을 병렬 말뭉치(parallel corpus)로 구성하여 XML (eXtensible Markup Language) 파일로 작성한 결과이다. 각 창자별로 채록한 사설은 내용과 장단의 대응 관계를 고려하여서 고유 번호를 부여하여서 대응시켰는데, <그림 2>의 예에서 IBU의 사설 중에서 "버큼이 북쩍 물농월이 뒤때려" 항목은 046.0이라는 코드를 부여받고, 다른 창자의 사설 중에서 이에 대응되는 항목들을 병렬로 구성하였으며, 대응되는 항목이 없는 경우에는 JHS, KTK, PBS, SMK처럼 046.0이라는 코드 항목에 대응되는 것이 없는 것으로 표시하였다. 이렇게 구성된 자료는 사설 비교를 위해서 관계형 데이터베이스 관리 시스템(RDBMS) 자료로 변환하였다. 변환된 자료는 16명의 창자에 대해서 모두 76개의 대응쌍으로 구성되어서, 이 연구를 위해서 구축한 사설 비교 대응 자료는 76×16의 테이블로 구성되었다. 사설의 대응 단위를 비교하기 위해서는 창자별 대응쌍을 구성해야 한다. 모두 16명의 창자를 대상으로 삼았기 때문에, 이 중에서 두 명씩 대응쌍을 구성하는 방법은 120가지가 존재한다.[7] 그 결과 비교를 측정해야 할 대응쌍(s, t)는 모두 120×76=9.120개이다.

7) 16개의 개체에서 순서에 상관없이 2가지를 추출하는 조합은 ${}_{16}C_2 = 120$이다.



<그림 2> 수궁가 '고고천변' 사설 병렬 대응 자료

3.1.1. 거리 측정

사설의 거리 비교에는 최소편집거리(minimum edit distance)라고도 불리는 레벤시타인 거리(Levenshtein distance, LD)[8] 측정법을 사용한다. LD는

8) 이 방법은 Levenshtein(1966)에서 제시된 방법이다. 두 어휘의 최소 편집 거리를 계산하는 데 적용한 예는 Jurafsky & Martin(2000)에서 상세히 설명되어 있고, 이 방법을 확대해서 방언의 차이를 측정하기 위해 적용한 예는 Heeringa(2004)에 자세히 설명되어 있다.

두 개의 스트링이 있을 때 스트링 s를 스트링 t로 바꾸는 데 사용되는 비용을 측정하는 방법이다. 비용의 측정에는 삽입(insertion), 삭제(deletion), 대치(substitution) 세 가지 비용을 계산한다. <표 2>는 고고천변 사설 중에서 대응 코드 [030.0]에 해당하는 항목의 창자별 대응 자료이다. 이 중에서 IBU와 JKS 두 창자의 사설의 거리를 LD를 이용해서 측정하는 방법을 설명해 보기로 한다.

<표 2> 사설의 대응

창자코드	창자	고고천변 사설 [030.0] 항목의 창자별 대응
IBU	임방울	소호천자 기관허던 만수문전에 봉황새
JHS	정회석	
JKS	정광수	소호천자 기관허던 만수문전으 봉황새
JTD	조통달	소호천자 기관허던 만수문전에 봉황새
KCH	김창환	
KSY	김수연	소천자 기관허던 만수문전으 봉황새
KTK	강도근	소호천자 기관허던 만수문전으 봉황새
KYJ	김영자	소호청양 기관허던 만수문전에 봉황새
KYS	김연수	소호시절으 기관허던 만수문장으 봉황새
NHS	남해성	소호천자 기관허던 만수문전의 봉황새
OJS	오정숙	소호시절에 기관허던 만수문전에 봉황새
PBS	박봉술	
PCW	박초월	소호천자 기관허던 만수문전으 봉황새
PTJ	박동진	소천자 기관허던 만수문전에 봉황새
SMK	송만갑	
SSS	송순섭	

LD는 동적 프로그래밍(dynamic programming) 기법을 사용하여서 원천 스트링(source string)을 목표 스트링(target string)으로 변환하기 위해 소요

되는 비용을 계산함으로써 두 스트링의 유사도를 계산하는 알고리듬이다. <표 2>에서 IBU의 사설을 원천 스트링으로 설정하고 IBU의 사설을 JKS의 사설로 변환하는 데 소요되는 비용을 계산해 보면 <표 3>과 같다. IBU에는 존재하는 '호'가 JKS에는 없기 때문에 IBU를 JKS로 변환하려면 '호'를 삭제해야 하며, 이 연산을 D로 표시하였다. 또한 IBU의 사설에 존재하는 '에'는 JKS의 사설에서는 '으'에 대응되어야 하기 때문에 대치 연산을 수행해야 하고 이 연산을 S로 표시하여야 한다. 따라서 IBU의 사설을 JKS의 사설로 변환하기 위해서는 삭제(D)와 대치(S) 연산이 각각 한 번씩 발생했다. '=' 기호는 두 문자가 동일하기 때문에 비용이 발생하지 않았다는 것을 표시한 것이다.

<표 3> IBU ⇒ JKS 변환 비용

소	호	천	자		기	관	허	던		만	수	문	전	에		봉	황	새
소		천	자		기	관	허	던		만	수	문	전	으		봉	황	새
=	D	=	=	=	=	=	=	=	=	=	=	=	=	S	=	=	=	=

이번에는 반대로 JKS의 사설을 IBU의 사설로 변환해 보자. JKS의 사설에 없는 '호'가 IBU의 사설에서는 삽입되어 있고, JKS의 '으'가 IBU에서는 '에'로 대치되어 있다. 따라서 이 경우에는 삽입과 대치 두 가지의 연산이 발생했다. IBU ⇒ JKS, JKS ⇒ IBU의 변환 결과가 동일한 값을 유지하려면, 삭제와 삽입 연산에 소요되는 비용은 동일한 값으로 설정하는 것이 일반적이다.

<표 4> JKS ⇒ IBU 변환 비용

소		천	자		기	관	허	던		만	수	문	전	으		봉	황	새
소	호	천	자		기	관	허	던		만	수	문	전	에		봉	황	새
=	I	=	=	=	=	=	=	=	=	=	=	=	=	S	=	=	=	=

이 방법에 의하면, 어떤 스트링 *S*와 *T*의 거리는, *S*를 스트링 *T*로 변환할 때 필요한 삽입, 삭제, 대치 연산의 합으로 정의된다. 과정을 동적 프로그래밍 기법으로 구현한 것이 LD 알고리듬이다. LD 알고리듬은 <표 5>와 같다.[9] 이 연구에서는 del-cost(삭제 비용), ins-cost(삽입 비용), subst-cost(대치

<표 5> Levenshtein Distance 알고리듬

```
int LevenshteinDistance(char s[1..m], char t[1..n])
{
    //d is a table with m+1 rows and n+1 columns
    declare int d[0..m, 0..n]
    for i from 0 to m
        d[i, 0] := i
    for j from 0 to n
        d[0, j] := j

    for j from 1 to n
    {
        for i from 1 to m
        {
            if (s[i] == t[j]) then
                d[i, j] := d[i-1, j-1]
            else
                d[i, j] := min
                        (
                        d[i-1, j] + del-cost,
                        d[i, j-1] + ins-cost,
                        d[i-1, j-1] +subst-cost
                        )
        }
    }
    return d[m, n]
}
```

9) 동적 프로그래밍 방법을 적용한 Levenshtein distance 알고리듬과 이 알고리듬에 대한 설명은 Kondrak(2002: 25), Heeringa(2004: 127)에 설명되어 있다. 이 연구에서는 이보다는 더 간결한 형태로 제시되어 있는 위키피디아(Wikipedia)의 자료를 변형하여 제시하였다. (http://www.wikipedia.org/Levenshtein_distance 참조)

비용)을 모두 1로 설정하였다. LD는 두 스트링을 비교하면서 이전의 비교 결과가 다음 비교에 다시 사용되도록 함으로써 효율적으로 계산을 수행하는 알고리듬이다.

<표 6>은 LD 알고리듬을 이용해서 두 스트링을 비교하는 과정이다.[10] 비교 결과는 상단에서부터 하단으로, 좌측에서 우측으로 채워지고 최종적으로 d[m, n]의 값인 2가 두 스트링의 거리이다.

<표 6> LD 알고리듬을 이용한 두 스트링의 거리 비교

	#	소	호	천	자	기	관	허	던	만	수	문	전	에	봉	황	새
#	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
소	1	**0**	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
천	2	1	**1**	**1**	2	3	4	5	6	7	8	9	10	11	12	13	14
자	3	2	2	2	**1**	2	3	4	5	6	7	8	9	10	11	12	13
기	4	3	3	3	2	**1**	2	3	4	5	6	7	8	9	10	11	12
관	5	4	4	4	3	2	**1**	2	3	4	5	6	7	8	9	10	11
허	6	5	5	5	4	3	2	**1**	2	3	4	5	6	7	8	9	10
던	7	6	6	6	5	4	3	2	**1**	2	3	4	5	6	7	8	9
만	8	7	7	7	6	5	4	3	2	**1**	2	3	4	5	6	7	8
수	9	8	8	8	7	6	5	4	3	2	**1**	2	3	4	5	6	7
문	10	9	9	9	8	7	6	5	4	3	2	**1**	2	3	4	5	6
전	11	10	10	10	9	8	7	6	5	4	3	2	**1**	2	3	4	5
으	12	11	11	11	10	9	8	7	6	5	4	3	2	**2**	2	3	4
봉	13	12	12	12	11	10	9	8	7	6	5	4	3	3	**2**	3	3
황	14	13	13	13	12	11	10	9	8	7	6	5	4	4	3	**2**	3
새	15	14	14	14	13	12	11	10	9	8	7	6	5	5	4	3	**2**

10) 설명의 편의를 위해서 공백 문자(space)는 제거하고 비교한다.

LD 알고리듬을 적용하여서 9,120개의 스트링 쌍의 거리를 측정하면서, 그 결과값들을 보정해주는 정규화 과정을 거쳐야 한다. LD 알고리듬으로 두 스트링의 거리를 측정하는 경우, 한 스트링의 길이가 12이고, 다른 스트링의 길이가 0이면 두 스트링의 거리는 12가 된다. 그리고 두 스트링의 차이가 비교 대상이 되는 스트링의 길이에 따라서 달라지기 때문에 측정값들을 비교 대상이 되는 두 스트링의 길이를 이용해서 정규화 해 줘야 거리 측정값이 비교 대상이 되는 스트링의 길이에 의해서 영향을 받지 않게 된다. 정규화는 다음과 같이 수행하였다.

(1) LD 측정값의 정규화

$$norm\,Val_i = \frac{Distance(s_i, t_i)}{MAX(length(s_i), length(t_i))}$$

비교 대상이 되는 두 스트링을 s_i, t_i라고 했을 때, 두 스트링 중에서 길이가 긴 스트링의 길이로 측정 거리를 나눠주면, 그 결과는 언제나 $0 \leq norm\,Val_i \leq 1$의 범위에 있게 된다.

(2) LD 측정값의 평균

$$\sum_{k=1}^{n} \frac{norm\,Val_i}{n} = \frac{\sum_{k=1}^{n} norm\,Val_i}{n}$$

(1)을 적용해서 정규화한 값들의 평균을 구해서 두 창자의 거리를 나타내면, 16명의 창자의 거리는 16×16의 대칭행렬로 구성되며, 그 결과는 <표 7>과 같다.

<표 7> 16명 창자의 LD 거리 대칭 행렬

	IBU	JKS	JHS	JTD	KCH	KSY	KTK	KYJ	KYS	NHS	OJS	PBS	PCW	PTJ	SMK	SSS
IBU	0	0.1275	0.3206	0.1141	0.5179	0.118	0.1762	0.1243	0.2177	0.1112	0.1657	0.5726	0.1151	0.2544	0.515	0.4092
JKS	0.1275	0	0.3366	0.1622	0.5134	0.159	0.1898	0.062	0.2132	0.1672	0.2123	0.5383	0.1543	0.25	0.4926	0.4523
JHS	0.3206	0.3366	0	0.3276	0.5002	0.3345	0.2784	0.3263	0.4124	0.3378	0.388	0.4035	0.3216	0.3699	0.4208	0.3941
JTD	0.1141	0.1622	0.3276	0	0.5009	0.078	0.1595	0.1555	0.2193	0.0691	0.1778	0.5467	0.0611	0.2234	0.4869	0.3741
KCH	0.5179	0.5134	0.5002	0.5009	0	0.5035	0.5333	0.5109	0.5254	0.5034	0.5783	0.5427	0.5075	0.5415	0.4563	0.4831
KSY	0.118	0.159	0.3345	0.078	0.5035	0	0.1605	0.1578	0.2016	0.0714	0.1709	0.5392	0.0597	0.2216	0.4942	0.3761
KTK	0.1762	0.1898	0.2784	0.1595	0.5333	0.1605	0	0.1899	0.2607	0.164	0.2405	0.4817	0.1667	0.2737	0.4729	0.4432
KYJ	0.1243	0.062	0.3263	0.1555	0.5109	0.1578	0.1899	0	0.2057	0.1486	0.2031	0.5381	0.1522	0.2548	0.5001	0.4409
KYS	0.2177	0.2132	0.4124	0.2193	0.5254	0.2016	0.2607	0.2057	0	0.2111	0.0884	0.5673	0.2026	0.3224	0.5345	0.4827
NHS	0.1112	0.1672	0.3378	0.0691	0.5034	0.0714	0.164	0.1486	0.2111	0	0.1815	0.5382	0.0588	0.2273	0.4878	0.3732
OJS	0.1657	0.2123	0.388	0.1778	0.5783	0.1709	0.2405	0.2031	0.0884	0.1815	0	0.6223	0.1746	0.3238	0.5603	0.4608
PBS	0.5726	0.5383	0.4035	0.5467	0.5427	0.5392	0.4817	0.5381	0.5673	0.5382	0.6223	0	0.5403	0.5871	0.2755	0.2877
PCW	0.1151	0.1543	0.3216	0.0611	0.5075	0.0597	0.1667	0.1522	0.2026	0.0588	0.1746	0.5403	0	0.2238	0.4829	0.3759
PTJ	0.2544	0.25	0.3699	0.2234	0.5415	0.2216	0.2737	0.2548	0.3224	0.2273	0.3238	0.5871	0.2238	0	0.5649	0.4772
SMK	0.515	0.4926	0.4208	0.4869	0.4563	0.4942	0.4729	0.5001	0.5345	0.4878	0.5603	0.2755	0.4829	0.5649	0	0.2179
SSS	0.4092	0.4523	0.3941	0.3741	0.4831	0.3761	0.4432	0.4409	0.4827	0.3732	0.4608	0.2877	0.3759	0.4772	0.2179	0

3.1.2. 군집분석과 결과 해석

다변량 자료(multivariate data sets)를 분석하면서 이미 알려진 또는 정해진 몇 개의 부류로 나누는 기법을 분류(classification)라고 하고, 미리 알려지지 않은 수의 부류로 분할하는 기법을 군집 분석(clustering analysis)이라고 한다.[11] 레벤시타인 거리를 적용해서 16명 창자의 사설들 간의 거리를 측정한 결과가 <표 7>에 제시되어 있기 때문에 이 결과에 군집 분석을 적용해 보도록 한다. 군집 분석에는 R을 사용한다.[12] <표 7>의 자료를 이용해서 창자 간 거리를 바탕으로 군집(cluster)을 나누고, 또한 계통분기[13]를 분석하면 각각 <그림 3>과 <그림 4>와 같다.

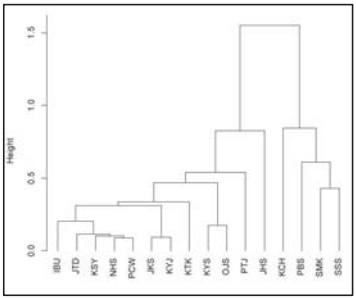
<그림 3과> <그림 4>의 결과를 <그림 1>의 수궁가 전승 계보와 비교해서 해석해보도록 하자.

---

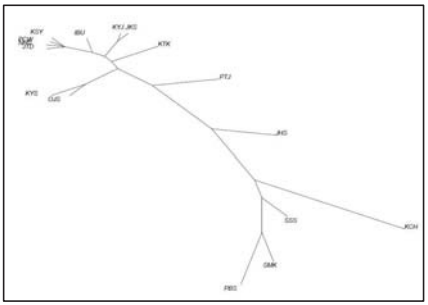
11) Bilisoly(2008: 219).

12) R에 대해서는 R Development Team(2009)를, 그리고 텍스트 마이닝 방법에서 활용되는 텍스트의 군집 분석은 Bilosoly(2008)을 참조하라. 이 연구에서는 언어 자료의 분석을 위한 R 활용법인 Baayen(2008)을 함께 참고하였다.

13) 계통 분기는 Paradis(2004)에서 제시된 방법을 사용하였다.



<그림 3> 사설 간 거리에 따른 군집



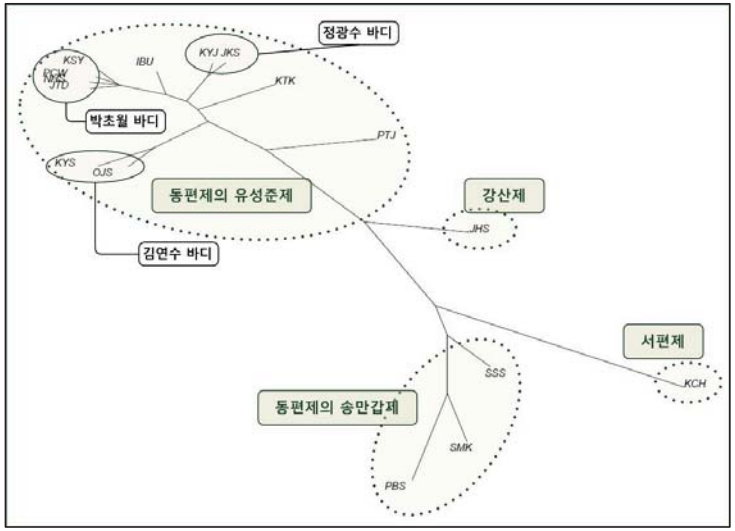
<그림 4> 사설 간 거리에 따른 계통 분기

'고고천변' 대목의 사설은 크게 수궁가 전승계보와 동일하게 동편제의 송만갑제, 동편제의 유성준제, 강산제, 서편제의 4개의 군집이 나눠진다. 이 중 동편제에 속하는 송만갑제와 유성준제는 같은 동편제임에도 불구하고 유성준제와 송만갑제의 거리보다는 유성준제와 강산제의 거리가, 그리고 송만갑제와 서편제의 거리가 더 가깝게 측정되었다. 강산제(JHS)와 서편제(KCH)는 각각 개별적인 군집을 이루고 있는데, <그림 3>과 <그림 4>를 종합해서 해석해 보면, 강산제의 경우는 동편제의 유성준제와 거리가 가깝고, 서편제는 동편제의 송만갑제와 거리가 가깝다.

송만갑제에 속하는 송만갑(SMK), 박봉술(PBS), 송순섭(SSS)은 한 군집을 이루고 있기는 하나, 서로 간의 거리는 {KSY, PCW, NHS, JTD}, {KYS, OJS}, {KYJ, JKS} 군집의 각 구성요소들 사이의 거리와 비교해 보면 그 거리가 상대적으로 멀게 나타나고 있어 이들 간의 사설 차이가 심한 것을 알 수 있다. 송만갑제의 전승 계보가 송만갑(SMK) → 박봉술(PBS) → 송순섭(SSS)임을 감안하면, 송순섭의 사설은 그의 스승인 박봉술(PBS)의 사설보다 스승의 스승인 송만갑(SMK)의 사설과 가깝다는 것을 알 수 있는데, 이는 박봉술(PBS)이 전승 과정에서 변이 혹은 탈락시킨 사설을 송순섭(SSS)이 복원하여 사설을 짠 것으로 해석된다.[14]

14) 박봉술(PBS)의 사설은 분량상으로 볼 때, 송만갑(SMK)의 사설보다 현저하게 적은데, 이는 탈락된 사설이 많기 때문이다.

유성준제에서는 크게 보면 박동진(PTJ)과 그 외의 창자들이 각각의 군집을 이루고 있는데, 그 거리도 다른 창자들 간의 거리보다도 상당히 멀다. 박동진(PTJ)을 제외한 유성준제의 창자들은 <그림 4>와 <그림 5>에서 볼 수 있듯이, 각각의 거리가 상대적으로 가깝게 나타나고 있어 사설이 친연성이 있음을 보여준다. 이들 창자들은 다시 [KYS, OJS]와 그 외 창자들이 각각의 군집을 이루고 있는데, [KYS, OJS]가 다른 군집으로 나타나는 것은 김연수(KYS)의 사설 개작에서 비롯된 것으로 여겨진다. 김연수(KYS)는 당대에 활동한 명창들 중에서는 드물게 중등교육을 받은 지식인으로 문자속이 넉넉하여 신재효의 사설을 받아들이고 그 외의 사설들도 소리의 이면에 맞게 개작하여 사설을 구성한 것으로 알려져 있다.



<그림 5> 사설 간 거리와 판소리 수궁가 전승 계보

한편 <그림 5>를 보면, 유성준제에 속하는 임방울(IBU), 강도근(KTK) 역시 앞서 언급한 김연수(KYS), 박동진(PDJ), 정광수(JKS)와 마찬가지로 유성준제 내에서는 각각 별개의 군집에 속하는 것으로 그 거리가 측정된다.

최혜진(2006: 424)에서는 "정광수와 임방울은 유성준의 제자였으나 이들 또한 자신만의 소리길을 닦아서 사설과 소리에 각각의 특징을 보유하고 있는 명창들"이라고 말하고 있는데, 이들 뿐만 아니라 강도근(KTK), 김연수(KYS), 박동진(PTJ)도 역시 자신의 특징적인 사설을 보유하고 있는 것으로 분석된다.

유성준제의 박초월 바디인 박초월(PCW), 남해성(NHS), 조통달(JTD), 김수연(KSY), 정광수 바디인 정광수(JKS), 김영자(KYJ), 김연수 바디인 김연수(KYS), 오정숙(OJS)은 각각 바디 내에서 거리가 매우 가까운 모습을 보이는데, 이는 박초월, 정광수, 김연수를 사사한 창자들이 현재 활동하고 있는 창자들이라는 점을 감안하면, 전대의 창자들보다는 스승의 사설을 거의 그대로 이어받고 있음을 보여준다. 이들 바디 중에서는 김연수 바디에 속하는 김연수(KYS)와 오정숙(OJS)의 거리가 가장 멀고, 박초월 바디에 속하는 창자들의 거리가 가장 가깝다. 그리고 박초월 바디 내에서는 남해성(NHS)이 스승인 박초월(PCW)과 가장 가깝다.

3.2. 음악어법의 비교

3.1절에서 사설을 대상으로 비교를 하였는데, 이번에는 각 창자의 음악어법을 비교해 보도록 한다. 우선 16명의 소리를 정간보[15]로 채보하고, 각각의 정간보를 모두 3.1절에서와 마찬가지로 120개의 대응쌍을 구성하여서 비교하였다. 사설의 비교와의 차이점은, 사설 비교에서는 스트링 비교 알고리듬을 사용하였지만 음악어법의 비교에서는 정간보의 각 박을 1:1로 대응시켜서 그 동일성과 차이성을 직접 비교하였다. 이 연구에서 채보한 정간보의 예는 <표 8>과 <표 9>에 그 일부가 제시되었다. 비교 방법은, 각 장단(행) 속의 각 박(열)을 문자와 마찬가지로 취급하여 비교하고, 한 장단의

15) 판소리 음악어법에서 가장 중요한 것은 장단이다. 따라서 아무리 소리를 잘 하더라도 장단이 맞지 않으면 소리로서 결격인 것으로 여겨졌다. 판소리 정간보는 사설과 장단과 관계, 각 장단 내에서의 사설과 박의 관계를 가장 잘 보여주는 악보이기 때문에 판소리의 음악적 특성을 가장 잘 보여준다고 할 수 있다.

비교 결과를 취합해서 그 평균을 이용하여 각 창자의 거리를 측정하였다. 비교를 할 때, 해당 장단의 대응쌍이 존재하는 경우에만 비교했기 때문에, 평균의 측정은 대응쌍의 크기에 따라 조정하였다.

<표 8> 송만갑 '고고천변' 정간보(앞부분)

박자	1	2	3	4	5	6	7	8	9	10	11	12
1	고	--	고	천	--	변	홍	일	광	△		
2	부	--	상	--	--	으	높	이	떠	△		
3	양	--	--	--	--	곡	잦	은	안	--	--	개
4	월	--	--	봉	-으	로	돌	--	고	△		
5	어	--	장	--	--	촌	개	--	짖	고	△	
6	회	안	--	--	--	봉	구	름	--	떠	△	
7	노	--	화	는	--	다	눈	되	고	△		
8	부	평	은	둥	--	둥	높	--	이	떠	--	--
9	--	--	--	어	--	룡	잠	--	자	고	△	
10	잘	--	새	펄	--	펄	날	아	들어	△		
11	동	--	정	여	-천	으	파	시	--	추	△	
12	금	--	색	추	파	가	여	-그	라	△		

<표 9> 박초월 '고고천변' 정간보(앞부분)

박자	1	2	3	4	5	6	7	8	9	10	11	12
1	고	고	--	천	--	변	일	륜	홍	△		
2	부	--	상	--	--	으	높	--	이	떠	--	△
3	양	--	곡	--	--	으	잦	은	--	안	개	--
4	월	--	봉	--	으	로	돌	--	고	돌아	△	
5	△		어	장	--	촌	개	--	짖	고	△	
6	△		회	안	--	봉	구	-름	이	떴	-구	나
7	노	화	--	난	--	다	눈	--	되	고	--	△
8	부	△	평	--	--	은	물	-에	둥	--	--	실
9	어	--	--	룡	--	은	잠	자	고	△		
10	잘	새	는	펄	--	펄	날아	든	--	다	--	△
11	동	--	정	여	천	으	파	시	--	--	--	추
12	금	--	색	추	파	가	여	-기	라	△		

이렇게 구성된 정간보를 창자별로 비교하기 위해서 정간보를 장단과 박 단위로 비교하고 그 결과를 종합해서 창자별 거리를 측정한다.

(3) 정간보 장단의 비교

$$\delta = \sum_{k=1}^{m} dist(a_k, b_k)$$

$dist(a,b)$ 함수는 정간보의 대응되는 박이 일치하는 경우에는 0, 그렇지 않으면 1을 반환하도록 하였다. 한 장단은 12박으로 구성되어 있기 때문에 $m=12$이고, 따라서 $0 \leq \delta \leq 12$이다. 박의 수가 동일하기 때문에 사설에서처럼 길이에 따른 정규화는 하지 않는다.

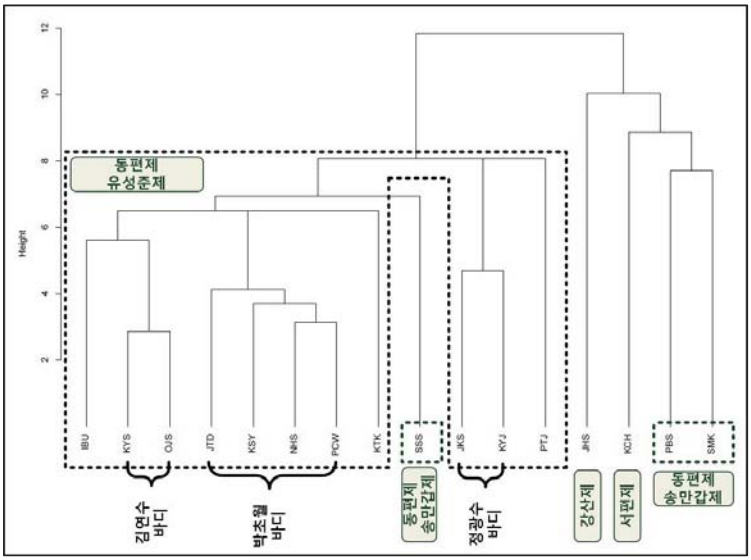
(4) 창자의 거리 비교

$$\Delta = \frac{1}{n}\sum_{i=1}^{n} \delta_i$$

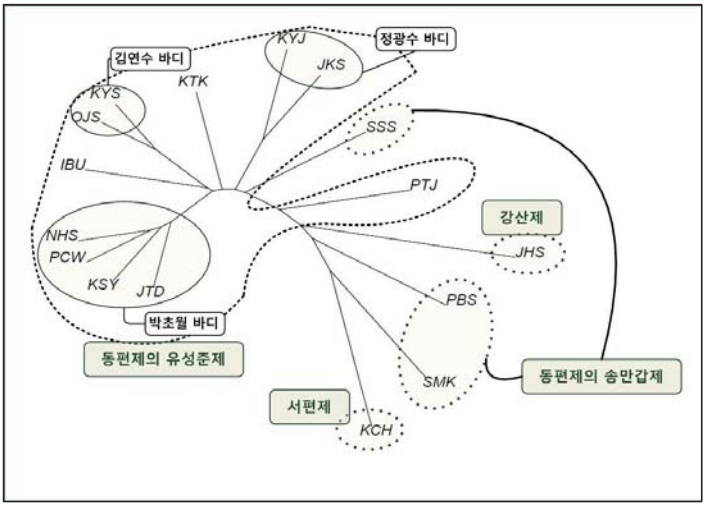
창자별로 대응되는 장단이 있을 때에만 장단 비교를 하기 때문에, (4)에서 n의 크기는 대응되는 창자마다 서로 다른 값이 적용된다. 사설의 비교와 마찬가지로 음악어법의 비교 결과를 창자별 거리를 표시한 대칭 행렬로 구성한 뒤, 군집 분석과 계통 분기 연산을 수행하였다. <그림 6>과 <그림 7>은 이 결과에 <그림 1>의 수궁가 전승 계보를 설명으로 첨가한 것이다.

‘고고천변’의 음악어법 분석에서 가장 먼저 주목되는 점은, 동편제 송만갑제에 속하는 송순섭(SSS)이 같은 소리제인 송만갑(SMK), 박봉술(PBS)과 같은 군집으로 분류되지 않고, 박동진(PTJ)보다도 더 유성준제의 다른 창자들과 거리가 가깝게 나타난다는 점이다. 이러한 모습을 사설 분석의 결과와 종합하여 해석해보면, 송순섭(SSS)은 송만갑(SMK), 박봉술(PBS)의 사설을 계승하면서도 소리는 유성준제의 영향을 많이 받은 것으로 판단된다.

강산제(JHS)와, 서편제(KCH) 그리고 송만갑제의 송만갑(SMK), 박봉술(PBS)은 크게 한 군집을 이루고 있다. 하지만 상호간의 거리는 상당히 먼



<그림 6> 음악어법(장단) 비교 결과의 장단 분석



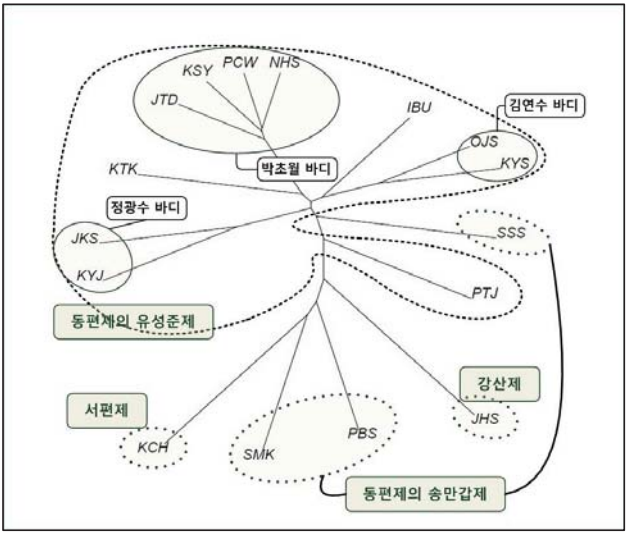
<그림 7> 음악어법(장단) 비교 결과의 계통 분기

것으로 보아 각각의 특징적인 음악어법을 보유하고 있는 것으로 판단된다. 여기에서 특징적인 점은 사설 비교에서 강산제(JHS)가 거리가 멀지만 동편제 유성준제와 군집을 이루고 있는 것과는 달리 송만갑제인 {SMK, PBS}와 군집을 이루고 있다는 점이다.

유성준제에서는 사설 분석 결과와 마찬가지로 박동진(PTJ)과 그 외의 창자들이 각각의 군집을 이루고 있지만 그 거리는 다른 창자들 간의 거리보다 상당히 멀다. 이러한 사설과 음악어법 분석의 결과는 박동진(PTJ)이 유성준에게 배웠다고 전해지기는 하나 전승계보가 확실하지 않다는 기존의 논의를 뒷받침한다 하겠다. 박동진(PTJ)을 제외한 유성준제의 창자들은 <그림 6>, <그림 7>에서 볼 수 있듯이, 각각의 거리가 상대적으로 가깝게 나타나고 있어 사설 분석 결과와 동일한 모습을 보인다. 또한 임방울(IBU), 강도근(KTK), 박동진(PTJ) 유성준제 내에서 각각 별개의 군집에 속하는 것으로 그 거리가 측정되는 것으로 나타나 사설 분석 결과와 마찬가지로 특징적인 음악어법을 보유하고 있는 것으로 분석된다.

유성준제의 박초월 바디인 박초월(PCW), 남해성(NHS), 조통달(JTD), 김연수(KSY), 정광수 바디인 정광수(JKS), 김영자(KYJ), 김연수 바디인 김연수(KYS), 오정숙(OJS)은 사설과 마찬가지로 각각의 군집을 형성하고 있고, 각 바디 내에서 거리 역시 상대적으로 가까운 모습을 보인다. 이로 볼 때, 남해성(NHS), 조통달(JTD), 김수연(KSY), 김영자(KYJ), 오정숙(OJS)은 사설과 마찬가지로 음악어법에 있어서도 전대의 창자들보다는 스승의 음악어법을 거의 그대로 이어받고 있음을 알 수 있다. 박초월 바디 내에서는 사설과 마찬가지로 남해성(NHS)이 스승인 박초월(PCW)과 가장 가깝다.

이번에는 장단뿐만 아니라, 조, 리듬패턴, 붙임새를 포함시켜서 종합적으로 비교해 보았다.



<그림 8> 정간보 음악어법의 종합적 비교에 따른 창자별 거리

(4) 정간보 음악어법의 종합적 비교

$$\delta' = \sum_{k=1}^{m} dist(a_k, b_k) + k + r + b$$

k는 조, r은 리듬패턴, b는 붙임새이다. 이때, $k+r+b$의 값이 δ'의 $1/2$을 넘지 않도록 조정하였다. 이렇게 $k+r+b$의 값을 δ'의 $1/2$이 넘지 않도록 조정한 이유는, 음악어법을 반영하는 '조', '리듬패턴', '붙임새'는 한 장단의 12박 전체에 반영되어 있기 때문이다. 즉, 한 장단 내에서 박에 따라 배열된 사설이 있고, 그 사설에 초분절적으로 얹혀있는 정보가 '조', '리듬패턴', '붙임새'이기 때문이다. 이 분석에 따른 결과가 <그림 8>에 제시되었다. 장단 이외에 붙임새, 리듬패턴, 조 등의 음악적 요소를 추가해서 분석한 결과는 장단만을 이용해서 분석한 결과와 크게 다르지 않다. 역시, 송순섭과 박동진의 소리는 <그림 7>의 분석에서처럼 소속 유파에서 거리가 먼 것으로 측정되었다.

4. 맺음말

이 연구에서는 '전승과 변모'가 일어난 텍스트 간의 거리를 측정해서 분류(classification)하거나 군집(cluster)을 나누는 방법론을 실험해 보고, 판소리 수궁가의 '고고천변' 대목에 적용해서 그 결과가 기존의 판소리 전승 계보와 어느 정도 차이가 있고, 또한 전승 계보에서 동일한 유파로 분류되는 창자들 사이에는 어느 정도의 차이가 있는지를 측정해 보았다. 이 연구에서 다룬 유파는 크게 동편제 송만갑제, 유성준제, 서편제, 강산제이다. 전승과 변모라는 과정이 늘 그렇듯이, 각 유파별로 드러나는 차이가 균일하지 않다는 점이 이 연구를 통해서 확실하게 드러나게 되었다. 또한 유파의 분류와는 차이가 나는 현상들, 즉 동편제유성준제의 박동진과 동편제 송만갑제의 송순섭의 사설, 음악어법에 대한 종합적 분석을 통해서 이들은 자신이 속한 유파에서는 상대적으로 차이가 많이 나는 창자라는 것을 알게 되었다. 물론 이러한 유사성과 차이는 이미 알려진 사실들이 많지만, 그 차이가 구체적으로 얼마나 되는지를 정밀하게 측정하는 방법론은 처음 시도되었다.

여기서 사용한 방법론은 방언 연구, 이본 연구, 텍스트 비평 등의 다양한 분야에서 적용되었거나 적용될 수 있는 방법론이다. 이 방법의 적용 범위를 확장해서 고소설의 이본연구,[16] 또는 계통 관계에 존재하는 언어들의 거리 측정, 방언 연구 등에 적용하면서 인문학 분야에서의 전자 텍스트 구축과 그 활용의 범위를 넓히는 데 이 연구의 의의가 있다.

16) 이윤석(1997)에서는 홍길동전 필사본의 이본연구를 통해서 정본을 확립하는 것이 중요함을 강조하였다. 앞으로 홍길동전과 같은 다양한 이본이 존재하는 고소설의 분류에도 텍스트 군집 분석 방법을 적용해서 이본연구의 영역을 확장해 볼 계획이다.

참고문헌

권오경(2008),「고고천변(皐臯天邊)>의 존재양상과 기능 고찰」,『어문학』 제87호, pp. 317-344, 한국어문학회.

류수열(2002),「<수궁가> 소재 노정기의 존립과 변이: '고고천변'과 '범피중류' 및 '혼령상봉' 대목의 비교」,『판소리연구』 제14집, pp. 81-99, 판소리학회.

김미선(2001),『유성준제 <수궁가> 연구』, 이화여자대학교 석사학위논문.

김진영・김동건・김미선(2008),『정간보와 함께하는 김수연 창본 수궁가』, 이회문화사.

박황(1973),『판소리소사』, 신구문화사.

박황(1987),『판소리 이백년사』, 사사연.

안재원(2008),「서양고전문헌학의 방법론: 문헌 계보도, 편집, 번역, 주해」,『규장각』 32, pp. 257-282, 서울대학교 규장각 한국학연구원.

이윤석(1997),「새로 소개하는 <홍길동전> 이본 몇 가지」,『문학한글』 13, pp. 67-91, 한글학회.

인권환(1986),「수궁가의 형성과 창자의 전승 계보」,『배달말』 11, pp. 139-174, 배달말학회.

인권환(1992),「수궁가 동편제와 강산제」,『민족문화연구』 25호, pp. 39-69, 고려대학교 민족문화연구소.

Baayen, R. H.(2008), *Analyzing Linguistic Data: A practical introduction to statistics using R*, Cambridge: Cambridge University Press.

Barbrook, A. C., Howe, C. J., Blake, N. and Robinson, P.(1998), "The phylogeny of *The Canterbury Tales*", Nature, 394, p. 839, Macmillian Publishers Ltd.

Bilisoly, R.(2008), *Practical Text Mining with Perl*, John Wiley & Sons, Inc.

Csernel, M. and Patte, F.(2007), "Critical Edition of Sanskrit Texts", *Proceedings of First International Sanskrit Computational Linguistics Symposium*, Paris, Oct. pp. 103-121.

Heeringa, Wilbert(2004), *Measuring Dialect Pronunciation Differences using Levenshtein Distance*,

Ph.D. thesis, University of Groningen.

Jurafsky, D. and Martin, J. H.(2000), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall.

Kondrak, Grzegorz(2002), *Algorithms for Language Reconstruction*, Ph.D. thesis, University of Toronto.

Levenshtein, V. I.(1966), Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8), 707-710.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M.(2005), *Cluster Analysis Basics and Extensions; unpublished*.

Paradis E., Claude J. and Strimmer K.(2004), "APE: analyses of phylogenetics and evolution in R language", *Bioinformatics* 20, pp. 289~290.

R Development Core Team(2009), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org

원고 접수일: 2009년 9월 30일
심사 완료일: 2009년 11월 10일
게재 확정일: 2009년 11월 30일

ABSTRACT

A Study of Measuring Text Distances using the Hierarchical Clustering Method in Application to Pansori Narratives

Choi, Woonho · Kim Dong Keon

In this study, a computational method was used to measure the distance values between texts, especially to measure the variations which took place during the transmission of texts. "Gogochenbyeon (a scene of rabbit marvelling at the beautiful scenary of land world)" is called 'nundaemog' of Sugungga. 'Nundaemog' is one of the most essential and important parts of a Pansori, like arias of an opera. So many Pansori singers recorded those 'nundaemog', and therefore, 'gogochenbyeon' has been recorded by several famous singers. We selected 16 albums of 16 singers, and transcribed the narratives from those albums to the texts, and also put the musical expressions (rhythm, rhythm pattern, key, beat) on the Korean traditional musical score 'Jungganbo'. From these raw materials, the text-distances of the transcribed narratives were measured using Levenshtein distance, and the distances were normalized by the length of the narrative strings. And then the texts were clustered using hierarchical clustering methods and turned into the phylogenetic trees using neighbor-joining methods. After the result, we found some differences between the lineage

of the Pansori transmission and the real narratives that some singers like Park Tongjin and Song Sunseob are somewhat distant from their pupils or masters.