JCS Journal of Conservation Science

1991-2021 | (사)한국문화재보존과학회 30주년 기념
문화재를 살리는 기술
역사의 빛이 되는 보존과학

Research Article

# 옛한글 문서의 전자문서화와 정보공유 방법 제안

# Digitization of Old Korean Texts with Obsolete Korean Characters and Suggestion for Improvement of Information Sharing

김하영[1], 유우식[2,3,*]

[1]한국학중앙연구원 장서각, [2]웨이퍼마스터스, 미국 캘리포니아주 더블린시, [3]경북대학교 인문학술원

Ha Young Kim[1], Woo Sik Yoo[2,3,*]

[1]*Jangseogak, The Academy of Korean Studies, Seongnam 13455, Korea*

[2]*WaferMasters, Inc., Dublin, CA 94568, USA*

[3]*Institute of Humanities Studies, Kyungpook National University, Daegu 41566, Korea*

**Received** March 22, 2021
**Revised** May 26, 2021
**Accepted** May 26, 2021

***Corresponding author**
E-mail: woosik.yoo@wafermasters.com
Phone: +1-650-796-8396

Journal of Conservation Science 2021;37(3):255-269

https://doi.org/10.12654/JCS.2021.37.3.06

pISSN: 1225-5459, eISSN: 2287-9781



**초 록** 옛한글로 저술된 자료는 활자 인쇄본, 목판 인쇄본, 필사본, 고소설, 서간 등 방대한 자료가 한국학중앙연구원 장서각을 비롯하여 많은 기관에 소장되어 있다. 옛한글을 전산정보화하기 위해서는 수작업에 의한 '입력'과정이 필요하다. 옛한글 문서의 전자문서화 작업이 오랫동안 진행되어 왔으나 옛한글을 전공한 연구자 개인의 노력으로 옛한글을 읽고 입력하여 전자자료화 되고 있는 실정이다. 연구자의 숙련도가 개인적인 작업능력의 향상에 머무르고 기술의 축적으로 이어지지 못한다. 현재까지 극히 일부분의 옛한글 문서만이 소개되고 대부분의 자료는 수장고에 보관되어 있는 상태이다. 어렵게 전자문서화된 옛한글 고문서도 전자기기 간의 호환성 문제로 정보 공유 및 표시에도 어려움이 있다. 옛한글 문서의 전자문서화의 작업효율을 높이고 전자문서화 기술의 축적을 위해서는 옛한글의 입력, 표시, 저장 방법의 개선을 비롯하여 옛한글 문서의 이미지 분석을 통한 광학적 문자인식(OCR)의 개발이 필요하다.

**중심어** 옛한글, 옛한글 문서, 옛한글 글꼴의 비호환성, 전자문서화, 광학적 문자인식(OCR), 이미지분석

**ABSTRACT** A vast amount of materials—such as prints, woodblock prints, manuscripts, old novels, and letters—written in old Korean and using old grammar and/or obsolete characters, are collected in many institutions, including the Jangseogak at the Academy of Korean Studies. Digitization of these texts has required a prolonged manual inputting process. Individual researchers, who majored in old Korean, have read and typed the characters into electronic documents, which depends upon individual skill, effort, and approach, and is particularly limiting because none can be significantly increased. To date, only a small proportion of the old Korean document collections, currently kept in storage, have been digitized and made available to the public. Even the electronic formats of the texts prove difficult to displaying correctly, due to the incompatibility between the old Korean characters and the character set on today's electronic devices. To improve the techniques and efficiency of digitizing old Korean texts, it is necessary to develop optical character recognition (OCR), which will analyze images of old Korean documents, as well as input, display, and storage methods.

**Key Words** Old Korean characters, Old Korean documents, Old Korean font incompatibility, Digitization, Optical Character Recognition (OCR), Image analysis

## 1. 서 론

옛한글로 저술된 자료는 활자 인쇄본, 목판 인쇄본, 필사본, 고소설, 서간 등의 형태로 방대한 자료가 한국학중앙연구원 장서각을 비롯하여 많은 기관에 소장되어 있다. 옛한글 자료는 우리 말과 글의 변화과정과 시대상을 보여주는 소중한 문화유산이라고 할 수 있다. 옛한글로 저술된 자료에는 현대의 한글에서 사용되지 않는 자음, 모음, 받침 등이 사용되고 있으며 현대어처럼 띄어쓰기도 없이 위에서 아래로 쓰여진 종서형식을 취하고 오른쪽에서 왼

쪽으로 쓰여 있다. 옛한글 자료의 전자 정보화 및 전자문서화에 대한 필요성은 오래전부터 인식되어 왔으나(Kim, 1990) 정보처리 환경에 따라 전자문서화를 위한 입력자체가 불가능하거나 전자문서화된 자료도 정보처리 환경 간의 호환성의 제약으로 정보의 공유에 어려움이 발생하고 있다. 이러한 상황은 현재까지 해결되지 못한 채 이어지고 있다.

현재까지도 옛한글 자료의 전자문서화는 옛한글을 전공한 연구자 개인의 노력으로 옛한글을 읽고 수작업으로 입력하는 것이 유일한 방법이다. 옛한글의 전자정보화를 위한 입력과정은 현재 사용하지 않는 자음, 모음, 받침 등을 포함한 옛한글을 표준 키보드로 입력하는 것은 매우 복잡하여 많은 시간과 노력이 필요하다. 발음할 수 없거나 뜻을 알 수 없는 옛한글 문서를 수작업으로 입력하는 것은 입력의 난이도가 훨씬 높다. 옛한글 전공자의 수가 절대적으로 부족하고 새롭게 전자문서화에 참여할 수 있는 옛한글 전공자 육성의 전망도 그다지 밝은 편은 아니어서 극히 일부분의 옛한글 문서만이 소개되고 대부분의 자료는 수장고에 보관되어 있는 상태이다. 소장되어 있는 옛한글 자료의 전자문서화 작업의 기술축적과 고효율화를 통하여 옛한글 자료의 내용에 관한 심화연구가 이루어질 수 있는 환경을 만들어갈 필요가 있다. 인공지능(artificial intelligence, AI) 기반의 연구에 대한 기대도 높아지고 있으나 옛한글의 경우에는 전공자들조차 문자의 판독에 어려움을 겪는 경우가 적지 않아 자동문자인식에 필요한 옛한글 문자와 글꼴의 데이터베이스를 구축하는 것부터 장벽에 부딪히고 있는 실정이다. AI를 활용하려면 모든 옛한글에 대한 판독이 이루어져 옛한글의 자동 판독에 필요한 데이터베이스의 구축이 우선되어야 한다. 한국학중앙연구원 장서각에 소장된 옛한글의 문집자료는 총 455종류로 그중에서 5.1%에 해당하는 23종류의 자료만이 원문 텍스트로 입력이 완료된 상태이다. 대부분이 필사본으로 수려한 서체의 왕실자료부터 조악한 수준의 민간자료까지 다양한 서체가 존재하여 활자체의 광학적 문자인식(Optical Character Recognition, OCR) (Wikipedia, 2021a) 보다 난이도가 높다.

본 연구에서는 옛한글 자료의 전자문서화와 정보공유의 어려움의 구체적인 사례를 통하여 그 원인을 살펴보고 기존의 수작업에 의한 옛한글 문서 입력의 대안으로 광학적 문자인식 (OCR)방법을 제안한다. 이미지 분석을 통한 옛한글 자료의 광학적 문자인식에 필요한 기술적인 고려사항에 관하여 고찰한다. 본 연구에서는 AI를 활용한 옛한글의 광학적 문자 인식법의 개발에 앞서 서체의 개인차가 큰 필사본 옛한글 자료의 광학적 문자인식의 기술적 문제점의 파악과 광학적 문자인식의 요소기술의 개발과 시연에 관한 내용에 국한하여 소개한다.

## 2. 옛한글 자료의 전자문서화

옛한글 자료를 전자문서화하여 내용을 보존, 연구, 소개하기 위해서는 옛한글의 '입력'과정이 필수적이다. 모든 작업이 옛한글 연구자들이 수작업으로 입력이 이루어진다. 옛한글은 활자인쇄, 목판인쇄, 필사본 등의 다양한 형태로 존재하며 현대 한글에서 사용되지 않는 자음, 모음, 기호, 철자 등이 사용되어 발음하지 못하는 것도 많고 의미를 이해하기도 어려운 경우가 대부분이다. 따라서 수작업으로 입력하는 과정도 현대 한글을 한글 자판(keyboard)으로 입력하듯 쉽게 입력할 수 없어 작은 양의 옛한글 문서를 전자문서화하더라도 많은 시간과 노력이 필요하다. 필사본의 경우에는 필체가 각기 다르고 흘림체로 쓰여진 것이 많아 옛한글 전공자들의 경우에도 글자를 읽어내는 것조차 쉽지 않다. 활자로 인쇄된 옛한글 자료에도 한글과 한자가 함께 쓰여진 것과 한글 글자의 크기가 다른 것들이 함께 사용된 자료도 많다. 이러한 다양한 문제점들이 옛한글 문서의 전자문서화에 장애요인으로 작용한다. 옛한글 자료의 내용을 효율적으로 전자 문서화할 수 있는 새로운 입력방법의 개발이 필요하다.

### 2.1. 현대 한글과 옛한글

옛한글 원문 정보를 전자문서화하여 공유하기 위해서는 옛한글 자모(子母)를 정확하게 판독하여 입력할 수 있어야 한다. 현대 한글과 옛한글에서는 사용되나 현대 한글에서는 사용되지 않는 사용되는 자모의 차이를 Table 1에 정리하였다. 현대 한글에서 사용되지 않는 자음, 모음, 기호가 옛한글에서는 237개나 사용되고 있다. 이것은 현재까지의 연구에서 확인된 것이고 앞으로 더 많은 옛한글 자료를 조사하게 되면 새로운 자모가 추가될 가능성도 배제할 수 없다.

### 2.2. 한글 전자문서화의 어려움

아스키(ASCII: American Standard Code for Information Interchange, 미국 정보교환 표준부호) (Wikipedia, 2021b)는 영문 알파벳(alphabet)을 사용하는 대표적인 문자 부호화(encoding) 방법이다. 아스키는 컴퓨터, 인쇄기, 통신 장비를 비롯한 문자를 사용하는 각종 전자 장치에서 사용된다. 아스키는 1바이트(1 byte = 8 bit) 중에서 하위 7비트

**Table 1**. List of current and old Hangul (Korean) characters

| | | | |
|---|---|---|---|
| Current Hangul Characters | Single | Consonants (14) | ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ |
| | | Vowels (10) | ㅏ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅣ |
| | Double | Consonants (16) | ㄲ ㄳ ㄵ ㄶ ㄸ ㄺ ㄻ ㄼ ㄽ ㄾ ㄿ ㅀ ㅃ ㅄ ㅆ ㅉ |
| | | Vowels (11) | ㅐ ㅒ ㅔ ㅖ ㅘ ㅙ ㅚ ㅝ ㅞ ㅟ ㅢ |
| Old Obsolete Hangul Characters | Single | Consonants (9) | ㅿ ㆁ ㅱ ㅲ ㅳ ㅴ ㅵ ㅶ ㆆ |
| | | Vowel (1) | ㆍ |
| | | Symbols (2) | ㆍ ᠄ |
| | Double | Consonants (154) | ㄱㄱ ㄱㄴ ㄱㄷ ㄱㄹ ㄱㅁ ㄱㅂ ㄱㅅ ㄱㅇ ㄱㅈ ㄱㅊ ㄱㅋ ㄱㅌ ㄱㅍ ㄱㅎ ㄲㅏ ㄲㅓ ㄲㅗ ㄲㅜ ㄲㅡ ㄴㄱ ㄴㄲ ㄴㄴ ㄴㄷ ㄴㅂ ㄴㅅ ㄴㅈ ㄴㅎ ㄷㄱ ㄷㄲ ㄷㄷ ㄷㅂ ㄷㅅ ㄷㅈ ㄹㄱ ㄹㄲ ㄹㄴ ㄹㄷ ㄹㄹ ㄹㅁ ㄹㅂ ㄹㅅ ㄹㅈ ㄹㅋ ㄹㅌ ㄹㅍ ㄹㅎ ㅁㄱ ㅁㄲ ㅁㄴ ㅁㄷ ㅁㄹ ㅁㅁ ㅁㅂ ㅁㅅ ㅁㅈ ㅁㅋ ㅁㅌ ㅁㅍ ㅁㅎ ㅂㄱ ㅂㄲ ㅂㄴ ㅂㄷ ㅂㄹ ㅂㅁ ㅂㅂ ㅂㅅ ㅂㅈ ㅂㅋ ㅂㅌ ㅂㅍ ㅂㅎ ㅅㄱ ㅅㄲ ㅅㄴ ㅅㄷ ㅅㄹ ㅅㅁ ㅅㅂ ㅅㅅ ㅅㅈ ㅅㅋ ㅅㅌ ㅅㅍ ㅅㅎ ㅇㄱ ㅇㄲ ㅇㄴ ㅇㄷ ㅇㄹ ㅇㅁ ㅇㅂ ㅇㅅ ㅇㅈ ㅇㅋ ㅇㅌ ㅇㅍ ㅇㅎ ㅈㄱ ㅈㄲ ㅈㄴ ㅈㄷ ㅈㄹ ㅈㅁ ㅈㅂ ㅈㅅ ㅈㅈ ㅈㅋ ㅈㅌ ㅈㅍ ㅈㅎ ㅊㄱ ㅊㄲ ㅊㄴ ㅊㄷ ㅊㄹ ㅊㅁ ㅊㅂ ㅊㅅ ㅊㅈ ㅊㅋ ㅊㅌ ㅊㅍ ㅊㅎ ㅋㄱ ㅋㄲ ㅋㄴ ㅋㄷ ㅋㄹ ㅋㅁ ㅋㅂ ㅋㅅ ㅋㅈ ㅋㅋ ㅋㅌ ㅋㅍ ㅋㅎ ㅌㄱ ㅌㄲ ㅌㄴ ㅌㄷ ㅌㄹ ㅌㅁ ㅌㅂ ㅌㅅ ㅌㅈ ㅌㅋ ㅌㅌ ㅌㅍ ㅌㅎ ㅍㄱ ㅍㄲ ㅍㄴ ㅍㄷ ㅍㄹ ㅍㅁ ㅍㅂ ㅍㅅ ㅍㅈ ㅍㅋ ㅍㅌ ㅍㅍ ㅍㅎ ㅎㄱ ㅎㄲ ㅎㄴ ㅎㄷ ㅎㄹ ㅎㅁ ㅎㅂ ㅎㅅ ㅎㅈ ㅎㅋ ㅎㅌ ㅎㅍ ㅎㅎ |
| | | Vowels (71) | ㅏㅏ ㅏㅐ ㅏㅑ ㅏㅒ ㅏㅓ ㅏㅔ ㅏㅕ ㅏㅖ ㅏㅗ ㅏㅘ ㅏㅙ ㅏㅚ ㅏㅛ ㅏㅜ ㅏㅝ ㅏㅞ ㅏㅟ ㅏㅠ ㅏㅡ ㅏㅢ ㅏㅣ ㅐㅏ ㅐㅐ ㅐㅑ ㅐㅒ ㅐㅓ ㅐㅔ ㅐㅕ ㅐㅖ ㅐㅗ ㅐㅘ ㅐㅙ ㅐㅚ ㅐㅛ ㅐㅜ ㅐㅝ ㅐㅞ ㅐㅟ ㅐㅠ ㅐㅡ ㅐㅢ ㅐㅣ ㅑㅏ ㅑㅐ ㅑㅑ ㅑㅒ ㅑㅓ ㅑㅔ ㅑㅕ ㅑㅖ ㅑㅗ ㅑㅘ ㅑㅙ ㅑㅚ ㅑㅛ ㅑㅜ ㅑㅝ ㅑㅞ ㅑㅟ ㅑㅠ ㅑㅡ ㅑㅢ ㅑㅣ ㅒㅏ ㅒㅐ ㅒㅑ ㅒㅒ ㅒㅓ ㅒㅔ ㅒㅕ ㅒㅖ ㅒㅗ ㅒㅘ ㅒㅙ ㅒㅚ ㅒㅛ ㅒㅜ ㅒㅝ ㅒㅞ ㅒㅟ ㅒㅠ ㅒㅡ ㅒㅢ ㅒㅣ ㅓㅏ ㅓㅐ ㅓㅑ ㅓㅒ ㅓㅓ ㅓㅔ ㅓㅕ ㅓㅖ ㅓㅗ ㅓㅘ ㅓㅙ ㅓㅚ ㅓㅛ ㅓㅜ ㅓㅝ ㅓㅞ ㅓㅟ ㅓㅠ ㅓㅡ ㅓㅢ ㅓㅣ ㅔㅏ ㅔㅐ ㅔㅑ ㅔㅒ ㅔㅓ ㅔㅔ ㅔㅕ ㅔㅖ ㅔㅗ ㅔㅘ ㅔㅙ ㅔㅚ ㅔㅛ ㅔㅜ ㅔㅝ ㅔㅞ ㅔㅟ ㅔㅠ ㅔㅡ ㅔㅢ ㅔㅣ ㅕㅏ ㅕㅐ ㅕㅑ ㅕㅒ ㅕㅓ ㅕㅔ ㅕㅕ ㅕㅖ ㅕㅗ ㅕㅘ ㅕㅙ ㅕㅚ ㅕㅛ ㅕㅜ ㅕㅝ ㅕㅞ ㅕㅟ ㅕㅠ ㅕㅡ ㅕㅢ ㅕㅣ ㅖㅏ ㅖㅐ ㅖㅑ ㅖㅒ ㅖㅓ ㅖㅔ ㅖㅕ ㅖㅖ ㅖㅗ ㅖㅘ ㅖㅙ ㅖㅚ ㅖㅛ ㅖㅜ ㅖㅝ ㅖㅞ ㅖㅟ ㅖㅠ ㅖㅡ ㅖㅢ ㅖㅣ ㅗㅏ ㅗㅐ ㅗㅑ ㅗㅒ ㅗㅓ ㅗㅔ ㅗㅕ ㅗㅖ ㅗㅗ ㅗㅘ ㅗㅙ ㅗㅚ ㅗㅛ ㅗㅜ ㅗㅝ ㅗㅞ ㅗㅟ ㅗㅠ ㅗㅡ ㅗㅢ ㅗㅣ ㅘㅏ ㅘㅐ ㅘㅑ ㅘㅒ ㅘㅓ ㅘㅔ ㅘㅕ ㅘㅖ ㅘㅗ ㅘㅘ ㅘㅙ ㅘㅚ ㅘㅛ ㅘㅜ ㅘㅝ ㅘㅞ ㅘㅟ ㅘㅠ ㅘㅡ ㅘㅢ ㅘㅣ ㅙㅏ ㅙㅐ ㅙㅑ ㅙㅒ ㅙㅓ ㅙㅔ ㅙㅕ ㅙㅖ ㅙㅗ ㅙㅘ ㅙㅙ ㅙㅚ ㅙㅛ ㅙㅜ ㅙㅝ ㅙㅞ ㅙㅟ ㅙㅠ ㅙㅡ ㅙㅢ ㅙㅣ ㅚㅏ ㅚㅐ ㅚㅑ ㅚㅒ ㅚㅓ ㅚㅔ ㅚㅕ ㅚㅖ ㅚㅗ ㅚㅘ ㅚㅙ ㅚㅚ ㅚㅛ ㅚㅜ ㅚㅝ ㅚㅞ ㅚㅟ ㅚㅠ ㅚㅡ ㅚㅢ ㅚㅣ ㅛㅏ ㅛㅐ ㅛㅑ ㅛㅒ ㅛㅓ ㅛㅔ ㅛㅕ ㅛㅖ ㅛㅗ ㅛㅘ ㅛㅙ ㅛㅚ ㅛㅛ ㅛㅜ ㅛㅝ ㅛㅞ ㅛㅟ ㅛㅠ ㅛㅡ ㅛㅢ ㅛㅣ ㅜㅏ ㅜㅐ ㅜㅑ ㅜㅒ ㅜㅓ ㅜㅔ ㅜㅕ ㅜㅖ ㅜㅗ ㅜㅘ ㅜㅙ ㅜㅚ ㅜㅛ ㅜㅜ ㅜㅝ ㅜㅞ ㅜㅟ ㅜㅠ ㅜㅡ ㅜㅢ ㅜㅣ ㅝㅏ ㅝㅐ ㅝㅑ ㅝㅒ ㅝㅓ ㅝㅔ ㅝㅕ ㅝㅖ ㅝㅗ ㅝㅘ ㅝㅙ ㅝㅚ ㅝㅛ ㅝㅜ ㅝㅝ ㅝㅞ ㅝㅟ ㅝㅠ ㅝㅡ ㅝㅢ ㅝㅣ ㅞㅏ ㅞㅐ ㅞㅑ ㅞㅒ ㅞㅓ ㅞㅔ ㅞㅕ ㅞㅖ ㅞㅗ ㅞㅘ ㅞㅙ ㅞㅚ ㅞㅛ ㅞㅜ ㅞㅝ ㅞㅞ ㅞㅟ ㅞㅠ ㅞㅡ ㅞㅢ ㅞㅣ ㅟㅏ ㅟㅐ ㅟㅑ ㅟㅒ ㅟㅓ ㅟㅔ ㅟㅕ ㅟㅖ ㅟㅗ ㅟㅘ ㅟㅙ ㅟㅚ ㅟㅛ ㅟㅜ ㅟㅝ ㅟㅞ ㅟㅟ ㅟㅠ ㅟㅡ ㅟㅢ ㅟㅣ ㅠㅏ ㅠㅐ ㅠㅑ ㅠㅒ ㅠㅓ ㅠㅔ ㅠㅕ ㅠㅖ ㅠㅗ ㅠㅘ ㅠㅙ ㅠㅚ ㅠㅛ ㅠㅜ ㅠㅝ ㅠㅞ ㅠㅟ ㅠㅠ ㅠㅡ ㅠㅢ ㅠㅣ ㅡㅏ ㅡㅐ ㅡㅑ ㅡㅒ ㅡㅓ ㅡㅔ ㅡㅕ ㅡㅖ ㅡㅗ ㅡㅘ ㅡㅙ ㅡㅚ ㅡㅛ ㅡㅜ ㅡㅝ ㅡㅞ ㅡㅟ ㅡㅠ ㅡㅡ ㅡㅢ ㅡㅣ ㅢㅏ ㅢㅐ ㅢㅑ ㅢㅒ ㅢㅓ ㅢㅔ ㅢㅕ ㅢㅖ ㅢㅗ ㅢㅘ ㅢㅙ ㅢㅚ ㅢㅛ ㅢㅜ ㅢㅝ ㅢㅞ ㅢㅟ ㅢㅠ ㅢㅡ ㅢㅢ ㅢㅣ ㅣㅏ ㅣㅐ ㅣㅑ ㅣㅒ ㅣㅓ ㅣㅔ ㅣㅕ ㅣㅖ ㅣㅗ ㅣㅘ ㅣㅙ ㅣㅚ ㅣㅛ ㅣㅜ ㅣㅝ ㅣㅞ ㅣㅟ ㅣㅠ ㅣㅡ ㅣㅢ ㅣㅣ |

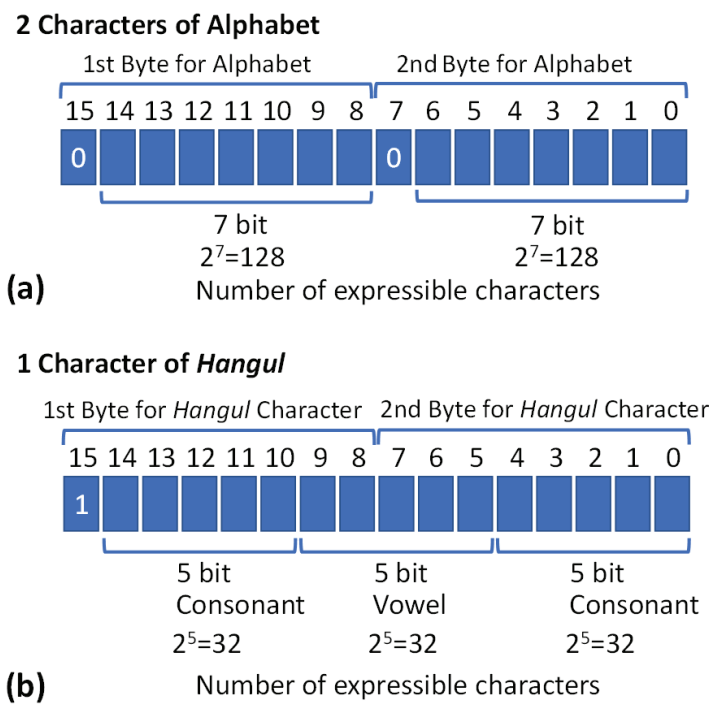


**Figure 1**. Structure comparison between (a) a 1 byte alphabet and (b) a 2 byte Korean character.

(bit)를 사용한 문자 부호화 방법으로, 52개의 영문 알파벳 대소문자와, 10개의 숫자, 32개의 특수 문자, 하나의 공백 문자와 33개의 출력 불가능한 제어 문자들로 총128개로 이루어져 있다. 1바이트 (8비트)단위의 데이터에서 앞에 0을 사용하고 하위 7비트로 128 ($2^7$=128)가지의 문자를 표시한다. 영문 알파벳을 대상으로 하는 경우 128가지의 조합으로 모든 문자를 표현할 수 있기 때문에 첫 번째 비트를 사용하지 않았지만 현재는 최상위 1비트는 오류 검출

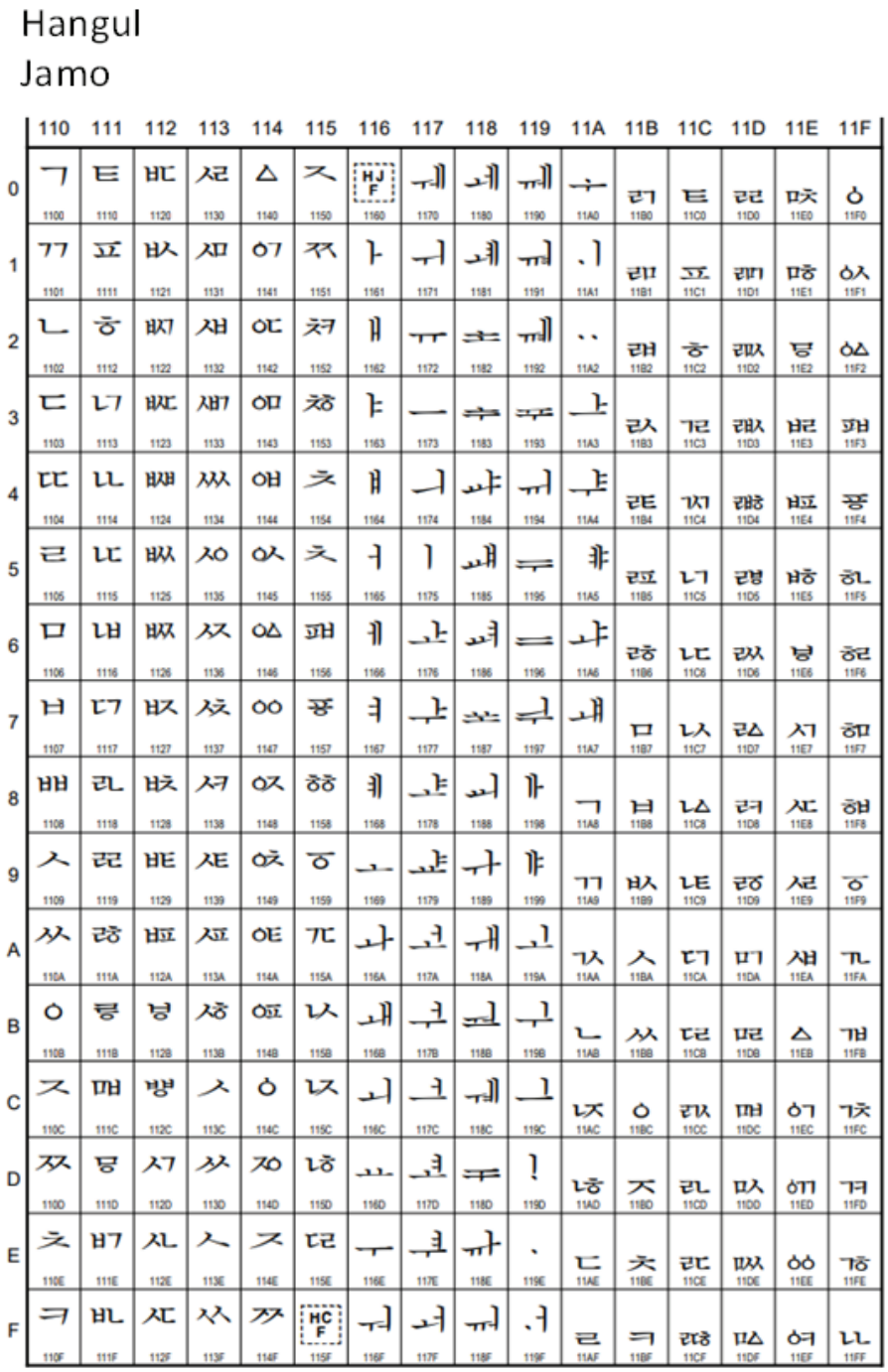


**Figure 2**. Unicode Hangul Jamo (Korean consonants and vowels) table (Unicode, 2021).

용 신호(Parity bit)로 사용되고 있다. 영문의 경우에는 26개의 알파벳의 대문자와 소문자를 합쳐도 52개 밖에 되지 않고 풀어쓰기로 되어 있기 때문에 7비트의 정보로 모든 문자를 표현할 수 있지만 한글의 경우에는 1바이트(8비트)의 부호로는 모든 문자를 표현할 수 없어 새로운 부호 체계가 필요하다.

한글의 부호화 방식은 크게 완성형과 조합형의 두 가지로 나뉜다(Kim, 1990). 완성형은 한글의 낱자를 독립된 문자로 인식하여 각각의 글자에 부호(code)를 부여하는 방식이고 조합형은 한글의 창제원리를 반영하여 초성, 중

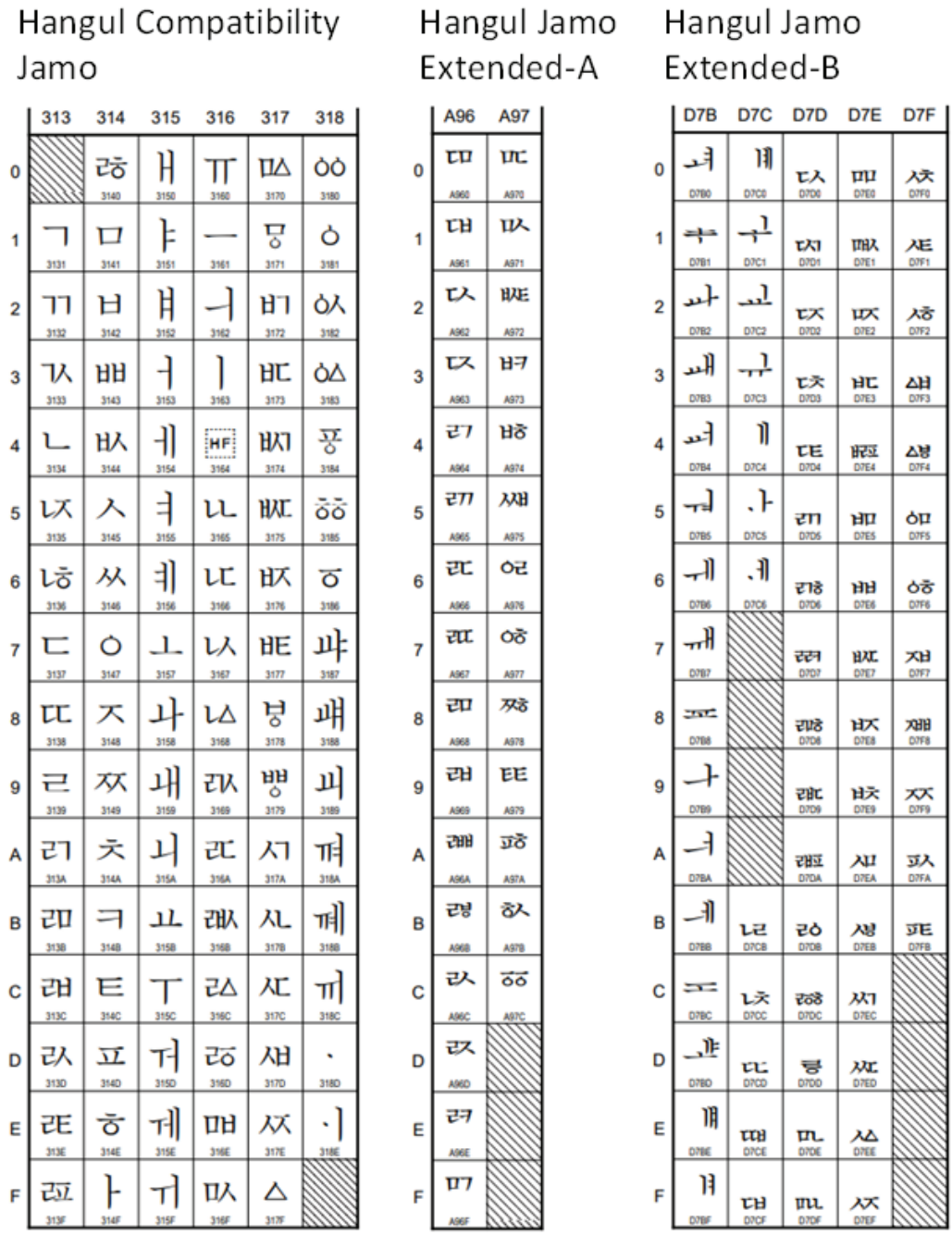


**Figure 3**. Unicode tables for Hangul Compatibility Jamo, Hangul Jamo Extended-A and Hangul Jamo Extended-B for old Korean documents (Unicode 2021).

성, 종성을 독립된 문자로 인식하여 자음과 모음의 조합으로 글자를 만들어내는 방식이다. 조합형은 한글의 창제원리가 잘 반영되었고 간단한 계산으로 초성, 중성, 종성을 분리할 수 있는 장점이 있지만 2바이트(16비트)를 최상위 1비트를 1로 하고 나머지 15비트를 5비트 단위로 분리해서 해석해야 하는 정보 처리상의 부담과 다른 문자체계들과 호환이 안 된다는 단점이 있다. 완성형은 한글의 창제원리를 무시한다는 단점이 있었지만 정보처리 능력의 제한 때문에 조합형의 단점이 큰 부담으로 작용하여 완성형이 국가표준으로 채택되어 사용되어 왔다. Figure 1에 1 바이트 알파벳과 2바이트 한글 조합형 문자의 신호구성과 표현 가능한 문자수를 정리하여 표시하였다.

정보처리능력에 부족했던 한글 전산화의 초기에 채택되었던 초기 완성형에서는 현대 한글로 표현할 수 있는 글자 중에서 사용 빈도가 높은 2,350자만을 지원하여 현대 한글의 자모음으로 조합가능한 한글 중에서 많은 글자가 표현되지 못했으며 심지어는 표준어에 있는 글자들도 전자문서상에서 표현될 수 없었다. '가'로 시작되는 초기 완성형에서 표현 가능한 조합은 '가각간갇갈갉갊감갑값갓갔강갖갗같갚갛'의 18자에 제한되었다. 초기 완성형으로는 비교적 간단한 조합형 글자인 '갘'자도 표현이 불가능하다. 이러한 문제점을 해결하기 위하여 유니코드(Unicode) 통합 완성형이라는 이름의 부호체계를 도입하여 기존의 2,350자료 표현이 불가능했던 8,822자를 추가하여 현대 한글에서 사용되는 11,172자를 모두 표현할 수 있게 되었다(Wikipedia, 2021c; 2021d). 그러나 기존의 초기완성형과 별도로 추가 지정되어 한글문자의 코드 값만으로는 문자가 논리적으로 정렬되지 않는 문제가 있는 상태로 현재까지 사용되고 있다.

현재의 유니코드에는 (통합)완성형의 11,172자와 Figure 2에 소개한 모든 자모의 조합형 한글 낱자가 모두 수록되어 있어 현대 한글과 옛한글을 모두 표현할 수 있고(Wikipedia, 2021c; 2021d; Unicode, 2021) 2바이트의 조합형 한글은 사용되지 않는다. 일반적인 현대 한글 문서에는 (통합)완성형을 사용하면 되지만, 옛한글 등 특수한 목적의 문서를 작성하기 위해서는 조합형(첫가끝)의 한글 낱자를 사용해야 한다. 조합형의 경우, 한글 자모 한글자가 유니코드 한 글자(3바이트)로 취급되기 때문에 데이터 길이가 3배로 늘어나는 문제가 발생하기 때문에 일반적인 현대 한글을 표현할 때는 이용되지 않고, 주로 옛한글을 표시할 때에만 이용된다. Figure 3과 Figure 4에 유니코드에 사용되는 한글 자모의 부호표를 예시하였다(Wikipedia, 2021d; Unicode, 2021). (통합)완성형과 조합형 낱자는 한 글자씩 따로 부호화되어 있고 그 조합은 무려 160만 자 이상에 달한다. 풀어쓰기식의 알파벳에 비해 한글의 전산정보화의 어려움이 여기에 있다. 한자의 수가 약 5만자이고 이체자를 포함해도 88,884자에 불과한 점과 비교해도 한글 문서의 전자정보화의 어려움은 쉽게 상상할 수 있다.

### 2.3. 옛한글 전자문서화를 위한 입력방법

옛한글의 입력에는 다양한 방법이 존재하며 크게 세 가지 유형으로 구별할 수 있다. 첫 번째 방법은 Figure 4에 예시한 옛한글 자판을 이용한 온라인 한글 입력기(OHI)(Online Hangul Input, 2021)와 Figure 5에 예시한 국립국어원이 제공하는 마우스로 원하는 글자를 선택하여 입력하는 옛한글 입력기이다. 두 번째 방법은 운영체계의 시스템에서 자판으로 입력된 신호를 옛한글로 변환하는 IME(Input Method Editor)로 마이크로소프트(Microsoft)가 제공하는 입력기와 날개셋 입력기(Wikipedia, 2021e) 등이 있다. 세 번째로는 국내에서 사용되는 가장 보편적인 방법으로 아래아 한글에서만 사용할 수 있는 흔글 입력기이다. 이처럼 옛한글의 입력방법마다 작동환경, 자판의 배열, 연구자 개인적인 선호도가 다르며 호환성에도 문제가 있다. 옛한글의 전자문서화에 성공하더라도 모니터나 프린터의 환경설정이 맞지 않으면 올바르게 표시되지 않는 등의 문제가 발생한다.

옛한글 문서는 활자나 목판으로 인쇄된 서적류부터 낱장으로 필사된 것까지 다양한 형태로 존재한다. 순 한글로 쓰여진 것, 한글과 한자가 함께 쓰여진 것, 한자에 한글을 병기한 것, 한글에 한자를 병기한 것, 한자와 한글의 배열과 크기가 복잡하게 쓰여진 것, 한글 편지, 한글 문서 등이 있다. 옛한글의 경우 현대 한글과 철자와 의미가 다른 경우가 많아 옛한글을 소리 내어 읽을 수 없는 경우도 많고 읽을 수 있다고 해도 그 뜻을 이해하기가 쉽지 않아 현대 한글을 입력하듯이 생각나는 대로 또는 읽은 대로 암송하면서 입력하는 것은 불가능에 가깝다. 흘림체 필사본의 경우에는 더욱 그러하다. 옛한글 자료의 연구에 있어서 입력작업이 차지하는 비중이 커서 자료의 내용을 연구하는 데 많은 부담으로 작용하고 어렵게 익힌 옛한글의 독해능력이 연구자 개인의 능력으로 머무를 뿐 계승되지 못하는 문제점이 있다. 이러한 문제를 해결하기 위해서는 옛한글 자료의 이미지를 바탕으로 광학적 문자인식(OCR)기술의 개발을 통한 새로운 옛한글 자동 입력방식의 구현이 필요하다.


**Figure 4**. A keyboard image for typing old Korean characters (Online Hangul Input, 2021).


**Figure 5**. Screen captured image of old Korean character input service online (National Institute of Korean Language, 2021).

### 2.4. 옛한글 자료의 온라인 열람과 검색

옛한글 자료를 온라인으로 검색하고 열람이 가능해지면 이상적이겠으나 옛한글을 필요로 하는 수요자체가 많지 않아 국제적으로 널리 사용되는 검색엔진에서 옛한글로 검색하는 것은 현재로서는 불가능하다. 우선 옛한글을 입력하는 것도 쉽지 않고 호환성의 문제로 모니터 화면상에서 옛한글로 표시하거나 프린터로 옛한글을 인쇄하는 것 또한 쉽지 않다.

현실적인 방법으로는 Figure 6에 예시한 한국학중앙연구원의 디지털장서각에서 온라인으로 제공하는 옛한글 자료의 이미지와 더불어 서지/해제, 본문(텍스트)의 XML(Extensible Markup Language) 자료가 좋은 예가 될 수 있다(The Academy of Korean Studies, 2021).

유이양문록(뉴니냥문녹[劉李兩門錄])의 첫 장의 이미지와 원문을 옛한글로 표기하고 있으며 한글과 더불어 한자도 병기하고 읽기 쉽게 띄어쓰기를 한 가로쓰기로 표시하고 있다. 이러한 온라인 검색과 열람을 가능하게 하려면 옛한글 원문의 입력과 독해까지 완성되어야 한다. 옛한글을 현대 한글로 번역한 자료를 열람할 수 있게 하는 것도 가능할 것이다. 이미지 정보는 손쉽게 제공할 수 있으나 옛한글 원문정보를 제공하려고 하면 옛한글 원문의 입력과정이 필수적이다. 연구자들 개인의 노력에 의한 옛한글 입력과정의 부담을 덜어 본문해석 등의 옛한글 자료의 내용연구에 보다 많은 시간을 할애할 수 있도록 하는 노력이 필요하다.

Figure 6의 오른쪽 옛한글 본문(기사보기, xml)에서 병기된 한자와 띄어쓰기를 생략하여 옛한글을 지원하는 글꼴(font)로 표시한 경우와 옛한글을 지원하지 않는 궁서체로 표시한 경우의 차이를 Figure 7에 예시하였다. 같은 내용의 본문을 현대 한글과 같이 가로쓰기로 예시한 것을 Figure 8에 표시하였다. 이처럼 옛한글이 바르게 입력된 전자문서라고 하더라도 전자기기의 사용환경에 따라서 의도한 대로 표시되지 않는다면 열람의 목적을 달성할 수 없게 된다. 따라서 어떤 전자기기의 운영체계의 환경 하에서도 호환성으로 인한 문제에 구애됨이 없이 본문을 열람할 수 있도록 이미지 파일로 변환해서 표시해 주는 방법도 함께 고려하는 것이 바람직하다(Figure 9). 세로쓰기, 가로쓰기의 변환도 가능하게 하고 한자와 띄어쓰기를 추가하거나 생략하는 것도 선택할 수 있게 하면 열람자의 다양한 요구에 대응할 수 있을 것이다(Figure 10). 이러한 기능을 구현하려면 열람 및 검색시스템의 구상단계에서부터 다양한 서비스를 제공하기 위하여 필요한 기능을 정리하여 설계하는 실행에 옮기는 작업이 필요하다.

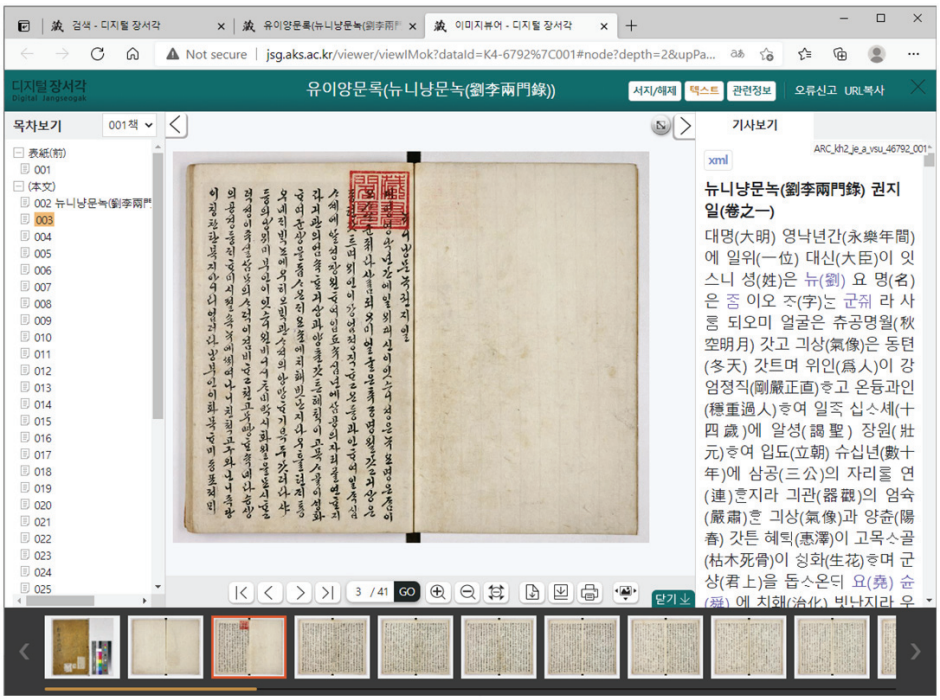


**Figure 6**. Screen captured image of web service from Digital Archive of Jangseogak, The Academy of Korean Studies.

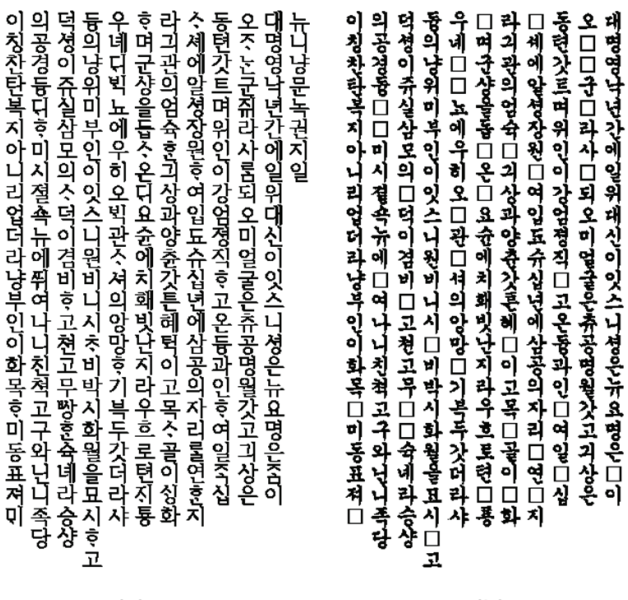


**Figure 7**. Old Hangul text in vertical text direction. (a) correctly displayed old Korean text and (b) incorrectly displayed old Korean text due to incompatibility.

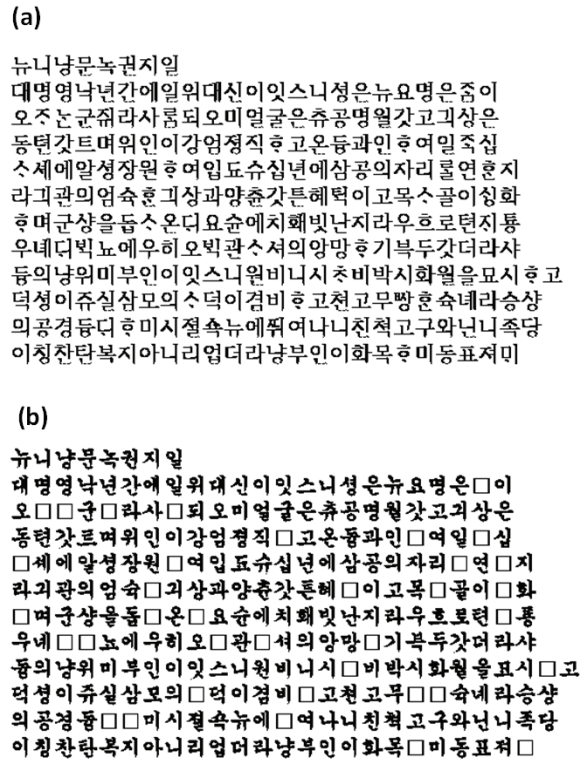


**Figure 8**. Old Hangul text in horizontal text direction. (a) correctly displayed old Korean text and (b) incorrectly displayed old Korean text due to incompatibility.


**Figure 9**. (a) Old Hangul text image after removing unnecessary mark (red library seal) and (b) vertically displayed old Hangul text.


**Figure 10**. Old Hangul text with supplemental information in Chinese characters and intentional spacings between words. (a) horizontally displayed text and (b) vertically displayed text.

## 3. 한글의 광학적 문자인식(OCR)

본 연구에서는 옛한글 자료의 이미지로부터 광학적 문자인식(OCR)기술을 통하여 옛한글 입력방법을 편리하게 할 수 있는지에 관하여 그 가능성을 조사하였다. 사진 또는 스캔 이미지에서 임의의 부분의 한글을 OCR을 통해서 한글을 추출할 수 있도록 하는 기능을 구현하는 것을 목표로 연구를 진행하였다. 이미지 프로세싱 소프트웨어로는 WaferMasters, Inc.(California, USA.)의 PicMan (Kim *et al*., 2019; Yoo, 2020; Yoo and Yoo, 2021)을 사용하여 선택된 영역 내의 한글을 가로 또는 세로방향으로 읽어낼 수 있는 기능을 구현할 수 있는지를 조사하였다. Figure 11에 한국학중앙연구원 앞의 버스 정거장의 표지판의 사진에서 PicMan으로 한글을 추출한 예를 들었다. 활자체의 현대 한글의 경우에는 글자의 크기, 해상도, 명암대비가 확보되면 세로쓰기, 가로쓰기 모두 문제없이 한글을 추출할 수 있었다. 띄어쓰기가 있는 경우, 구독점이나 기호가 있는 경우, 궁서체처럼 글씨의 굵기에 변화가 큰 경우, 세로쓰기의 경우, 글씨가 작은 경우, 명암대비가 작은 경우 등은 한글을 잘못 읽는 경우가 증가하는 경향이 나타났다. 이러한 현상은 PicMan에서 한글 OCR에 사용한 데이터베이스가 가로쓰기의 활자체를 중심으로 만들어졌기 때문이다.

Figure 12에 마이크로소프트(Microsoft사)의 Word에서 바탕체, 맑은 고딕체, 궁서체, 굴림체로 폰트 크기(font size)를 11pt와 16pt로 보통 활자체와 굵은 활자체로 글자와 글자 간에 인위적으로 간격을 띄어 (a) 가로 또는 (b) 세로방향으로 표시한 화면을 96dpi (dot per inch)의 이미지로 저장한 것과 (c) 한국학 중앙연구원 장서각 왕실문헌 자료인 유이양문록(뉴니냥문녹(劉李兩門錄))의 첫 장의 마지막 부분의 세로쓰기 부분을 OCR로 문자인식을 시도한 결과를 정리하였다. 활자체의 경우, 가로쓰기와 세로쓰기 모두 16종류의 테스트 패턴에서 2종류에서만 100%의 인식률을 얻었으므로 12.5%의 성공률을 얻은 셈이다. 글자단위로 인식률을 계산하면 약 85%의 인식률을 얻을 수 있었다. 글자의 크기에 따라서 'ㅇ'을 'ㅁ'으로 읽거나 'ㅏ'를 'ㅑ'로 인식하는 등 비슷한 모양의 음소로 인식하는 경우가 많았다. 또한 '앙'을 세로읽기로 할 경우 '0 1 0' 또는 '0 1 °'등의 숫자 또는 숫자와 기호의 조합으로 인식하는 경우도 있었다. 활자체의 경우에도 글자의 크기가 작거나 글자의 모양에 멋을 낸 경우와 음소(자음, 모음)끼리 너무 근접해 있는 경우 문자 인식의 오류가 증가하는 경향이 있었다. 글자의 크기, 이미지의 해상도가 OCR의 성능에 장애가 되지 않는 범위를 조사하여 활용하게 되면 활자체의 문자 인식률을 높일 수 있을 것이다. 필사본의 경우 글자체가 수려하고 고른 왕실문헌의 경우에도 데이터베이스가 준비되어 있지 않아 거의 모든 글자를 올바르게 인식하지 못하였다. 필사본의 특징상 글자와

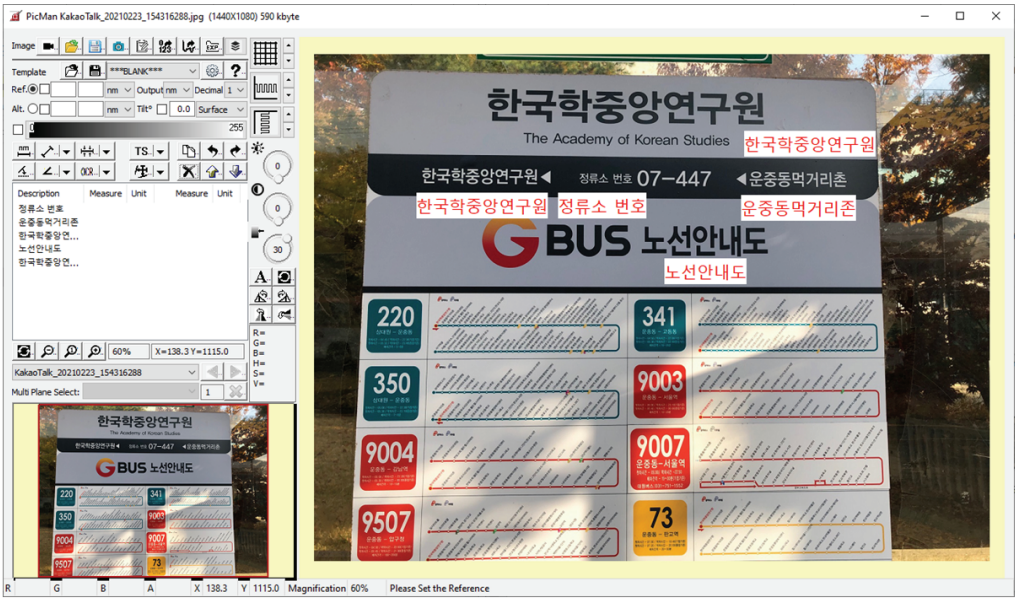


**Figure 11**. An example of Optical Character Recognition (OCR) of Korean characters from a photograph of a posted sign at the bus stop in front of The Academy of Korean Studies using image analysis software, PicMan.

글자 사이에도 가는 먹선으로 연결되어 있어 글자와 글자를 구분하는 것에도 어려움이 있으며 옛한글 자모의 데이터베이스가 준비되지 않아 발생한 문제이다. 민간자료의 경우에는 조악한 서체의 자료가 많고 일부는 활자 인쇄본도 있어 OCR인식의 대상을 결정하여 개발을 진행하는 것이 필요해 보인다. 참고로 현재 한국학중앙연구원 장서각에는 옛한글 문집이 455종이 소장되어 있으며 왕실자료가 107종, 민간자료가 348종이다. 원문 텍스트까지 입력된 자료는 모두 왕실문헌으로 23종에 불과하다. 왕실자료의 21.5%, 전체자료의 5.1%만이 원문 텍스트의 입력이 완료된 상태이다. 옛한글 자료를 소장한 다른 기관의 경우까지 고려하면 옛한글 문서의 전자문서화 작업의 양은 옛

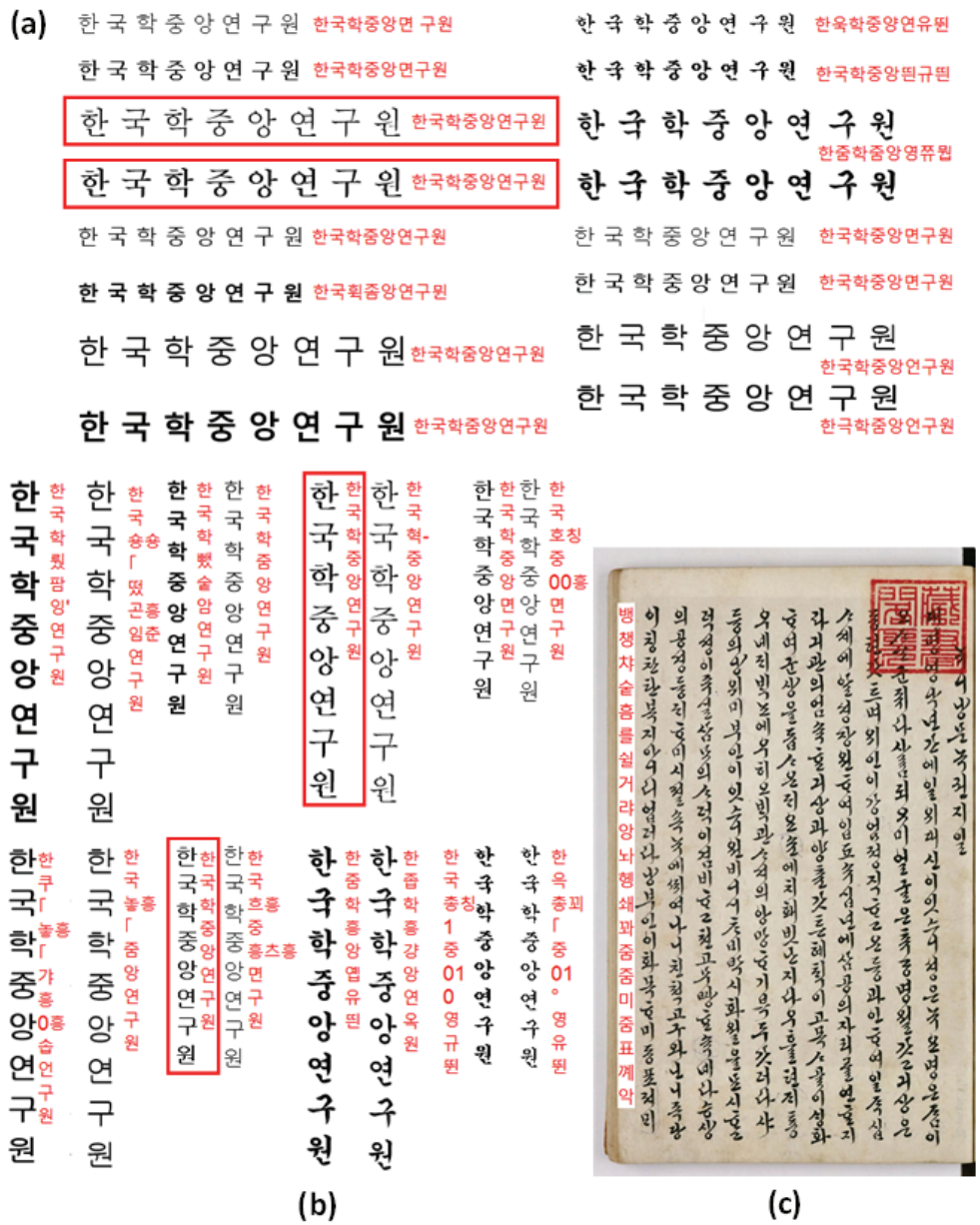


**Figure 12**. Trial examples of Optical Character Recognition (OCR) of Korean characters with various fonts (Batang, Malgun Gothic, Gungsuh and Gulim), sizes (11 pt and 16 pt) and thicknesses (normal and bold) in (a) horizontal direction and (b) vertical directions and (c) an old Korean hand written document in vertical direction using image analysis software, PicMan. Perfectly recognized words were highlighted by red boxes.

한글 연구자들의 개인적인 노력으로 수작업에 의한 입력만으로는 감당하기 어려운 실정이다.

비교를 위하여 OCR로 알파벳, 일본어, 일본어 한자, 중국어 번자체(繁字體), 중국어 간자체(簡字體)의 문자 인식도 시도해 보았다. 문자의 종류에 따라서 종류에 맞는 데이터베이스를 선택해야 하는 번거로움은 있었지만 활자체의 경우 약 70~80% 정도의 인식률을 얻을 수 있었다. 인식률이 너무 낮으면 문서의 수정작업에 더 많은 작업시간을 필요로 하기 때문에 실무에서 OCR기술을 활용하려면 70% 이상의 인식률의 정확도가 요구된다. 앞에서도 언급한 바와 같이 옛한글 문서의 형태는 활자나 목판으로 인쇄된 서적류부터 낱장으로 필사된 것까지 다양하다. 문서의 내용도 순 한글로 쓰여진 것, 한글과 한자가 함께 쓰여진 것, 한자에 한글을 병기한 것, 한글에 한자를 병기한 것, 한자와 한글의 배열과 크기가 복잡하게 쓰여진 것, 한글 편지, 한글 문서 등이 존재하므로 PicMan에서 임의의 지정된 영역의 한글과 한자의 문자정보를 추출할 수 있는 기능은 매우 유용하게 활용될 수 있을 것으로 기대된다. 글자단위, 단어단위, 행 또는 열 단위, 문장단위 또는 페이지 단위로 영역을 지정하여 문자를 추출하게 되면 문자의 오인식률을 줄이고 작업 효율을 높이는 데 유리하게 작용하게 된다.

## 4. 고찰 및 결론

OCR에 의한 한자 및 한자 초서의 판독에 관한 연구는 오래전부터 국내외의 여러 기관에서 진행되어 왔다. 연구자 개개인의 판독능력에 의존하지 않고 데이터베이스를 축적하여 활용함으로써 작업 능률을 높이고 오류를 줄이기 위한 노력의 일환으로 전자기술과 정보화기술을 활용한 좋은 사례이다. 경북대학교 Digital Lachiveum (Kyungpook National University, 2021)에서 한자 초서체 문서의 탈초작업과 번역작업이 시도되어 일정한 성과를 얻고 있으며 일본의 나라문화재연구소(奈良文化財研究所)에서는 목간(木簡)과 흘림체 문자의 해독 시스템으로 개발된 MOJIZO (Nara National Research Institute for Cultural Properties, 2021)라고 하는 온라인 서비스를 통해서 나라문화재연구소가 소장하고 있는 데이터와 일본의 동경대학(東京大学) 사료편찬소(史料編纂所)가 소장하고 있는 데이터를 사용하여 업로드된 이미지와 가장 유사한 글자의 후보군을 표시해 주고 있다. 경북대학교의 Digital Lachiveum은 OCR로 판독하고자 하는 이미지를 한 번에 업로드하면 자동으로 글자와 글자를 분리하여 판독하고 해석 예문을 제시한다. 나라문화재연구소의 MOJIZO의 경우에는 글자를 하나씩 분리해서 업로드해 주어야 한다. 두 경우 모두 아직까지는 매우 만족스러운 결과는 얻지 못하고 있지만 앞으로 지속적으로 데이터를 축적해 가게 되면 상당한 수준의 개선이 이루어 질 것으로 기대된다.

본 연구에서도 전자기술과 정보화기술을 옛한글 자료에 적용하여 입력작업효율의 향상과 옛한글 판독기술 축적의 가능성을 살펴보았다. 옛한글로 저술된 다양한 자료의 형태와 전자문서화와 공유의 어려움에 관하여 옛한글의 입력방식, 전자기기 간의 비호환성으로 인한 여러 가지 문제에 관하여 살펴보았다. 전자문서화의 가장 큰 걸림돌인 옛한글의 입력방법을 기존의 옛한글 연구자 개인의 노력에 의한 수동 입력방법에서 옛한글 문서의 스캔 또는 사진 이미지로부터 이미지 프로세싱 소프트웨어에 옛한글을 OCR로 인식하여 활용하는 방법을 제안하고 그 첫 단계로 이미지 상의 임의의 영역 내의 한글을 추출하는 기능을 구현하여 광학적 문자인식기능의 가능성을 시험하였다. 한글 및 옛한글 이미지를 사용하여 OCR로 한글 인식을 시험한 결과 활자체의 현대 한글의 경우에는 글자의 크기, 해상도, 명암대비가 확보되면 세로쓰기와 가로쓰기 모두 문제없이 한글을 추출할 수 있었다. 띄어쓰기가 있는 경우, 구독점이나 기호가 있는 경우, 궁서체처럼 글씨의 굵기에 변화가 큰 경우, 세로쓰기의 경우, 글씨가 작은 경우와 명암대비가 작은 경우 등은 추출된 한글에 오류가 발생하는 비율이 높아지는 경향이 나타났다.

옛한글의 경우에는 아직 학습 데이터가 준비되지 않아 읽을 수 없었으나 옛한글 문서 중에서 현대 한글과 같은 글자의 경우에는 큰 문제없이 한글을 읽어낼 수 있었다. 옛한글과 한자가 함께 사용된 문서의 경우에도 OCR기능을 적용할 수 있는지를 확인하기 위하여 일본어의 한자와 중국어 번자체와 간자체도 OCR로 인식이 가능한지 시험하여 실용적으로 활용할 수 있는 수준의 결과를 얻는 데 성공하였다. 다만 한글과 한자가 함께 사용된 문서의 경우 한글과 한자의 영역을 미리 지정해서 적용할 학습된 데이터베이스를 지정해줘야 하는 번거로움이 개선해야 할 문제점으로 남아 있다.

옛한글의 경우 전자기기 간의 호환성의 문제로 전자문서화된 옛한글이라도 올바르게 표시되지 못하는 경우가 빈번하게 발생하는데 이러한 경우에는 전자문서화된 옛한글 문서를 Figure 9과 Figure 10에 예시한 바와 같이 텍스트와 텍스트를 이미지로 변환하여 제공하는 방법으로 비호환성의 문제를 해결하는 방법을 제안하였다.

옛한글을 OCR로 인식하기 위해서는 지금까지 전자문서로 변환된 옛한글 자료와 그 원문 이미지를 활용하여 옛한글의 OCR인식에 필요한 학습 데이터베이스를 현대

한글의 OCR인식에 활용하고 있는 학습 데이터베이스처럼 구축하여 활용하면 충분히 실현이 가능할 것으로 생각된다. 많은 학습자료를 사용하게 되면 옛한글의 판독률이 높아지게 되어 옛한글 연구자 개개인의 판독능력에 의존하는 한계를 극복하고 옛한글 자료의 판독기술의 축적으로 이어질 수 있을 것으로 기대된다. 이러한 작업은 개인 단위의 노력으로 성취할 수 있는 성격은 아니지만 중장기적인 안목을 가지고 기획하고 실행에 옮기면 충분히 가능할 것으로 생각된다. 이러한 노력이 방대한 옛한글 자료의 내용을 옛한글 연구자들만의 학문적 호기심의 대상이 아닌 문화적 콘텐츠로 활용될 수 있는 계기를 마련하는데 반드시 필요하다고 할 수 있을 것이다. 옛한글 문서의 판독이 진행되어 많은 옛한글 문자와 글꼴에 대한 이미지 정보가 확보되면 AI기술을 적용하여 옛한글 문자의 이미지 정보만으로도 옛한글 문자의 판독이 전문가의 수작업에 의한 입력에 버금가는 상당한 수준까지 이루어 질 수 있을 것으로 기대된다. 그렇게 하기 위해서는 무엇보다 전문가에 의한 옛한글 문자의 정확한 판독과 판독된 옛한글 문자 이미지의 데이터베이스의 구축이 이루어져야 한다. 이것은 AI에서는 판독된 정답을 알고 있는 이미지를 이용한 학습을 통한 데이터베이스의 구축이 선행되어야 하기 때문이다.

이미지 분석 소프트웨어 PicMan의 다양한 기능을 활용하게 되면 OCR에 의한 문자인식 이외에도 문서 상의 글자의 윤곽을 추출하여 글꼴에 관한 정보도 자료로 수집하여 문화재 정보의 전자매체 상의 기록과 보존이 가능하다. 하나의 소프트웨어를 사용하여 OCR에 의한 문자인식을 통한 전자문서화와 교정기능과 더불어 문서의 변색, 손상 등의 정보를 정량적으로 평가하고 기록할 수 있어 보존과학 분야에서의 활용이 가능하다(Kim *et al*., 2019; Yoo, 2020; Yoo and Yoo, 2021; Yoo *et al*., 2021). 수리 전의 사전조사와 기록은 물론 수리 전후의 변화를 비교하고 기록할 수 있는 장점도 있다. 문화재 보존과학 분야의 응용에 있어서도 현장에서 필요한 맞춤형 기능을 추가하고 보완해 가면 매우 유용하게 활용될 수 있을 것으로 기대된다.

## 사 사

본 연구를 수행함에 있어서 많은 격려와 지원을 해 주신 한국학중앙연구원 장서각 왕실문헌연구실 김덕수 실장님께 감사드립니다. 기술적인 협력을 주신 미국 WaferMasters, Inc.의 강기택 씨와 김정곤 박사님, 프로젝트의 전반적인 진행과정에 도움을 주신 (주)휴니텍 유지현님께도 사의를 표합니다.

## REFERENCES

Kim, G., Kim, J.G., Kang K. and Yoo, W.S., 2019, Image-based quantitative analysis of foxing stains on old printed paper documents. Heritage, 2(3), 2665-2677. (in English)

Kim, H.G., 1990, A study on the composition of *Hunminjeongeum* code system and computer processing method for the computerization of classical data in Korean and Korean literature, Korean Culture Research, 23, 145-187. (in Korean)

Kyungpook National University, 2021, http://www.dila.co.kr/bbs/write.php?bo_table=opentrans (March 22, 2021)

Nara National Research Institute for Cultural Properties, 2021, https://mojizo.nabunken.go.jp/ (March 22, 2021)

National Institute of Korean Language, 2021, Old Korean Input System, https://www.korean.go.kr/common/oldHangeul.do (March 22, 2021)

Online Hangul Input, 2021, Online Hangul Input - Dubeolsik Old *Hangul* Keyboard, https://pat.im/1179 (March 22,2021)

The Academy of Korean Studies, 2021, *Yu-Yi Yangmunrok*, http://jsg.aks.ac.kr/viewer/viewIMok?dataId=K4-6792%7C001#node?depth=2&upPath=001&dataId=001 (March 22, 2021)

Unicode, 2021, Hangul Jamo, https://www.unicode.org/charts/PDF/U1100.pdf (March 22, 2021)

Wikipedia, 2021a, Optical character recognition, https://en.wikipedia.org/wiki/Optical_character_recognition (march 22, 2021)

Wikipedia, 2021b, ASCII, https://en.wikipedia.org/wiki/ASCII, (March 22, 2021)

Wikipedia, 2021c, Korean language and computers, https://en.wikipedia.org/wiki/Korean_language_and_computers (March 22, 2021)

Wikipedia, 2021d, Hangul Jamo (Unicode block), https://en.wikipedia.org/wiki/Hangul_Jamo_(Unicode_block) (March 22, 2021)

Wikipedia, 2021e, Nalgaeset *Hangul* Input, https://ko.wikipedia.org/wiki/%EB%82%A0%EA%B0%9C%EC%85%8B_%ED%95%9C%EA%B8%80_%EC%9E%85%EB%A0%A5%EA%B8%B0 (in Korean) (March 22, 2021)

Yoo, W.S., 2020, Comparison of outlines by image analysis for derivation of objective validation results: “Ito Hirobumi’s characters on the foundation stone” of the Main Building of Bank of Korea. Journal of Conservation Science, 36(6), 511-518. (in Korean with English abstract)

Yoo, Y. and Yoo, W.S., 2021, Digital image comparisons

for investigating aging effects and artificial modifications using image analysis software. Journal of Conservation Science, 37(1), 1-12.

Yoo, W.S., Yoo, S.S., Yoo, B H. and Yoo, S.J., 2021, Investigation on the conservation status of the 50-year-old "Yu Kil-Chun Archives" and an effective and practical method of preserving and sharing contents. Journal of Conservation Science, 37(2), 167-178. (in Korean with English abstract)