한글 필사본 음식조리서 말뭉치 구축을 위한 마크업 방안 연구

안의정· 박진양· 남길임*†
경북대학교

Eui-jeong Ahn, Jin-yang Park, and Kil-im Nam. 2008. A Study on the Markup Scheme for Building the Corpora of Korean Culinary Manuscripts. *Language and Information 12.2*, 95–114. This study aims at establishing a markup system for 17-19th century culinary manuscripts. To achieve this aim, we, in section 2, look into various theoretical considerations regarding encoding large-scale historical corpora. In section 3, we identify and analyze the characteristics of textual theme and structure of our source text. Section 4 proposes a markup scheme based on the XML standard for bibliographical and structural markups for the corpus as well as the grammatical annotations. We show that it is highly desirable to use XML-based markup system since it is extremely powerful and flexible in its expressiveness and scalable. The markup scheme we suggest is a modified and extended version of the TEI-P5 to accommodate the textual and linguistic characteristics of pre-modern Korean culinary manuscripts. (**Kyungbook National University**)

Key words: 필사본 음식조리서 (culinary manuscript), 말뭉치 (corpus), 마크업 (markup), TEI, XML

1. 서론

이 연구는 조선 후기 한글 필사본 음식조리서를 기반으로 조리 용어 검색 시스템을 개발하기 위한 연구 과제의 하위 연구로, 본 연구에서는 말뭉치 구축에서 가장 기본적인 과제라 할 수 있는 마크업 방안에 대해 음식조리서 텍스트를 대상으로 논의하고자 한다. '마크업 (markup)'은 드러나지 않는 텍스트의 속성을 부호화하여 텍스트에 삽입하는 것을 말하는데, 말뭉치 구축에 있어서는 텍스트의 속성을 체계적이고 명시적으로

* 702-701 대구광역시 북구 산격동 1370 경북대학교 인문대학 국어국문학과, Email: ejahn@lex.yonsei.ac.kr, parkjinyang@empal.com, nki@knu.ac.kr

† 이 연구는 학술진흥재단(2007년 기초연구과제-인문사회 분야)의 지원을 받았으며, 과제명은 '조선시대 한글 음식조리서 연구를 통한 조리 용어 통합 검색 시스템 개발'이다.(KRF-2007-322-A00050) 논문을 꼼꼼히 검토하고 유익한 도움말을 주신 심사위원분들께 진심으로 감사드린다.

드러내는 데 결정적인 역할을 한다.[1] 특히 본 연구의 주요 대상인 필사본 음식조리서는 현대 문어 텍스트와 달리, 필사본이라는 특수성을 가진 동시에 역사 자료이며, 특정 시기의 '음식조리서'라는 텍스트 속성으로 인해 텍스트의 제작 방식과 역사성, 텍스트 주제적·구조적 특성들이 마크업 과정에 적절하게 반영되어야 할 필요가 있다.

필사본이란 제작 방식에 따른 분류로 손으로 쓰인 책을 의미하며, 이 중 한글로 표기된 자료를 한글 필사본이라 한다.[2] 그 동안 국어사 문헌의 말뭉치 구축은 '21세기 세종계획(이하 '세종계획')'의 특수 말뭉치 구축 분과의 세부 과제 중 하나인 '역사 자료 말뭉치 개발' 부문에서 이루어진 바 있다. 세종계획의 결과물인 역사 자료 말뭉치는 15세기부터 개화기까지의 중요한 국어사 문헌이 원시 말뭉치 또는 형태 주석 말뭉치로 구축되어 있으나, 주로 판본 역사 자료와 각 시기의 주요 저작물을 대상으로 하였기 때문에, 한글 필사본 자료나 부녀자들의 생활문 등 다양한 텍스트 장르들이 포함되지 못하였다.

한글 필사본 자료는 텍스트의 내용이 다양하고 당대 현실을 반영하는 사실적인 텍스트라는 점에서 방대한 자료를 대상으로 하는 국어사 연구에서 중요한 위치를 차지한다. 특히 본 연구의 대상인 필사본 한글 음식조리서는 전통 생활 어휘나 조리 용어가 풍부하여, 국어학이나 조리학 전공자뿐 아니라 이 분야에 관심을 가지는 일반인들에게도 유용한 자료로 활용될 수 있다는 활용상의 강점이 있다. 이러한 한글 필사본 자료의 마크업 방법에 대한 연구는 다양한 장르의 국어사 자료의 구축을 위한 기초 연구로서, 전산 기술을 활용하는 국어사 연구의 한 전기를 마련한다는 의의를 가진다. 본 연구에서는 마크업 언어로 XML(eXtensible Markup Language)을 이용하고자 하며, 형태 정보 주석뿐만 아니라 의미 정보 주석이나 기타 미시적 정보들까지도 체계적으로 하나의 파일 내에 포함되도록 구조화할 것이다. 또, 향후 다른 연구자들이 기본적인 주석 내용을 자신의 목적에 맞게 가공하여 연구하거나, 보다 심화된 주석으로 확장하는 것이 용이하도록 설계할 것이다.

본 연구는 조리 용어 검색 시스템을 개발하기 위해 필요한 마크업 방안을 논의함으로써, 향후 검색 시스템의 효율성과 접근성을 최대화하는 데 기여하고자 한다. 본 연구의 마크업은 필사본 음식조리서의 말뭉치 구축 단계에 따라 다음의 세 가지 단계로 구분하여 논의될 수 있다. 첫째, 전문학자들이 판독하고 해석한 필사본 자료를 원시 말뭉치로 가공하는 단계, 둘째, 원시 말뭉치의 효율적인 활용을 위해 현대어역 말뭉치를 병렬 말뭉치로 구축하는 단계, 셋째, 형태·의미 주석 말뭉치로 구축하는 단계가 그것이다.

이를 위해 2장에서는 마크업에 대한 이론적인 논의로, 마크업의 정의와 중요성, 요건 등을 살필 것이다. 3장에서는 본 연구가 대상으로 하는 한글 필사본 음식조리서 텍스트의 특성에 대해 검토하고, 이러한 특성을 반영하기 위한 마크업 방안에 대해 논

[1] 마크업의 정의와 요건에 대해서는 2장에서 보충함.
[2] 그 밖의 제작 방식에는 금석문, 판본, 활자본, 신활자본이 있다. (국립국어연구원, 2000, 60)

의할 것이다. 마지막으로 4장에서는 2장과 3장에서 연구한 마크업 방안을 실제 말뭉치 구축 과정에서 어떻게 구현하였는지에 대해 기술하기로 한다.

2. 마크업에 대한 이론적 논의

2.1 마크업의 정의와 요건

마크업(markup)이란, 텍스트 내용 외의 요소에 대한 해석을 명확하게 표현하고 전달하기 위해 표시하는 체계적인 방법을 말하는데, 문헌정보처리연구회 (2002, 16)에서 기술한 바와 같이 이 용어는 역사적으로 식자공이나 타자수에게 특정 단락을 어떻게 구성하고 인쇄할 것인가를 지시하기 위해 텍스트 내에 해제나 기타 표지를 표시하는 데에서 유래하였다.

마크업에 사용되는 일련의 규칙들을 정의한 것을 마크업 언어(markup language)라 한다. SGML(Standard Generalized Markup Langauge)[3]을 반영한 TEI, 그리고 최근 널리 활용되고 있는 XML(eXtensible Markup Language) 등이 대규모 전자 텍스트 자료의 구축을 위한 마크업 언어의 대표적인 예이다.[4] 대규모 말뭉치 구축에서 마크업 언어는 매우 중요하게 연구되어야 하는 부분으로, 영국국가말뭉치(British National Corpus, 이하 BNC) 구축에서도 5개의 실무집단[5]의 하나에 'encoding and markup(인코딩과 마크업)' 파트가 있을 정도로 말뭉치 구축에서 중요하게 다루어지고 있다. 마크업이 된 말뭉치 자료가 언어학적으로 효율적으로 이용되기 위해서는 다음과 같은 기본적인 요건을 갖추어야 한다.

첫째, 마크업이 된 자료는 자료로 구축될 수 있는 다양한 언어 자료의 특성을 빠짐없이 반영할 수 있도록 포괄적이고 자세하게 규정되어 있어야 한다. 일례로 TEI에서는 "모든 언어, 모든 시대, 모든 문학 장르나 모든 유형의 텍스트에 적용된다."는 특성을 명시하고 있다.[6]

둘째, 텍스트의 외형적인 정보도 필요하지만 이보다는 텍스트의 내용적인 특성이나 언어학적 내용을 더 상세히 규정하는 것이 필요하다. 단, 이러한 마크업 언어의 규정을 너무도 엄격히 반영하게 되면, 즉 본문 속에 마크업 태그들을 너무 많이 포함하게

[3] SGML은 텍스트의 제목, 본문, 주석 등의 위치를 나타내기 위해 별도로 주석하는 메타언어로 텍스트의 내용과 구조를 정의하는 데 유용하다. TEI 지침은 SGML을 이용하여 전자 문서의 표준적인 구조를 정한 것인데, 유럽을 비롯하여 한국어 말뭉치 구축에서 적극적으로 도입이 되었다.

[4] 'TEI'(Text Encoding Initiative)란, 1987년에 시작한 국제적 연구 협의체로, 전세계 연구자들의 다양한 전자 문서 교환을 위하여 표준 가이드라인을 규정하여 제공하기 위해 만들어졌다. XML은 SGML의 부분 집합으로 SGML보다 단순한 언어이며, 이기황 (2007, 149-150)에서 기술한 바와 같이 그 특징에 있어서 유연하고 확장성이 좋아 전자 문서뿐 아니라 다양한 종류의 언어 자원을 다루는 언어 공학에서도 이용된다. 2007년에 배포된 TEI-P5에서는 주석 방식에 있어서 XML이 반영되었다.

[5] Burnard (2007, 74)에 의하면 5개의 실무 파트에는 "permissions(허가), design criteria(설계), enrichment and annotation(주석), encoding and markup(인코딩과 마크업), retrieval software(검색 소프트웨어)"가 있다. 첫째 파트인 '허가'란 텍스트를 말뭉치로 개발하기 위해 저작권자로부터 텍스트 사용 허가를 받는 것을 말한다.

[6] 문헌정보처리연구회 (2002, 1) 참조. 그 밖의 TEI에 대한 자세한 설명은 이 책과 강범모 (1998)을 참고할 것.

되면, 특별한 도구 없이 원문을 읽어가는 것이 매우 어려울 수 있다.[7] 따라서 규정된 마크업 언어의 규정을 대체적으로 따르되, 현 실정에 맞게 응용하여 반영하는 것도 필요하다.

셋째, 자료의 특성을 왜곡하여 해석하는 부분이 들어 있지 않아야 한다. 말뭉치는 여러 연구자들이 상호 교환하여 연구하는 것이 가능하며, 어떤 경우는 원문의 매체를 소유하지 않은 상태에서 2차적인 말뭉치 텍스트만을 입수하여 연구해야 할 경우도 있다. 따라서 원문을 임의적으로 판단하여 입력하거나 왜곡된 정보를 반영하게 되는 경우에는 잘못된 언어학적 연구 결과를 산출할 수가 있는 것이다. 실제로 원문의 입수가 쉽지 않은 필사본 자료나 1차 매체가 음성인 경우가 이에 해당하는데, 특히 필사본의 경우 명백한 오기나 누락된 표기 등 자료의 특성으로 인해 자료 자체에 대해 해석하는 부분에 대한 한계를 정하기가 어려운 경우가 있다.

넷째, 향후 자료 이용과 관련하여 언어학적 주석의 확장 가능성을 고려해야 한다. 원시 말뭉치로 구축된 자료는 형태 주석, 의미 주석 등의 언어학적 주석이 부가될 수 있으며, 통사 주석과 같이 단어 단위 이상의 구 단위나 문장 단위 주석을 필요로 하는, 확장의 가능성이 존재하기 마련이다.

이상에서 살펴본 마크업 요건들은 본 연구의 주요 대상인 한글 필사본 음식조리서의 특성을 고려할 때, 실제 마크업 과정에서 논의의 대상이 되었던 부분이다. 역사 자료이며 필사본이라는 텍스트 정보는 첫째, 셋째의 요건과 관련되며, '음식조리서'라는 전문 영역 텍스트의 특성은 둘째의 요건과 관련된다. 또한 본 연구가 원시 말뭉치, 주석 말뭉치 구축 및 이를 통한 통합 검색 시스템 개발이라는 순차적인 연장선상에서 진행되므로 이러한 연구의 진행 과정은 넷째 요건인 주석의 확장 가능성과 밀접한 관련을 가진다. 본 연구에서 가장 중점을 둔 부분은 주석의 확장 가능성인데, 이는 지금까지 세종계획을 비롯한 한국어 말뭉치 구축 부분에서 적극적으로 고려되지 못했던 부분이다. 이에 대해서는 4장에서 상세히 논의될 것이다.

2.2 말뭉치 마크업의 쟁점

여기서는 2.1의 마크업의 요건을 중심으로 기존 말뭉치 마크업의 특성을 살펴봄으로써 말뭉치 마크업의 쟁점과 해결 방안에 대해 논의하고자 한다.

우선, 국내 말뭉치 마크업의 대표적인 사례로, SGML을 도입한 국내 최초의 말뭉치인 연세말뭉치와 TEI-P3를 확장·변형하여 활용한 세종계획 말뭉치를 들 수 있다. 전자의 경우, SGML의 요건에 맞춘 몇 개의 태그를 한글 자소로 만들어 이용하였는데, 예를 들어 '제목'은 한글 자소로 "<ㅈㅁ>"과 같이 표기하는 방법을 택하였다.[8] 하지만 이는 단순한 문서 정보만을 수록하는 한계를 가진 점, 한글 자소를 사용함으로써

[7] 이를 위해 마크업 기호는 보이지 않도록 하고 이 기호가 의미하는 바를 시각적으로 표현한 뷰어(viewer)를 개발하기도 한다.

[8] 이에 대한 자세한 설명은 안의정 (1999, 10) 참고.

국제적인 규약인 TEI와의 통일성이 없다는 점 등의 문제로 오래 사용되지는 못했다.

한편, 세종계획이 따르고 있는 당시 TEI 체계는 1994년에 TEI에서 제정하여 발표한 것으로, 모든 문서 유형을 포괄하고 있기는 하지만 언어학적 주석에 대한 부분은 소략하게 기술되어 있고, 주석 말뭉치의 마크업에 대해서도 규정하고 있지 않다는 문제점이 있었다. 따라서 이를 따르고 있는 세종계획 말뭉치의 경우도 원시 말뭉치는 TEI 체계를 따랐지만, 형태 주석 말뭉치의 경우 어절 번호가 붙은 원시적인 수직형으로 되어 있어서, 형태 주석 이상의 의미 정보나 어절 이상의 단위에 대한 주석 정보 확장이 용이하지 않다는 한계가 있었다. 세종계획 이후의 여러 논의에서 TEI 체계는 주석 말뭉치 구축과 말뭉치 주석 정보의 확장성에 있어 취약하다는 문제점이 지적되어 왔다. 대표적으로 이기황 (2007)에서는 "(세종 말뭉치는) 현재의 주석 형식에 맞는 응용 소프트웨어를 사용하여야만 하며, 주석의 확장에 대처하기가 힘들다는 단점이 있다." 고 지적하면서, 말뭉치 활용을 극대화하기 위해서 구조화된 문서에 대한 일반적이고 확장성이 보장되며 개방된 마크업을 제공하는 XML을 채용하는 것이 바람직하다고 하였다.

위의 기존 국내 마크업 사례의 문제점을 종합하면, 지금까지 국내 마크업은 국제적 통일성의 준수, 주석 정보 확장 가능성, 응용 소프트웨어 활용의 제한성 등의 문제를 가지고 있음을 알 수 있다. 한편 이에 더하여, 기존 마크업의 문제는 원시 말뭉치와 병렬 말뭉치, 또는 원시 말뭉치와 주석 말뭉치가 별도의 파일로 존재한다는 점이다. 따라서 대표적인 예인 세종계획 말뭉치의 경우, 향후에 파일을 수정하거나 변형할 경우 원시 말뭉치와 주석 말뭉치를 각각 수정해야 하는 등 파일 관리에 어려움이 많았다. 별도의 정보이지만 유기적인 관계에 있는 형태 주석 정보, 의미 주석 정보의 효율적인 관리를 위해서는, 이들에 대한 각각의 주석 말뭉치가 하나의 파일 내에 존재해야 할 필요가 있다.

본 연구에서는 XML 체계를 도입함으로써 위의 문제를 해결하고자 한다. 즉, 원문과 현대어역, 그리고 언어학적 주석 정보로 구성된 말뭉치가 정보의 확장, 각종 프로그램에 유연하고, 이들이 한 파일 내에서 존재하여 파일 관리와 활용에 용이하도록 만들기 위해서는 XML 체계를 도입할 필요가 있는데, 그 상세한 구조에 대해서는 4장에서 깊이 논의할 것이다.

한편, 처음부터 XML 체계를 도입한 것은 아니지만, 최근 이미 구축된 모든 문서를 BNC-XML이라 하여 XML 마크업으로 전환한 BNC의 사례는 본 연구에서 참조할 만하다. BNC-XML의 구어 텍스트 예를 보이면 [그림 1]과 같다.[9]

Burnard (2007, 99)에서는 [그림 1]의 BNC-XML을 위한 언어학적 주석이 세 가지 측면에서 개선되었다고 하였는데[10], 그 중 추가 요소인 <mw> 태그를 이용하여

[9] [그림 1]의 자료는 "http://www.natcorp.ox.ac.uk/XMLedition/URG/cdifsp.html#cdif22" 에서 옮겨왔다.

[10] 세 가지 개선 사항은 다음과 같다.

```
<s n="5490">
 <event desc="radio on"/>
 <pause dur="34"/>
 <w c5="PNP" hw="you" pos="PRON">You </w>
 <w c5="VVN" hw="get" pos="VERB">got</w>
 <w c5="TO0" hw="ta" pos="PREP">ta </w>
 <unclear/>
 <w c5="NN1" hw="radio" pos="SUBST">Radio </w>
 <w c5="CRD" hw="two" pos="ADJ">Two </w>
 <w c5="PRP" hw="with" pos="PREP">with </w>
 <w c5="DT0" hw="that" pos="ADJ">that</w>
 <c c5="PUN">.</c>
</s>
<s n="5491">
 <pause dur="6"/>
 <w c5="AJ0" hw="bloody" pos="ADJ">Bloody </w>
 <w c5="NN1" hw="pirate" pos="SUBST">pirate </w>
 <w c5="NN1" hw="station" pos="SUBST">station </w>
 <w c5="VM0" hw="would" pos="VERB">would</w>
 <w c5="XX0" hw="not" pos="ADV">n't </w>
 <w c5="PNP" hw="you" pos="PRON">you</w>
 <c c5="PUN">?</c>
</s>
```

[그림 1] BNC-XML 말뭉치의 예

다수어(multiword)를 명시적으로 마크업한다는 부분이 눈여겨 볼 만하다. <mw> 태그는 TEI 체계에서 취약한 관용 표현, 구 구성 등의 다중 어휘 단위(multi-lexical unit)에 대한 주석 체계에 유용하게 활용될 수 있다. 특히 본 연구에서는 형태 정보 주석뿐 아니라 조리 용어와 조리법 관련 표현을 대상으로 하여 의미 정보 주석 말뭉치도 개발하고자 하는데, 음식조리서 말뭉치에는 "개곰낫만큼(개암알 크기만큼씩), 주머니에 닛 내ᄃᆞᆺ 쳐(주머니에 잇꽃 물을 내듯 쳐)" 등과 같이 조리법과 관련된 표현에 구나 절 단위가 많으므로 이러한 요소를 참고할 만하다.

TEI-P5에서는 [그림 2]에서 보이는 바와 같이[11] 이전 버전과 다르게 XML 형식의 주석에 대해 규정되어 있다.

본 연구에서 목적으로 하고 있는 한글 필사본 말뭉치는 원시 말뭉치뿐 아니라 원

① 다수어(multiword)와 그 구성 요소는 <mw>, <w> 로써 명시적으로 태깅한다.

② C5 태그세트를 훨씬 단순화시킨 버전을 사용하여 추가 어휘부류 스키마를 개발하였다.

③ 수동으로 규정된 규칙을 기반으로 단어의 레마화가 자동 수행된다.

[11] 이 자료는 "http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf"의 528쪽에서 옮겨 왔다.

```
<s>
  <w ana="#AT0">The </w>
  <w ana="#NN1">victim</w>
  <w ana="#POS">'s</w>
  <w ana="#NN2">friends </w>
  <w ana="#VVD">told </w>
  <w ana="#NN2">police </w>
  <w ana="#CJT">that </w>
  <w ana="#NP0">Kruger </w>
  <w ana="#VVD">drove </w>
  <w ana="#PRP">into </w>
  <w ana="#AT0">the </w>
  <w ana="#NN1">quarry </w>
  <w ana="#CJC">and </w>
  <w ana="#AV0">never </w>
  <w ana="#VVD">surfaced</w>
</s>
```

[그림 2] TEI-P5의 형태소 주석 말뭉치의 예

시 말뭉치와 일대일의 문장 대응 관계를 가지도록 구축된 현대어역 말뭉치, 형태 주석 말뭉치와 의미 주석 말뭉치로 구성되므로, 말뭉치의 마크업은 유연성과 확장성이 보강된 XML을 기반으로 할 필요가 있다. 또한 본 연구의 궁극적인 목적이 검색 시스템 개발에 있느니만큼 검색을 위한 관용어, 연어 등의 다중 어휘 단위까지도 주석하기 위해서는, XML 체계를 활용하되 이를 연구 목적에 맞게 변형·재구성해야 할 필요가 있다. 이를 위하여 본 연구에서는 2007년에 배포된 TEI-P5를 확장·변형하고, 한글 필사본에 필요한 요소를 취하여 마크업에 이용하고자 한다.

3. 한글 필사본 음식조리서의 텍스트적 특성과 마크업 방안

3장에서는 한글 필사본 음식조리서의 텍스트 특성에 따른 마크업 방안에 대해 상세히 논의할 것이다. 각 절에서는 역사자료로서의 특성과 필사본으로서의 특성, 그 중에서도 개별 음식명을 중심으로 구성된 음식조리서의 특수한 텍스트 구조가 마크업 체계를 구성할 때 고려해야 할 중요한 변수임을 논의하게 될 것이다.

3.1 역사 자료로서의 특성

본 연구의 주요 대상인 한글 필사본 자료는 역사 자료로서 다음과 같은 텍스트 특성을 가지고 있다.[12]

[12] 홍윤표 (2006, 139)에서는 국어사 자료의 말뭉치 구축과 관련하여 형식상의 특성을 정리한 바가 있는데, 이 중 한글 필사본 말뭉치에도 해당되는 내용을 다시 정리하면 다음과 같다.

우선, 역사 자료 말뭉치 구축에서 가장 어려운 문제 중 하나로, 한글 필사본 음식조리서 역시 띄어쓰기가 되어 있지 않다. 따라서 입력 과정에서 작업자들의 기준을 통일할 띄어쓰기 방안을 마련해야 하는 번거로움이 있을 뿐만 아니라, 입력 단계 이후 주석 단계에서 띄어쓰기를 수정할 경우 원시 말뭉치와 주석 말뭉치 모두에 수정 부분이 반영되어야 한다는 제한이 있다. 본 연구에서는 입력 단계에서 기준 사전을 정하여 띄어쓰기의 통일성을 최대한 고려하였으나, 주석 말뭉치 개발 과정에서 자유롭게 수정이 가능하도록 하기 위하여, 원시 말뭉치와 주석 말뭉치가 한 파일 내에 존재하도록 설계하였다. 이러한 방안은 세종계획 말뭉치와 같이 원시 말뭉치와 주석 말뭉치가 별개의 파일로 구분되어 있어서 띄어쓰기의 수정이 두 파일 각각에서 이루어져야 하는 번거로움을 피하고자 한 것이다.

둘째, 한글 필사본도 다른 국어사 자료와 마찬가지로 원문은 문장부호가 사용되는 경우가 거의 없고, 문단 하나가 하나의 문장으로 되어 있을 만큼 문장이 길다.[13] 이러한 특징은 용례 검색기로 원문을 검색할 때 검색된 용례의 가독성을 떨어드리며, 문장 단위 검색이 용이하지 않다는 문제가 있다. 따라서 본 연구에서는 본문 내에 문장 수준의 임의적인 최소 단위를 설정하여 원시 말뭉치와 현대어역 말뭉치의 문장 단위 구분을 통일하여 주석하였다. 이러한 결정은 최소 단위를 행이나 구로 하여 더 작은 언어학적 단위로 설정하는 것이 조리법 과정이나 전체 텍스트를 해독하는 데 문장 단위보다 유용하지 않다는 판단에 따른 것이다.

마지막으로, 본고에서 대상으로 하는 자료는 17~19세기 음식조리서이므로 근대국어의 문법적 특징을 비롯한 이 시기의 국어의 전반적인 특징이 말뭉치 구축 시에 고려되어야 한다는 점이다. 국어사 연구에서 살펴볼 때 근대국어는 문체면에서 종결어미가 매우 적으며, 현대국어에서는 나타나지 않는 다양한 유형의 이형태가 존재한

ㄱ. 입력자의 기준에 따라 띄어쓰기를 하였기 때문에 일정한 규칙에 따라 입력되어 있지 않다.

ㄴ. 대부분 원시 말뭉치의 상태일 뿐, 주석 말뭉치가 많지 않다.

ㄷ. 문헌자료에 대한 헤더(header)가 충실하지 않다. 단지 입력자나 입력기의 이름, 입력일자, 입력 문헌명과 간행연도 등만이 간략히 기재되어 있다.

ㄹ. 문헌별로 입력된 자료의 파일 이름이 입력자마다 달라서 혼돈을 일으키고 있다.

ㅁ. 입력 자료에 문장 단위로 끊어지는 곳의 표지가 없어서 용례사전을 만드는 프로그램 등의 사용에 어려움이 많다.

[13] 텍스트에 따라 매문단의 시작이나 문장의 끝에 ○라는 특수기호가 있기도 한데, 본 연구에서는 세종계획 역사자료 말뭉치의 지침을 참고하여 (국립국어연구원, 2001, 367), 보통 다음과 같은 종결어미가 나타나는 곳을 문장 경계로 하였다.

i) 다, 니라

ii) 녀, ㄴ다, ㄴ가, 뇨

iii) 라

iv) 져

위의 어미들을 기준으로 했을 때 한글 필사본 음식조리서의 경우 심하면 하나의 조리법이 하나의 문장으로 기술되곤 한다.

다는 특징이 있다. 종결어미가 적다는 것은 문장이 전체적으로 길다는 것으로 앞서 기술한 두 번째 특성과도 관련이 있다. 그리고 다양한 유형의 이형태가 존재한다는 것은 본 연구의 자료가 한글 필사본이라는 특성과 17~19세기의 넓은 시대적 분포를 가진다는 점에 기인하는 특성이다. 이 중에서, 이형태가 많다는 것은 형태소 주석 작업 시에 분석이 까다롭고, 용례 검색기 개발에서 문제가 될 수 있는 부분으로 이에 대한 해결책이 필요하다. 본 연구에서는 문장 종결 지점의 문제와 이 형태의 주석 문제를 효과적으로 해결하기 위해, 문장 종결 지점에 대한 지침을 마련하고, 형태소 분석 작업 시 이형태들을 연결시키는 작업과 그 결과를 말뭉치 안에 포함시키는 마크업 방안을 연구하였다.[14]

3.2 한글 필사본 자료로서의 특성

3.1에서 다룬 역사 자료로서의 특징과 더불어 역사 자료 중, '필사본'으로서의 특성 역시 마크업 체계를 개발하는 데 중요한 변수로 작용한다. 아래 그림은 본 연구의 주요 대상인 26종의 음식조리서 중 하나인『주식시의(酒食是儀)』의 일부로, 필사본으로서의 특성을 드러내는 자료이다.

[그림 3] 음식조리서『주식시의(酒食是儀)』의 일부

[14] 이 부분에 대한 자세한 설명은 "4.2 주석 말뭉치의 마크업"을 참고.

[그림 3]에서 알 수 있듯이, 한글 필사본 자료는 간본이나 활자본보다 원문을 판독하기가 어렵다. 따라서 한글 초서체를 읽어본 경험이 있는 입력자가 국어학적 지식을 바탕으로 하여 원문을 판독해야 하는데, 이 과정에서 판독에 자신이 없는 부분이나 판독을 하여 입력은 했으나 불확실하게 판독한 부분이 생기고 여기에 주석을 달아놓기도 한다. 따라서 이와 같이 오기로 판단되는 부분이나 자신의 판독이 부정확하다고 판단되는 부분, 원문이 손상된 부분 등에 대한 정보가 마크업 과정에서 드러나야 할 필요가 있다.[15]

판독의 정확성 문제와 더불어, 필사본의 특성으로 들 수 있는 것은 필사본의 텍스트는 판본 등과 달리 필사자가 하나의 텍스트 내에 시기가 다르거나 내용이 다른 정보를 따로 부가하게 됨으로써 나타나는 특징이다. 즉, 본래의 텍스트 외에 1차로 필사한 후에 저자가 끼워 넣은 부분도 있는데, 예로 종이의 여백에 쓰인 글들은 본문과 다른 새로운 내용이 추가된 것이 있는가 하면, 색인이나 글씨 연습처럼 본문의 내용과 겹치는 것도 있다. 전자는 반드시 입력되어야 하는 부분이라면, 후자는 생략하여도 전체 텍스트를 이해하는 데에 문제가 없는 부분이다. 이 때 전자에도 여백에 있었다는 의미로 태그를 부착하여 마크업해야 할지, 아니면 특정한 태그 없이 해당 위치에 끼워 넣어 입력만 해야 할지 결정해야 한다. 본 연구에서는 별도의 태그를 마련하여 이러한 필사본 자료의 텍스트적 특성을 체계적으로 나타내고자 하였다.

3.3 음식조리서로서의 특성

[그림 3]에서 보는 것처럼, 음식조리서는 텍스트의 구조상 일반 텍스트처럼 장, 절, 문단 등의 구분이 되어 있는 것이 아니라, 대부분 '연계찜, 개장/ 약식법, 강정법/ 해삼 달이는 법, 인절미 굽는 법' 등과 같은 음식명 또는 '음식+법' 등을 소제목으로 하여 그에 따라 조리 방법이 잇따라오는 구조로 되어 있다. 텍스트에 따라서 음식명 앞에 "면병류, 어육류, 주국방문, 초 담는 법"과 같은 음식 부류명이 오기도 하는데, 이는 텍스트마다 차이가 있다. 즉, 음식조리서는 마치 사전 텍스트가 표제어(headword)와 어휘 내항(entry)으로 구성되어 있듯이, 음식명과 조리 방법이 번갈아 기술되는 구조를 가지고 있으며, 음식명은 반드시 나오지만 음식 부류명은 생략되기도 한다는 것이 텍스트 구조상의 가장 큰 특징이다.

따라서 본 연구에서는 마크업 체계에서 본문의 구분을 문단이나 음식 부류명에 의해서가 아니라, 음식명이나 조리법(recipe)의 소제목 즉 `<foodItem>`의 태그를 부착하여 이를 중심으로 구분하였다. 음식명을 태그로 표시하고 조리법을 본문의 기본

[15] 실제로 이러한 특징은 현대국어 자료 중에서도 음성 데이터를 전사하여 구축하는 구어 말뭉치에서 나타난다. 음성 데이터의 녹음 상태가 좋지 않은 부분에서는 잘 들리지 않기 때문에 내용의 흐름을 고려하여 추정하여 전사를 하게 되고, 전혀 들리지 않는 부분은 공백으로 둘 수밖에 없게 된다. 말뭉치 마크업에서는 이러한 곳에 부착하는 태그를 정해 놓고 있다. TEI에서는 구어 텍스트의 경우 잘 들리지 않아 불확실한 부분은 `<unclear>` 태그로 감싸고, 단어의 흔적은 있으나 전혀 들리지 않아 공백이 생기는 부분에는 `<omit>` 태그를 붙인다.

단위로 하게 되면, 추후 용례 검색기를 개발할 때 어휘 단위뿐 아니라 관심이 있는 음식명을 기준으로 하여 검색하는 데 용이하기 때문이다. 한편 음식조리서의 또 다른 특징은 대부분 자녀나 며느리와 같은 후대 사람들에게 요리하는 법을 들려주는 듯한 구어체로 되어 있다는 것이다. 구어체는 예기치 못한 이형태가 많다는 특징이 있다. 이는 형태 주석 말뭉치 개발에서 중요한 요소로 고려되어야 할 사항이다.

4. 말뭉치 유형에 따른 마크업의 실제

4장에서는 3장에서 정리한 자료의 특성을 반영하여, 음식조리서 말뭉치의 마크업 과정을 원시 말뭉치와 주석 말뭉치로 나누어 설명하고자 한다. 본 연구에서 원시 말뭉치와 주석 말뭉치는 하나의 파일에 존재하게 되는데, 이는 파일 관리의 효율성을 위해서이다. 또, 3.1에서 논의한 바와 같이, 주석 말뭉치를 구축하는 과정에서 잘못된 원문 판독 내용을 발견하여 수정하고 싶을 때, 원시 말뭉치로의 접근성을 높이기 위함이다.

4.1 원시 말뭉치의 마크업

원시 말뭉치의 마크업은 전체 구조와 헤더, 본문에 사용된 태그로 나누어 설명될 수 있다.

4.1.1 전체 구조. XML은 다음 (1)과 같은 기본 형식을 가지고 있다.

(1) `<name att1="value" att2="value">`[16]

여기서 'name'은 태그의 이름을 말하며, 태그들은 'att1, att2' 등과 같은 속성(attribute)을 지니고, 각각의 속성은" "에 표현한 바와 같은 속성값(value)을 갖게 된다.

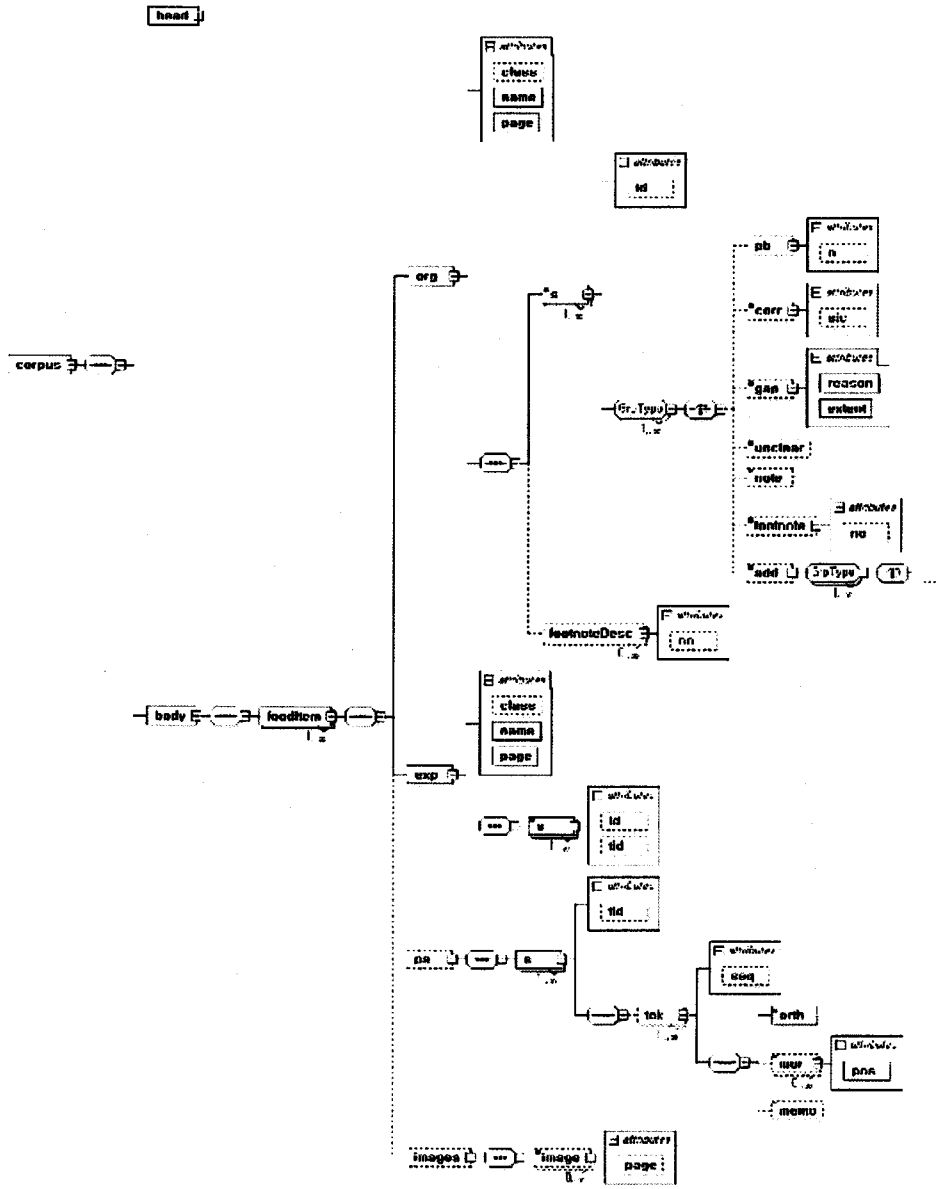
본 말뭉치에 사용된 태그들과 전체 구조를 도식화하면 [그림 4]와 같다.

[그림 4]에서 태그는 실선이나 점선으로 표현되어 있는데, 실선은 반드시 존재하여야 하는 필수 태그를 의미하고, 점선은 생략이 되어도 마크업 문법에 문제를 일으키지 않는 수의적인 태그를 의미한다.[17] 마찬가지로 속성들도 실선으로 표시된 것은 해당 태그에 반드시 있어야 하는 속성을 의미하고, 점선은 생략이 되어도 무방한 속성을 의미한다.

본문에 해당하는 <body> 태그 하위에는 <foodItem> 태그들이 위치한다. 보통의 말뭉치에서는 <p> 태그를 기본적인 구조 태그로 삼지만, 우리 자료에서는 대상 문헌이

[16] XML의 표준 규격에 대한 설명은 "http://www.w3.org/XML/" 을 참고할 것.

[17] 본 연구에서는 수행된 마크업 작업에 오류가 있는지를 검사하는 검사기를 만들어 최종 단계에서 사용하였다(프로그램 이름: FoodCorpusValidator). 이 검사기는 몇몇 태그를 입력하는 데에도 사용되었는데, 즉, 헤더에서 작업 수정 기록을 나타내는 태그나, <s> 태그의 일련번호 부착 등은 이 검사기를 이용하여 오류도 잡아내고 태깅 작업도 실시하였다. 이렇게 마크업 태깅 작업에서 도구를 사용하게 되면 일일이 입력할 때 생기는 오류를 방지할 수 있고, 수작업으로 찾아내기 어려운 오류를 발견하는 데 도움을 줄 수 있다.



[그림 4] 말뭉치의 전체 구조

문단에 맞춰 적혀 있지 않고 <p> 태그의 활용도도 높지 않기 때문에 사용하지 않았다. 대신 요리명 태그인 <foodItem> 태그를 기본 구조 태그로 설정하기로 하였다. 이 태그는 요리를 검색하는 인덱스 역할을 한다. 다시 말해 요리명만 검색하면 그에 해당하는 제목과 내용을 바로 찾을 수 있고 일반인을 위한 용례 검색기를 개발한다면

첫 화면에서 이 요리명들로 메뉴를 꾸며 찾을 수 있게 할 것이다.

<foodItem> 태그 하위에는 아래 그림과 같이 원전 텍스트를 부분인 <org> 태그와 이의 현대어역인 <exp> 태그가 위치하게 된다.

```
<foodItem>
 <org class="면병류" name="면" page="1a">
  <s id="0">
   <pb n="1a" />음식디미방</s>
  <s id="1">면병뉴</s>
  <s id="2">면</s>
  <s id="3">것모밀를 씨어 하ᄆᆞ이 ᄆᆞᆯ뇌디 말고 알마초 ᄆᆞᆯ뢰여</s>
  <s id="4">ᄠᆞᆯ을 조히 아아 디흘 제 미리 물 품겨 축축이 ᄒᆞ야 듯다가</s>
  <s id="5">디흘 제 녹도 거피ᄒᆞᆫ ᄠᆞᆯ 조히 시어 건져</s>
  <s id="6">믈 ᄲᅴ거든 모밀ᄠᆞᆯ 닷되예 물 부ᄅᆞᆫ 녹두 ᄒᆞᆫ 복ᄌᆞ식 섯거 지ᄒᆞᄃᆡ
   방하를 ᄀᆞ만ᄀᆞ만 지허 것ᄀᆞᆯ를 처 ᄇᆞ리고 키로 퍼 ᄇᆞ리고 키 그테 흰 ᄠᆞᆯ이
   나거든 그ᄅᆞᆯ 뫼화 다치 ᄒᆞ면 그 ᄀᆞᆯ리 ᄀᆞ장 희거든</s>
  <s id="7">면 ᄆᆞᆯ 제 더운물에 녹게 ᄆᆞ라 누ᄅᆞ면 비치 희고 조ᄒᆞᆫ 면이
   되ᄂᆞ니라</s>
  <s id="8">교틱ᄂᆞᆫ 싀면 교틱 ᄀᆞ치 ᄒᆞ라</s>
 </org>
 <exp class="면병류" name="면" page="1a">
  <s tid="0">음식디미방</s>
  <s tid="1">면병류</s>
  <s tid="2">면(麵) (메밀국수 만드는 법)</s>
  <s tid="3">겉메밀을 씻어 너무 많이 말리지 말고 알맞게 말려라</s>
  <s tid="4">메밀쌀의 잡것을 가려내고 깨끗이 해서 찧기 전에 미리 물을
   뿜어 축축이 해 두어라</s>
  <s tid="5">찧을 즈음에 녹두 알갱이의 껍질을 벗겨 깨끗이 씻은 다음 건져
   내라</s>
  <s tid="6">물이 빠지거든 메밀쌀 다섯 되에 물에 불린 녹두를 한 복자씩
   섞어 찧되, 방아를 살살 찧은 후 겉가루를 체로 치고 키로 까불어 키 끝에
   흰 부스러기가 나오거든 그것을 모아 나누면 그 가루가 아주 희다</s>
  <s tid="7">(그 가루로) 면을 반죽할 때에 더운 물에 녹게 반죽하여 누르면
   빛이 희고 깨끗한 면이 되느니라</s>
  <s tid="8">고명은 세면(실국수)의 고명과 같게 하라</s>
 </exp>
</foodItem>
```

[그림 5] 본문 마크업의 예(음식디미방)

이 말뭉치는 역사 자료 말뭉치이기 때문에 원문의 내용을 전공자가 아니라면 쉽게 이해할 수 없다. 따라서 원문의 내용 다음에는 현대어역 부분이 뒤따라오도록 구성하였다. <org>와 <exp>의 하위에는 <s> 태그가 각각 오는데, 이 태그는 다음의 (2)와 같이 현대어역을 참고하여 임의로 구분한 문장 수준의 단위에 붙였다.

(2) ㄱ. (주해문 입력) 겉메밀을 씻어 너무 많이 말리지 말고 알맞게 말려라. 메밀쌀의 잡것을 가려내고 깨끗이 해서 찧기 전에 미리 물을 뿜어 축축이 해 두어라. 찧을 즈음에 녹두 알갱이의 껍질을 벗겨 깨끗이 씻은 다음 건져내라. 물이 …

ㄴ. (원문 마크업)

<s id="3">것모밀ᄅᆞᆯ 씨어 하ᄆᆞ이 ᄆᆞᆯ뇌디 말고 알마초 ᄆᆞᆯ뢰여</s>
<s id="4">ᄲᆞᆯ을 조히 아아 디ᄒᆞᆯ 제 미리 물 품겨 축축이 ᄒᆞ야 듯다가</s>
<s id="5">디ᄒᆞᆯ 제 녹도 거피ᄒᆞᆫ ᄲᆞᆯ 조히 시어 건져</s>
<s id="6">믈 ᄲᅴ거든 … </s>

(음식디미방)

<s> 태그는 3.1에서 설명한 바와 같이, 한글 필사본 자료가 문장 구분이 없기 때문에 가독성과 용례 검색기 개발을 고려하여 도입한 인위적인 구조 태그라 할 수 있다. <s> 태그의 경계는 주해문의 해석에 따라 정해지기 때문에 (2ㄱ)처럼 주해문을 입력한 후에 이에 맞춰 원문인 (2ㄴ)에 붙이게 되며, 내용의 연결을 위해 일련번호를 붙여 정렬될 수 있도록 하였다.

본문은 [그림 3]에서 보인 바와 같이 음식명이 오고 뒤이어 조리법의 설명이 오는데, 음식명에 제목 태그를 별도로 달지 않았다. 그 대신에 <org>와 <exp> 태그의 'name' 속성 안에 음식명을 넣어서 활용할 예정이다. 필사본의 성격상 텍스트에 제목이 누락된 경우도 있었는데, 전체 구조에 어긋나지 않고 요리명이 중요한 인덱스 역할을 하므로 모두 살려서 넣었다.

<org>와 <exp> 태그에는 "name" 외에도 "class"와 "page"라는 속성이 있다. "class"는 "면병류, 어육류, 식초 담는 법" 등과 같이 음식명의 중분류에 해당하는데, 여러 가지 음식명을 단순하게 나열하는 것보다는 비슷한 유형끼리 묶어 그룹을 지을 때 사용하기 위해 마련한 것이다. 또, "page" 속성은 쪽 정보인데 해당 조리법이 문헌의 어떤 쪽에 걸쳐 나타나는가를 표시하는 것이다. 이 정보는 본문에 수작업으로 태깅된 쪽 정보인 <pb> 태그를 이용하여 자동으로 붙을 수 있도록 설계되었다. 쪽 정보는 몇 번째 쪽, 앞면(a), 뒷면(b)과 같이 표시하는 것이 일반적이다. 이 정보는 용례 검색기에서 출전을 표시할 때 필요하며, 국어사 자료에서는 필수적이다. 쪽 정보를 태깅할 때 만약 어절 내에 쪽 경계가 오는 경우 해당 어절을 다 입력한 후에 다음 쪽 표시를 해야 한다. 이렇게 해야만 쪽 태그로 인해 형태소나 어절이 분리되는 일이 없게 되고, 용례를 검색할 때 영향을 주지 않게 된다.

<exp> 아래에는 <pa>라는 태그 영역이 있는데, 이는 주석 말뭉치와 관련된 부분이므로 4.2에서 설명하기로 한다.

마지막에 위치하는 <images> 태그는 이미지 파일과의 연결을 위한 것으로, 이렇게 이미지 파일에 대한 정보를 말뭉치 파일 속에 넣는 이유는 원자료에 대해 확인하고 싶은 사용자를 위한 장치이다.

4.1.2 헤더. 헤더는 추후 말뭉치를 사용하거나 인용하는 경우 반드시 필요한 정보로서, 원전 텍스트에 대한 정보뿐 아니라, 말뭉치에 사용된 부호화, 입력·주석 과정, 작업자에 관한 자세한 기록을 말한다. 한글 필사본의 경우에는 헤더에 필사 연대, 필사자, 필사기, 서체, 기타 필사상의 특이점 등이 기록되어야 한다. 본 연구에서 정리한 <head>의 전체적인 모습은 [그림 6]과 같다.

헤더에는 보통 원전 텍스트의 서지 정보에 대한 자세한 설명과 파일 수정의 역사를 기록하게 된다. 또, 텍스트나 말뭉치의 분량, 텍스트별 특이 사항을 기록하기도 한다. 한글 필사본의 경우 동일 텍스트에 대해 다양한 문헌이 존재하는 경우가 많지 않고, 이들이 말뭉치로 모두 개발된 사례 또한 많지 않기 때문에 문헌간의 혼돈을 야기할 정도는 되지 못한다. 그러나 본 연구처럼 특정 분야로 한정하여 말뭉치를 구축하는 경우에는 '규곤요람'이나 '규합총서', '주식방문'처럼 동일한 내용의 이본이 입력될 수도 있다.[19] 이 경우 헤더 부분에서 판본의 다름을 분명히 해야 할 필요가 있다.

3.2에서 정리한 바와 같이 한글 필사본 자료는 형식이 규정되어 있지 않아서 본 텍스트 이외의 지면에도 필사가 되어 있는 곳이 있다. 이를 위해서 적절한 태그(<front>, <back>)를 부착하여 본 텍스트(<body>)와 구분하기도 하는데, 본 말뭉치에서는 [그림 6]처럼 <bibl> 태그의 하위에 <desc> 태그를 두어 이러한 내용들을 자유롭게 기술하였다.

4.1.3 본문. 본문에 사용된 태그는 대부분 원문 판독에 대한 것으로 다음과 같은 것들이 있다.

(3) ㄱ. 술이며 ᄇᆡ 얼면
⇒ 술이며 <corr sic="ᄇᆡ">ᄆᆡ</corr> 얼면

ㄴ. 고초ᄀᆞ{'ᄅᆞ' 누락}
⇒ <corr sic="고초ᄀᆞ">고초ᄀᆞᄅᆞ</corr>

ㄷ. <corr sic="">(복원 어절)</corr>

(4) <footnote> ····· </footnote>

[18] <head> 전에, 즉 말뭉치의 가장 앞부분에 위치하는 이 부분은 본 말뭉치의 마크업에 사용된 XML의 버전과 입력에 사용된 문자 인코딩 방식에 대한 정보를 나타낸다.

[19] 말뭉치 구축을 위해 입력된 이본들은 다음과 같은 것들이 있다. 이들을 <head>의 <title>에 다음과 같이 입력하였다.
규합총서(정양완본)/규합총서(동경대본)/규합총서(영남대본), 주식방문(개인소장)/주식방문(노가재공댁), 주방문(하생원댁)/주방문(김승지댁), ···

```
<?xml version="1.0" encoding="UTF-8"?>[18]
 <head>
  <bibl>
   <title> 규곤요람(연세대본)</title>
   <author>Unknown</author>
   <transcripter>Unknown</transcripter>
   <authorDate>1894</authorDate>
   <pageCount>21</pageCount>
    <desc> 본문 시작 전 표지에 다음과 같은 사항이 기록되어 있음 - 병신 오월/
건양원연 오월 초륙일/광셔병신 졍월 일/규곤요람 단 권/병신 오월 열일 릴이라/
건양원연 쵸하의 등출리라/飮食錄/음식록</desc>
  </bibl>
  <revision>
   <constructor> 경북대학교</constructor>
   <change>
    <date>2007-09</date>
    <worker> 송현주</worker>
    <workMemo>1 차 입력</workMemo>
   </change>
   <change>
    <date>2008-08</date>
    <worker> 이미향</worker>
    <workMemo>1 차 교정, 2 차 교정</workMemo>
   </change>
   <change>
    <date>2008-10</date>
    <worker> 백두현</worker>
    <workMemo>3 차 교정</workMemo>
   </change>
   <change>
    <date>2008-10</date>
    <worker> 안의정</worker>
    <workMemo>1 차, 2 차 마크업</workMemo>
   </change>
   <change>
    <date>2008-10</date>
    <worker> 김정아</worker>
    <workMemo>1 차 형태소 분석</workMemo>
   </change>
  </revision>
 </head>
```

[그림 6] 헤더 마크업의 예(규곤요람)

(5) <gap reason="판독불가" extent="약 2줄반 25글자" />[20]

(6) <add> ····· </add>

(3ㄱ)은 명백한 오기를 수정하였을 때 붙이는 태그로 한글 필사본에서는 오기가 자주 보인다. 이러한 오기를 수정하지 않고 오기 그대로 입력하게 되면 자료로 이용할 수 없기 때문에 수정하여 입력하여야 한다. 그러나 원문과 달라졌기 때문에 수정되었다는 기록도 <corr> 태그를 이용하여 남기는 것이다. 이 태그는 (3ㄴ), (3ㄷ)처럼 탈자나 어절 전체가 생략되었을 경우에도 사용할 수 있다.

(4)는 각주를 말하는데 판독자가 판독 과정에서 덧붙이고 싶은 내용을 남기는 데 사용되는 태그이다. 본문에는 이 태그만을 붙이고 각주의 내용은 [그림 4]에서 보듯이 <org>의 <s> 태그 다음에 위치하는 <footnoteDesc>에 입력한다. 각주 태그의 내용들은 이미지 파일을 살필 때 필요한 정보가 된다.[21]

(5)은 자료를 판독할 수 없는 경우에 붙이는 태그로 한글 필사본의 경우 원전의 상태가 좋지 않은 경우는 이 태그를 붙인다. "extent"의 속성값은 만약 원문 텍스트에서 실제 길이를 알 수 있다면 센티미터(cm)로 표현할 수도 있다.

(6)은 3.2에서 설명한 한글 필사본 자료의 특징인, 여백에 추가된 텍스트를 표시하기 위한 태그이다. TEI에서는 추가된 내용이 문자, 단어, 구 정도라면 <add> 태그를, 문장 이상의 비교적 긴 내용이 추가된 경우라면 <addSpan> 태그를 붙이도록 하여 구별하지만, 본 연구에서는 이 둘을 구분하는 것이 전체 개발 과정에서 중요하지 않으므로 <add> 태그 하나로 통일하여 태깅하였다.

이상과 같이 본 연구에서 말뭉치 본문에 사용한 태그는 모두 4가지뿐이다. 마크업 과정에서 태그 속의 속성의 수를 늘려 판독 과정을 더 자세히 설명할 수도 있다. 예를 들면 판독 내용에 자신이 있는 경우와 없는 경우의 정도성을 속성으로 표현할 수도 있지만, 이렇게 되면 한 태그의 길이가 너무 길어지고 마크업 작업도 어려워질 수 있으므로 해당 태그의 활용성을 고려하여 태그의 속성을 설정하는 것이 바람직하다.

4.2 주석 말뭉치의 마크업

이 절에서는 원시 말뭉치에 이어 주석 말뭉치의 구축을 위한 마크업 체계에 대해 설명하고자 한다. 앞서 논의한 바와 같이 본 연구의 마크업 체계에서는 원시 말뭉치와 주석 말뭉치는 한 파일 안에 존재하게 된다. 즉, [그림 4]에서 보이는 바와 같이 원시

[20] 태그는 시작 태그와 종료 태그의 짝을 맞춰야 하는데, 종료 태그를 생략하고 싶을 때에는 마지막에 "/"를 마지막에 삽입해야 한다.

[21] 각주 태그에는 다음과 같은 정보를 넣게 된다.

① 분명한 오기는 아니지만 오기로 추정되는 것에 대한 설명

② 현대어역의 보충 설명 예) 달게 : 열을 받아 뜨거워진 상태.

③ 문법적인 설명 예) 만이 : 많이. '만히'에서 'ㅎ'이 탈락된 표기.

말뭉치에 해당하는 <org>와 <exp> 다음에 <pa>가 나오게 되는데, 이는 형태 주석(Part of speech Analysis)을 의미한다.

주석 말뭉치의 내용은 조리법 단위로 붙게 되며 기본적인 태그의 구성은 [그림 7]과 같다.

```
<s tid="2">
  <tok seq="1">
    <orth>너</orth>
    <mor pos="MM">너</mor>
    <memo>너</memo>
  </tok>
</s>
<s tid="2">
  <tok seq="2">
    <orth>되</orth>
    <mor pos="NNB">되</mor>
    <memo>되</memo>
  </tok>
</s>
<s tid="2">
  <tok seq="3">
    <orth>가웃</orth>
    <mor pos="NNG">가웃</mor>
    <memo>반</memo>
  </tok>
</s>
<s tid="2">
  <tok seq="4">
    <orth>섞거</orth>
    <mor pos="VV">Y</mor>
    <mor pos="EC">어</mor>
    <memo>섞+어</memo>
  </tok>
</s>
```

[그림 7] 주석 말뭉치의 구조(술 만드는 법)

<s> 태그 내에 있는 "tid"는 원문의 <s> 태그와 연결하기 위한 것으로 현대어역(<exp> 태그)의 번호와 동일하다. <tok> 태그는 어절 단위에 붙는 태그이며 일련번호를 속성으로 갖게 된다. <tok> 내의 <orth> 태그는 원어절을 나타내고 <mor> 태그는 형태소 단위의 분석 결과를 나타내는데, 분석 표지는 "pos"의 속성값으로 표현하였다.

마지막에 나오는 <memo> 태그는 여러 가지 용도로 활용하기 위해 마련한 예비 태그인데, 본 연구에서는 3.1에서 기술한 다양한 유형의 이형태를 통합하기 위해 간

단하게 현대어 주석을 넣는 것으로 활용하였다. 예를 들어 "말리다" 라는 의미로 쓰이는 "말로이다, 말노이다, 말뢰다, 말뇌다" 등의 이형태들을 대상으로 그들의 통합형을 현대어로 정하여 <memo>에 "말리+다" 로 표기하는 것이다. 이렇게 처리하면 원어-현대어의 용어 대조표가 완성되어 활용될 수 있을 뿐만 아니라, 용례 검색기 내에서도 하나의 형태를 검색 조건으로 주게 되면 이들의 이형태의 용례까지도 한꺼번에 검색될 수 있다는 이점이 있다.

형태 주석 이후에는 구문 주석이나 의미 주석, 담화·화용 주석 등이 더 추가될 수 있는데, 이러한 주석은 <pa> 태그 다음에 영역을 만들어 주석 말뭉치를 확장할 수 있다.

5. 결론 및 남은 문제

지금까지 한글 필사본 음식조리서를 대상으로 원시 말뭉치와 주석 말뭉치의 마크업 방안에 대해 살펴보았다. 이를 위해 2장에서는 그간 역사 자료를 대상으로 이용되어 온 마크업 사례와 문제점을 조사·분석하였고, 3장에서는 본 연구가 대상으로 하는 텍스트의 주제적, 구조적 특성을 살펴보았다. 4장에서는 3장에서 살펴본 텍스트의 특성을 효율적으로 반영하기 위한 마크업 방안이, 전체 구조, 헤더, 본문, 주석 말뭉치로 나누어 제시되었다.

2장에서 살펴본 마크업 요건들과 본 연구의 주요 대상인 한글 필사본 음식조리서 텍스트의 특성을 고려할 때, 마크업 작업은 유연성과 확장성이 보강된 XML 체계를 기반으로 하는 것이 바람직하였고, 따라서 TEI-P5를 변형하여 한글 필사본에 필요한 요소를 분석함으로써 전체 연구 과제의 목적에 부합하는 마크업 방안을 제시하였다.

본 연구에서 제시한 마크업의 가장 큰 특징은 원시 말뭉치와 주석 말뭉치를 한 파일에 음식명 단위로 배치하고, 텍스트 구조의 이해에 필요하고 전체 연구 과제에서 유의미한 정보로 이용될 수 있는 태그들만을 구성하였으며, 각 태그마다 반드시 있어야만 하는 속성들을 설정한 것이다. 이 마크업 방안은 음식조리서 텍스트를 대상으로 하였지만, 다른 장르의 국어사 자료 마크업에도 활용되어 국어사 자료의 전산화에 기여할 것으로 생각된다.

이제 남은 문제는 이렇게 설정된 마크업이 본 연구 과제의 최종 목표인 검색기의 개발이나 그 밖의 연구에서 얼마나 효율적으로 활용될 수 있는가를 실제적으로 보이는 것인데, 이는 이미 개발된 XML 응용 소프트웨어를 이용하여 마크업이 된 말뭉치를 언어학적으로 활용하는 방법 등을 말한다.

< 참고문헌>

Burnard, Lou. 2007. BNC XML 소개. 국립국어원 (편) 저 *21세기 세종계획 최종 성과*

*발표회 자료집*에서. 국립국어원.

강범모. 1998. 한국학 문헌의 전산화를 위한 TEI 부호화 방안의 응용과 확장. 기술문서, 학술진흥재단. 연구결과보고서.

국립국어연구원. 2000. 21세기 세종 역사 자료 말뭉치 구축. 기술문서, 국립국어연구원.

국립국어연구원. 2001. 21세기 세종 역사 자료 말뭉치 구축. 기술문서, 국립국어연구원.

문헌정보처리연구회. 2002. TEI 가이드라인 (P3). 기술문서, 문헌정보처리연구회.

안의정. 1999. 한국어 입말뭉치 전사 방법 연구. 석사학위 논문, 연세대학교 대학원.

이기황. 2007. XML을 이용한 주석 말뭉치의 구조화와 활용. *제1회 한국 언어·문학·문화 국제학술대회*에서. 연세대학교 국어국문학과.

홍윤표. 2006. 국어사 연구를 위한 전자자료 구축의 현황과 과제. 임용기와 홍윤표 (편) 저 *국어사 연구 어디까지 와 있는가*에서. 태학사.

< 웹사이트>

http://www.natcorp.ox.ac.uk/XMLedition/URG/cdifsp.html#cdif22

http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

http://www.w3.org/XML/

접수 일자: 2008년 11월 5일

게재 결정: 2008년 12월 13일