국어 정보화와 새국어생활

이건식 ·한국학중앙연구원 어문생활사연구소

1. 서언

이 글은 《국어생활》과 《새국어생활》에서 국어 정보화 분야의 특집 주제로 다룬 내용을 게재 시기 적절성의 관점에서 분석하고, 디지털 정보 혁명 시대의 국어 생활에서 요구되는 국어 정보화 분야의 과제 일부를 제안하고자 한다.

국어 정보화의 영역은 대체적으로 국어 사용 정보 환경, 말뭉치 언어학(corpus linguistics), 전산 언어학(computational linguistics) 등으로 나뉜다. 컴퓨터를 국어 생활의 도구로 사용하는 문제를 다루는 국어 사용 정보 환경의 영역에 중점을 두어 이 글의 논의를 진행하고자 한다. 학문 연구 영역에 속하는 말뭉치 언어학과 전산 언어학의 영역은 제한을 두어 논의를 진행하고자 한다.

말뭉치 언어학은 언어학자의 직관보다는 언어 자료를 중시하고, 언어 단위의 출현 빈도를 계량적으로 측정하여 언어의 본질을 규명하려는 언어 연구 방법이다(Stefan, 2009:2). 그러나 국내에서는 이 언어 연구 방법론이 심도 있게 전개되지 못하고 있는 실정에 있어 이 글에서는 논의하지 않기로 한다. 전산 언어학의 경우에는 자연어 처리 컴퓨터 장치를

만들고자 하는 이론적인 문제보다는 자연어 처리 컴퓨터 장치를 일상 생활에 유익하게 활용하는 문제에 국한하여 논의를 진행하고자 한다. 예컨대 자연어 처리 컴퓨터 장치의 하나인 기계 번역기를 선택하고 사용하는 문제도 컴퓨터를 국어 생활의 도구로 활용하는 국어 사용 정보 환경의 영역에 속하는 것으로 판단하기 때문이다.

2. ≪새국어생활≫을 통해 본 국어 정보화

2.1. 특집 주제의 분야와 게재 횟수

≪국어생활≫은 1984년 10월에 창간호가 나온 이래 1990년까지 23차례 발간되었다. 그리고 ≪새국어생활≫은 1991년부터 지금까지 79차례 발간되었다. 창간호부터 줄곧 매 호마다 특집 주제 하나를 정하고 여러 편의 글을 게재하였다. 국립국어원에서 잠정적으로 마련한 분류안과 목록[1]에 따라 특집 주제의 분야별 게재 횟수를 제시하면 다음과 같다. 게재 횟수가 높은 순서로 분야를 제시한다. 다음에 제시한 분야별 게재 횟수는 중복 분류된 특집 주제를 포함한 것이다.[2]

---

1) 이 글에서 제시한 특집 주제의 분야별 게재 횟수는 국립국어원 담당자가 제공한 분류안과 목록에 기초한 것이다.

2) 국립국어원이 제공한 분류안과 목록은 부분적으로 특집 주제를 중복 분류하고 있다. 어휘/사전 분야에 분류된 '국어 속의 外來語(2)'는 국어 순화 분야, '호칭어(19)'는 국어 교육 분야, '북한의 국어 사전(3-4)'은 국어 정책 분야에도 포함되었다. 국어 교육 분야에 분류된 '어문 정책(20)'은 어문 규범 분야에도 포함되었다. 어휘/사전 분야의 '한국어의 어휘 의미망(17-3)'과 국어 정책 분야의 '21세기 세종 계획의 성과와 전망(19-1)'은 필자가 국어 정보화 분야로 중복 분류하였다. '19' 등의 괄호 속 숫자는 ≪국어생활≫의 19호를, '19-1' 등은 ≪새국어생활≫의 19권 1호를 나타낸다.

<표 1> 특집 주제 분야와 게재 횟수

분야	게재 횟수
어휘/사전	21
국어 정책/법과 제도/남북 언어 통일	19
국어 교육/국어 능력/언어 예절/대우법/문법	17
어문 규범/문자/표기	15
국어학사/문화 인물	11
국어 순화	9
관련 학문	6
방언	4
국어 정보화	6
합계	108

<표 1>에 제시된 9개 분야의 중요도가 동일하다면, 국어정보화 분야는 50% 정도만 다루어진 셈이다. 즉 12번 특집으로 다루어져야 하나 6번만 특집으로 다루어진 것이다. 그런데 9개 분야의 중요도는 모두 다를 것이다.

9개 분야의 중요도를 결정하는 것은 어려운 문제이다. 관점에 따라서 또 시대에 따라서 중요도가 다르게 결정될 것이다. 그런데 21세기를 흔히 디지털 정보 혁명의 시대라 말하고 있는 사실을 고려하면, 국어 정보화 분야가 특집 주제로 적게 다루어진 것이다. 디지털 정보 혁명으로 국어 분야에서 일어난 변화 가운데에 중요한 것들을 특집 주제로 다루어야 할 것이다. 특집 주제로 앞으로 다루어야 할 국어 정보화의 문제에 대해서는 제3장에서 기술하도록 한다.

2.2. 국어 정보화 특집 내용의 분석과 평가

2.2.1. 《국어생활》 11호(1987년 겨울)의 '국어와 컴퓨터'

'국어와 컴퓨터'를 특집 주제로 하여 6편의 글을 게재하였다. 한글 타자기, 워드프로세서 전용기, 컴퓨터 등의 국내 사용 현황을 소개하면서,

주로 한글 기계화의 중요성에 대해 논의하였다.

유경희의 '컴퓨터와 국어 생활'은 컴퓨터에 구현된 한글 구성 원리를 소개하였다. 이 글은 컴퓨터 분야에서만 논의되었던 한글 구성 원리를 국어 분야에 소개한 것이다. 김정수의 '국어 연구와 전산기'는 초성, 중성, 종성을 '불'과 같이 모아 쓰지 않고, 자모를 풀어 쓰고 왼쪽으로 45도 기울여 '  '과 같이 쓰는 방법을 제안하였다. 그리고 컴퓨터의 신속한 정보 검색 기능을 활용하여 국어 형태 연구에 컴퓨터가 긴요한 도구로 쓰일 수 있음을 간단하게 논의하였다. 송현의 '내가 쓰는 워드프로세서'는 문학 저작 도구로 컴퓨터를 사용한 경험을 언급한 것이다. 이 글은 작가(作家)의 작품 저작 도구로 컴퓨터가 필기 도구를 대신하게 될 것을 예고한 것이다.

컴퓨터가 사회 전 분야의 필수 도구로 확산되는 당시의 상황을 고려하면, 국어 분야에서 컴퓨터 사용을 강조한 이 특집은 적절한 시기에 논의가 진행된 것이다.

2.2.2. 《국어생활》 16호(1989년 봄)의 '컴퓨터와 국어 생활'

'컴퓨터와 국어 생활'을 특집 주제로 하여, 6편의 글을 게재하였다. 옛 글자를 컴퓨터에서 처리하는 방법을 소개하였고, 컴퓨터를 도구로 하는 국어 연구의 장점에 대하여 논의하였다.

金忠會의 '국어 자료 처리를 위한 개인용 컴퓨터의 시스템 설치에 대하여'와 洪允約의 '컴퓨터의 입문에서 활용까지'의 핵심적인 내용은 문서 편집기 '보석글'[3]에서 사용할 수 있는 옛 글자 목록과 옛 글자 사용 방법을 소개한 것이다. 金忠會(1989:54)는 옛 글자와 구결자를 포함한 1,224자의 '한자/고어 마스터 1.0'을 제안하였다. 이 목록은 한국어전산학회가

3) 문서 편집기 '보석글'은 소위 TGEDIT라 불린 것으로 1980년대 말과 1991년까지 국어 연구계에서 사용되었다. 1992년 HWP 2.0의 출현을 계기로 국어 연구계에서는 문서 편집기로 '보석글'을 사용하지 않고 HWP를 사용했다.

마련한 것으로 국어사 자료를 컴퓨터로 처리하는 기반을 만든 것이다. 鄭仁祥의 '資料 카드의 컴퓨터 活用 方法에 대하여'와 김병선의 '국어 문헌 자료의 처리 방법'은 국어 자료의 컴퓨터 정보 처리 방법을 소개하였다. 이 처리 방법이 종래의 종이 카드에 의한 자료 정리 방법을 효과적으로 대체할 수 있음을 언급하였다. 한편 金興圭의 '國文學 硏究에 있어서의 컴퓨터 활용'은 문학 작품의 문체에 대한 통계적인 방법의 연구가 컴퓨터로 가능할 것임을 소개하였다.

당시 국어사 연구계에서 컴퓨터를 사용하기 위한 관건은 옛 글자의 자유로운 입출력이었다. 이러한 문제를 해결한 사례를 중심으로 내용을 구성한 이 특집은 적절한 시기에 논의를 진행한 것이다.

2.2.3. 《새국어생활》 제9권 1호(1999년 봄)의 '국제 한자의 표준화'

'국제 한자의 표준화'를 특집 주제로 하여 6편의 글을 게재하였다. 당시까지 한·중·일 삼국 간에 한자의 국제 표준화의 논의에서 이루어진 것을 중심으로 특집의 내용을 구성하였다.

허성도의 '국학 연구용 상용 한자 제정안', 김홍규의 '국어 생활의 한자 사용 빈도 연구', 남윤진의 '국어사전 표제어의 한자 빈도' 등은 컴퓨터에서 우선적으로 구현해야 할 한자의 필수 목록에 대한 논의이다. 이재훈의 '국제표준문자코드 제안 한자 자형의 표준화에 대한 연구'는 한·중·일 삼국의 유니코드 통합 한자(CJK)의 한자 자형을 표준화하는 것에 대한 논의이다. 권인한(1999:89)의 '한국 한자음의 표준안 연구'는 유니코드 통합 한자(CJK)에 정의된 27,484자의 한자가 윈도우즈 등과 같은 범용 프로그램에 사용됨에 따라 한국이 제안한 17,184자의 한자음을 표준화한 것이다. 이준석과 이경원의 '한자 異體字典 편찬 연구'는 한국의 문헌 자료에 출현한 異體字를 정리하기 위한 방법을 제안한 것이다.

'국제 한자의 표준화' 특집은 미래 지향적인 주제가 아니라 과거의 주

제를 다룬 점에서 다소 아쉬운 점이 있다. 한자의 국제 표준화 문제가 한·중·일 삼국 간에 논의되기 전에 우리 한자 사용 실태에 대한 연구를 진행하였다면, 한·중·일 삼국의 유니코드 통합 한자(CJK)는 보다 우리의 실정에 적합하도록 정의되었을 것이다. 소위 CJK BMP로 불리는 한자 목록에는 20,902자의 한자가 정의되었는데, 유니코드 컨소시엄(2007: 411)에 따르면 이 한자 목록은 1992년까지 한·중·일 삼국 간에 합의된 것이다.

2.2.4. 《새국어생활》 제11권 2호(2001년 여름) '한국어 자료 정리 방안'

'한국어 자료 정리 방안'을 특집 주제로 4편의 글을 게재하였다. 그런데 이 특집이 국어 정보화를 염두에 두고 마련된 것은 아니다. 한국어 자료가 문화 유산임을 강조하고, 그 수집 방법과 정리 방안을 논의한 특집이다. 이호영의 '한국어 음성 자료의 수집과 정리'와 홍윤표의 '한국어 전자 자료의 수집과 정리 및 활용 방안'의 2편만이 국어 정보화와 관계되는 문제를 논의하였다.

홍윤표(2001:39-40)의 논의는 전자 자료가 가지는 효용 가치를 중시하여, 전자 자료의 표준화 구축 방법을 제안한 것이다. 홍윤표(2001:46-61)의 논의는 전자 자료의 메타데이터와 데이터의 구축 방법에 대한 구체적인 사례를 제시하여 전자 자료의 경우 표준화가 중요한 문제임을 강조하였다.

전자 자료가 사회 전 분야에서 개별적으로 구축되기 때문에 전자 자료의 메타데이터와 데이터는 독자적인 방식으로 구축된다. 그런데 전자 자료가 표준화된 형태의 메타데이터와 데이터로 구축된다면, 그 전자 자료는 별도의 수정 없이 다른 분야의 자료로 활용될 수 있을 것이다. 이 점에서 홍윤표(2001)의 전자 자료 표준화 제의는 우리가 해결하고 이루어 내야 할 과제라고 생각한다. 그리고 IMF로 인하여 1998년 '정보화 근로

사업'이 착수되어 우리 사회에서 전자 자료가 대규모적으로 전산화된 사정을 고려하면, 홍윤표(2001)의 전자 자료 구축에 대한 표준화 논의는 적절한 시기에 논의를 진행한 것이다.

2.2.5. 《새국어생활》 제17권 3호(2007년 가을)의 '한국어의 어휘 의미망'

'한국어의 어휘 의미망'을 특집 주제로 3편의 글을 게재하였다. 자연어 처리(National Language Processing)를 위해서 어휘 의미 관계망의 구축이 필요함을 강조한 특집이다.

윤애선의 '국내외 어휘 의미망의 구축과 활용'은 국내외 어휘 의미망 구축의 사례를 소개하고, 각 사례의 장점과 단점을 논의한 글이다. 옥철영(2007:30-35)의 '국어 어휘 의미망 구축의 개념과 사전 편찬'은 울산대학교 한국어처리연구실에서 개발한 '사용자 어휘 지능망(User-Word Intelligent Network)'의 구조 체계를 소개한 것이다. 그리고 옥철영(2007:36-47)은 '사용자 어휘 지능망'이 사전 편찬에 효과적으로 활용될 수 있음을 논의하였다. 이성헌의 '세종 전자 사전의 어휘 의미 부류 체계'는 21세기 세종 계획으로 개발된 세종 전자 사전이 채택한 어휘 의미 부류 체계를 설명한 것이었다.

김삼묘·유기영(2004:129)에 따르면, 4가지 종류의 형식 언어(Formal Langiage)[4]를 인식하는 4가지 종류의 컴퓨터 계산 장치(Automata)[5]를 만들 수 있다고 한다. 형식 언어를 인식하는 계산 장치는 해당 형식 언어의 의미를 기계적으로 이해할 수 있는 능력을 가지고 있다. 그런데 자연

4) 4가지 종류의 형식 언어는 구문 구조 기반 문법(Phrase Structured Grammar), 문맥 민감 문법(Context Sensitive Grammar), 문맥 자유 문법(Context Free Graammar), 정규 문법(Regular Grammer) 등이다.

5) 4가지 종류의 컴퓨터 계산 장치는 TM(Turing Machines), LBA (Linear-Bounded Automata), PDA(Pushdown Automata), FA(Finite State Autamat) 등이다.

언어(Natural Language)의 경우에는 현재 그러한 계산 장치가 구현된 사례가 없다. 다만 전산 언어학 분야에서는 기계 번역기 등의 개발 연구를 통해서 형식 언어의 계산 장치와 유사한 기능을 수행하는 계산 장치를 개발하려고 노력하고 있다.

컴퓨터 계산 장치를 활용하여, 디지털 정보를 추론하고, 검색하고, 분류하는 일은 디지털 정보 혁명 시대에 반드시 해결해야 할 과제이다. 이 점에서 이 특집은 의미 있는 논의를 소개한 것으로 평가된다. 다만 우리의 독자적 개념 인식이 한국어에 반영된 것이므로 전산 언어학자뿐만 아니라 국어학자도 이 문제 해결에 동참해야 할 것으로 생각된다.

2.2.6. 《새국어생활》 제19권 1호(2009년 봄)의 '21세기 세종 계획의 성과와 전망'

'21세기 세종 계획의 성과와 전망'을 특집 주제로 5편의 글을 게재하였다. 1998년부터 2007년까지 진행된 '21세기 세종 계획'의 사업 성과를 총괄(홍윤표:2009), 자료 구축(서상규:2009), 전자사전 개발(이성헌:2009), 언어 정보화(박형익:2009) 등의 분과로 나누어 정리하였고, '21세기 세종 계획' 사업을 계승할 후속 사업의 방향(홍종선·남경완:2009)에 대하여 논의하였다.

10년 동안 국어 정보화의 경험을 일목 요연하게 정리하여, 향후의 국어 정보화에 대한 논의의 기초 자료를 제공한 점에서 이 특집은 의의를 가진다. 앞으로 계획되어 착수될 국어 정보화 사업은 이 특집에 기초하여 보다 개선된 방안을 마련하여 추진될 것으로 기대된다.

홍윤표(2007)에 따르면, 10년 동안 추진된 '21세기 세종 계획'으로 6억 5천 4백만 어절의 말뭉치 자료가 구축되었고, 205만 단어를 포함한 전자사전이 구축되었다. 이러한 성과물은 국어에 대한 학문 연구의 중요한 수단으로 기여할 것이다. '21세기 세종 계획'으로 구축된 말뭉치 자료가 국

어 연구계에 배포되어 활용됨으로써 국어 연구의 수준이 향상되었음은 누구나 인정하는 바다.

3. ≪새국어생활≫을 위한 국어 정보화의 과제

3.1. 국어 사용 정보 환경의 표준화

컴퓨터가 국어 분야에 사용되기 시작하던 초창기인 1980년대에는 옛 글자, 한자 등의 자유로운 입출력이 컴퓨터 사용의 관건이었다. 그러나 1989년 HWP가 출시되고, 1991년의 HWP 2.0과 1999년의 윈도우즈 2000이 출시되어 유니코드가 일상화되면서, 컴퓨터에서 옛 글자, 한자 등을 자유롭게 입출력할 수 있게 되었다. 그런데 세계 각국의 문자를 종합한 유니코드의 특성이 전산화의 다른 문제를 야기하고 있다.

유니코드는 세계 각국의 문자 집합을 통합하였기 때문에 한국에서는 크게 2가지 문제가 발생한다.

하나의 문제는 한글 맞춤법에서 규정한 문장 부호를 전산 입력할 때 발생한다. 한글맞춤법에서는 문장 부호의 형태와 사용 방법을 규정했다. 그러나 문장 부호에 대응하는 유니코드 값을 규정하지 않았다. 이 결과로 가운뎃점 '·', 여는 소괄호 '(', 닫는 소괄호 ')', 여는 중괄호 '{', 닫는 중괄호 '}', 여는 대괄호 '[', 닫는 대괄호 ']', 빠짐표 '□', 빠짐표 '×' 등의 문장 부호는 상이한 유니코드로 입력될 가능성이 많다. 예컨대 문장에서 흔히 쓰는 가운뎃점은 라틴 문자 집합의 'U+B7(·)' 코드나 일반 구두점 집합의 'U+2024(·)'로 상이하게 입력될 가능성이 있다. 가운뎃점은 심지어 호환용 한글 자모인 'U+318D(·)'로 입력될 가능성이 있다. 실제로 가운뎃점은 가독성 때문에 'U+318D'로 흔히 입력된다. 전산 자료의 표준화를 위해서 문장 부호의 유니코드 값을 고시할 필요가 있다.

또 다른 문제는 유니코드 한·중·일 통합 한자 27,484자 중 17,184자

만 한국이 제출한 점이다. 권인한(1989:89)은 17,184자만 한자음을 정리하였다. 다른 나라에서 제안한 한자에 대해서 한자음을 정할 것인가의 문제에 대한 논의가 필요하다. HWP 2007은 4만여 자의 EXT_B 한자를 쓸 수 있도록 하였다. 이 한자의 대부분은 한자음이 없어 HWP 2007에서는 코드 값으로만 입력할 수 있다. 유니코드 한·중·일 통합 한자 8만여 자 모두에 한자음을 부여할 필요가 있는지에 대한 여론 수렴이 필요하고, 그에 따른 문제를 해결할 필요가 있다.

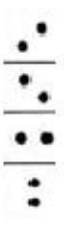
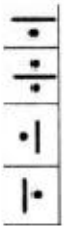
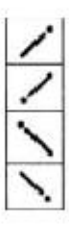
3.2. 국어 사용 정보 환경의 개선

완성형 옛 글자, 조합형 옛 글자, 구결자 등은 현재 HWP와 MS 워드에서 입력할 수 있다. 그런데 조합형 옛 글자와 구결자는 ISO나 유니코드에 정식으로 등록되지 않았다. 마이크로소프트 사에서 한국의 상황을 고려하여 MS 워드 2000을 출시할 때, 잠정적으로 그러한 문자 집합을 유니코드에 추가한 것이다. 따라서 완성형 옛 글자, 조합형 옛 글자, 구결자 등의 문자 집합을 ISO나 유니코드에 정식으로 등록하는 절차를 진행해야 할 것이다. 그리고 MS 워드 2000에서 잠정적으로 채택한 완성형 옛 글자, 조합형 옛 글자, 구결자 등의 구성 방법을 정리하여 소개할 필요가 있다. 완성형 옛 글자와 구결자는 2바이트로 구성되었고, 조합형 옛 글자는 6바이트로 구성되었는데, 이 사실을 모르는 프로그래머들이 완성형 옛 글자, 조합형 옛 글자, 구결자 등을 잘못 처리하는 경우가 발생하기 때문이다.

일본에서 '㆐, ㆑, ㆒, ㆓, ㆔' 등과 같은 훈독 지시 부호를 유니코드에 등록하였고, 중국에서는 '〡, 〢, 〣, 〤, 〥, 〦, 〧, 〨, 〩' 등과 같은 산가지[算木] 부호를 유니코드에 등록하였다. 이 점을 고려하면, 각필 구결 자료에 나타난 부호 구결자를 유니코드에 등록하는 작업을 진행할 필요가 있다. 부호 구결자에는 역독선, 지시선, 단점(·), 선, 쌍점, 눈썹, 느낌

표 등이 있음이 알려졌다. 다음은 장경준(2008:77)에서 가져온 선, 쌍점, 눈썹, 느낌표 등의 형태이다.

<표 2> 부호 구결자의 형태

선	쌍점	눈썹		느낌표	
					

<표 2>에 제시된 것을 포함하여 부호 구결자를 유니코드에 등록하는 문제를 논의할 필요가 있다.

한글 폰트의 형태적 변별성도 깊이 연구할 필요가 있다. '학장실'과 '화장실'의 형태적 유사성 때문에 대학의 건물 내에서 '화장실'을 가고자 하는 학생들이 '학장실' 문을 여는 경우가 있다는 이야기[6]는 웃고 넘길 문제가 아니다. 이러한 문제를 해결하고자 문화관광부에서 '문화부 바탕체'를 개발·보급한 바 있으나 한글 폰트의 형태적 변별성을 담보하는 논의를 심도 있게 전개할 필요가 있다고 생각한다.

3.3. 말뭉치 자료(corpus)의 표준화와 공동 이용

국어 말뭉치 자료(corpus)는 국어 연구의 출발점이다. 말뭉치 자료를 다루는 방법은 각 분야마다 다르지만 전통적 방식의 국어 연구자, 말뭉치 언어학자, 전산 언어학자 모두에게 말뭉치 자료는 학문 연구에 중요한 재

6) 이 사례는 홍윤표 선생님께서 말씀해 주신 것이다.

료가 된다. 그런데 디지털 정보 혁명 시대인 요즈음에는 말뭉치 자료가 흔히 전자 자료로 구축되거나 활용되고 있어서, 우리는 말뭉치 전자 자료 구축 방법의 표준화 방안과 구축된 전자 자료의 공동 이용 방안을 모색할 필요가 있다.

전자 자료의 경우에는 공동 활용을 위해서 표준화된 방식으로 구축되어야 한다. 이 조건을 만족할 경우에만 다른 사람이 구축한 전자 자료를 저렴한 비용으로 활용할 수 있기 때문이다. 그런데 구축된 전자 자료가 독자적인 방식으로 입력되어 있을 경우, 전자 자료를 재활용하기 위해서는 적지 않은 비용을 지불해야 한다.

그럼에도 불구하고 홍윤표(2001) 이외에는 전자 자료 구축의 표준화 방안을 제기하고 있지 못한 실정에 있다. 홍윤표(2001:47-49)가 '헤더'로 부르는 메타데이터와 '본문 입력 양식'으로 부르는 데이터 구축 방안을 출발점으로 삼아, 메타데이터와 데이터의 표준화된 양식에 대해 논의할 필요가 있다. 물론 메타데이터와 데이터가 논리적으로 명쾌하게 구분되어 구축되었다면, 그 형식을 저렴한 비용으로 자유롭게 변환할 수도 있다. 그러나 표준화된 지침을 가지지 못하면, 대개는 메타데이터와 데이터를 의미론적으로 분별할 수 없도록 구축하여 자료 형식의 변환에 막대한 비용을 지불하게 된다. 국어 연구 말뭉치 자료 구축을 오랫동안 수행한 국내의 경험을 집적하여, 전자 자료 구축의 표준화 방법을 체계적으로 마련할 필요가 있다.

홍윤표(2010:16-24)를 보면, 세종 말뭉치 자료, 국내 연구 기관의 말뭉치 자료, 북한과 중국 연변 지역에서 구축한 말뭉치 자료 등이 소개되어 있다. 세종 말뭉치 자료는 현재 누구나 다 자유롭게 이용할 수 있으나 후자의 2개 말뭉치 자료는 그렇지 못하다. 국내 연구 기관 말뭉치 자료의 경우, 우선은 말뭉치 자료를 구성한 문헌 목록을 작성하여 배포할 필요가 있다. 그리고 국가에서 비용을 지불하고, 그 말뭉치 자료를 전 국민 모두가 활용할 수 있도록 관련 정책을 수립할 필요가 있다.

3.4. 자연 언어 처리 연구 성과의 대국민적 소개

자연 언어 처리 연구는 인간의 언어를 컴퓨터 계산 장치로 이해하는 것을 궁극의 목적으로 추구한다. 그러나 현재 자연 언어 처리 연구는 인간의 언어를 완벽하게 이해하는 계산 장치를 만들지 못하고 있다. 그럼에도 불구하고 현재까지 이루어진 자연 언어 처리 연구의 성과는 우리의 생활을 풍요롭게 해 준다.

문서 작성기에 탑재된 맞춤법 검사기가 100% 완벽하지는 않지만 문서 작성 시 맞춤법 적용을 힘들어 하는 사람들에게는 매우 유용한 도구임에 틀림 없다. 한국어를 영어로 번역하거나 영어를 한국어로 번역하는 경우에 현재에는 그 기계 번역 결과가 만족스럽지 않다. 그러나 영어를 능숙하게 구사하지 못하는 사람들에게는 도움이 될 것이다. 스마트 폰으로 인터넷을 검색할 때, 음성으로 검색 문자열을 입력한 결과가 다소 만족스럽지 못한다고 하더라도 경우에 따라서 이러한 방법이 반드시 필요할 경우가 있다. 그리고 입력된 문서를 음성 발화로 변환시켜 주는 음성-텍스트 변환기의 변환 결과가 비록 만족스럽지 않다고 하더라도, 음성-텍스트 변환기는 시각 장애인에게는 빛과 같은 존재일 것이다.

맞춤법 검사기, 기계 번역기, 음성 인식기, 음성-텍스트 변환기 등이 상업 제품으로 출시되어, 공공 기관에서 각 제품의 장점과 단점을 평가하기에는 어려운 점이 있다. 그럼에도 불구하고 디지털 정보 혁명 시대의 국어 생활 개선을 위해, 국립국어원은 국민들에게 일상 생활에 유용한 자연 언어 처리 도구를 안내해 줄 의무가 있다고 생각한다.

4. 결언

《새국어생활》의 기존 국어 정보화 특집 내용을 분석하고 평가한 다음, 이 글은 국어 사용 정보 환경의 관점에 국한하여 《새국어생활》의

국어 정보화 분야에서 다루어야 할 과제에 대해 논의하였다.

국어에서 사용되는 모든 문자 집합이 컴퓨터에서 자유롭게 표현되고, 컴퓨터로 생산되는 문서와 데이터가 사회 구성원 간에 무리 없이 상호 교환할 수 있고, 컴퓨터로 기록된 우리 언어를 컴퓨터 장치로 손쉽게 이해할 수 있는 환경을 마련하는 것이 디지털 정보 혁명 시대에 걸맞은 우리의 새 국어 생활임을 강조하였다.

이러한 국어 사용 정보 환경을 심도 있게 준비하기 위해서는 국어를 사용하는 사람들의 국어 정보 처리 능력도 증진시켜야 하고, 자연어 처리를 위한 제반 방법 개발이 심도 있게 진행되어야 하나 이에 대해서는 지면의 제약으로 논의를 진행하지 못하였다.

참고문헌

권인한(1999), 한국 한자음의 표준안 연구, 《새국어생활》 제9권 1호, 87-100.

김병선(1999), 국어 문헌 자료의 처리 방법, 《국어생활》 16호, 109-125.

김삼묘·유기영(2005), 《계산모델:오토마타 및 형식언어》, 이한출판사.

김성익(1987), 인쇄와 전산 사식, 《국어생활》 11호, 45-49.

김정수(1987), 국어 연구와 전산기, 《국어생활》 11호, 38-44.

金貞欽(1987), 우리말 機械化의 어제·오늘·來日 《국어생활》 11호, 8-18.

金忠會(1999), 국어 자료 처리를 위한 개인용 컴퓨터의 시스템 설치에 대하여, 《국어생활》 16호, 4-60.

金興圭(1999), 國文學 硏究에 있어서의 컴퓨터 활용 《국어생활》 16호, 126-133.

김홍규(1999), 국어 생활의 한자 사용 빈도 연구, 《새국어생활》 제9권 1호, 17-48.

남기심(2001), 문화유산으로서의 국어, 《새국어생활》 제11권 2호, 5-10.

남윤진(1999), 국어사전 표제어의 한자 빈도, 《새국어생활》 제9권 1호, 49-68.

박동순(1987), 한글 컴퓨터화의 현실과 문제점, 《국어생활》 11호, 19-23.

박민규(1989), 어휘 조사의 전산 처리, 《국어생활》 16호, 134-147.

박형익(2009), 한민족 언어 정보화의 성과와 전망, 《새국어생활》 제19권 1호, 79-94.

서상규(2009), 국어 특수 자료 구축의 성과와 전망, 《새국어생활》 제19권 1호, 35-58.

송현(1987), 내가 쓰는 워드프로세서, 《국어생활》 11호, 50-55.

옥철영(2007), 국어 어휘 의미망 구축의 개념과 사전 편찬, 《새국어생활》 제17권 3호, 27-50.

유경희(1987), 컴퓨터와 국어 생활, 《국어생활》 11호, 24-37.

윤애선(2007), 국내외 어휘 의미망의 구축과 활용, 《새국어생활》 제17권 3호, 5-26.

이성헌(2007), 세종 전자사전의 어휘 의미 부류 체계, 《새국어생활》 제17권 3호, 51-68.

이성헌(2009), 세종 전자사전 개발의 성과와 전망, 《새국어생활》 제19권 1호, 59-78.

이재훈(1999), 국제 표준 문자 코드 제안 한자 자형의 표준화에 대한 연구,《새국어생활》 제9권 1호, 69-86.

이준석·이경원(1999), 한자 異體字典 편찬 연구, 《새국어생활》 제9권 1호, 101-120.

이호영(2001), 한국어 음성 자료의 수집과 정리, 《새국어생활》 제11권 2호, 11-22.

장경준(2008), 點吐口訣 硏究의 成果와 當面 課題, 《口訣硏究》 제21집, 구결학회, 67-98.

鄭仁祥(1989), 資料 카드의 컴퓨터 活用 方法에 대하여, 《국어생활》 16호, 82-108.

崔明玉(2001), 방언 자료의 수집과 정리, 《새국어생활》 제11권 2호, 23-36.

허성도(1999), 국학 연구용 상용 한자 제정안, 《새국어생활》 제9권 1호, 5-16.

洪允杓(1989), 컴퓨터의 입문에서 활용까지, 《국어생활》 16호, 61-81.

홍윤표(2001), 한국어 전자 자료의 수집과 정리 및 활용 방안, 《새국어생활》 제11권 2호, 37-76.

홍윤표(2009), 21세기 세종 계획 사업 성과 및 과제, 《새국어생활》 제19권 1호, 5-34.

홍윤표(2010), 어문 말뭉치 구축의 회고와 전망, 《차세대 어문 정보학의 전망》, 어문생활사연구소 2010년 제2차 국내학술회의 발표 논문집, 1-30.

홍종선·남경완(2009), 국어 정보화 사업의 미래와 전망, 《새국어생활》 제19권 1호, 95-117.

Stefan Th. Gries(2009), What is Corpus Linguistics?, *Language and Linguistics Compass* 3, Blackwell Publishing Ltd, 1-17.

The Unicode Consortium(2007), *The Unicode 5.0 STANDARD*, Pearson Educations Inc.