JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 22, No. 3, pp. 453-461, Mar. 2021

Check for updates

사전과 대량의 신문 기사를 활용한 남북한 어휘 대응에 대한 연구

이 현 아[1*] · 송 지 혜[2]
[1*]금오공과대학교 컴퓨터소프트웨어공학과 교수
[2]금오공과대학교 교양교직과정부 교수

A Study on Word Corresponding between South and North Korean Language using Dictionaries and News Article Corpus

Hyunah Lee[1*] · Ji-hye Song[2]
[1*]Department of Computer Software Engineering, Kumoh National University of Technology, Gumi, 39177, Korea
[2]Department of Liberal Arts and Teacher Training, Kumoh National University of Technology, Gumi, 39177, Korea

[요 약]

남북한은 같은 언어를 사용하지만 분단 이후 각기 다른 방향성과 과정으로 언어 규범을 정하고 언어 순화를 시행하면서 문법과 어휘에 많은 차이점이 생겼다. 본 연구에서는 자동으로 수집된 남북한 사전과 대량의 신문 기사를 활용하여 남북한의 어휘 대응을 시도하고 그 특징을 분석한다. 사전에서 획득한 북한 단어는 남한어 사용자도 이해할 수 있는 경우도 많았으나, 44% 가량의 단어는 남한어 사용자가 이해하기 어려운 단어로 나타났다. 북한의 <로동신문>과 <우리민족끼리>를 분석한 결과에서는 약 90%의 단어가 남한 사전이나 뉴스에도 출현하는 것으로 나타나 실제 문서에서는 남북한 단어의 공통성이 더 강한 것으로 분석되었다.

[Abstract]

The two Koreas use the same language, but after the division, many differences in grammar and vocabulary occurred as language norms were set and language refined in different directions and processes. This study attempts to correspond words of the two Koreans by using the automatically collected South and North Korean dictionaries and a large number of newspaper articles and analyzes their characteristics. In many cases, North Korean words obtained from dictionaries could be understood by South Korean users, but about 44% of the words were difficult for South Korean users to understand. As a result of analyzing North Korea's <Rodong Shinmun> and <uriminzokkiri>, about 90% of words appear in South Korean dictionaries and news, so the commonality between North and South Korean words is stronger actual documents.

색인어 : 어휘 대응, 남북한 언어 차이, 표준국어대사전, 형태소 분석, 북한 뉴스 기사

Key word : Word Corresponding, Difference between North Korean and South Korean, Pyojun-Gugeo-Daesajeon, POS tagger, North Korean News Article

http://dx.doi.org/10.9728/dcs.2021.22.3.453

Received 15 December 2020; **Revised** 20 January 2021
Accepted 20 January 2021

***Corresponding Author; Hyunah Lee**

Tel: +82-54-478-7546
E-mail: halee@kumoh.ac.kr

Ⅰ. 서 론

남북한은 같은 언어를 사용하지만 분단 이후 각기 다른 방향성과 과정으로 언어 규범을 정하고 언어 순화를 시행하면서 문법과 어휘에 많은 차이점이 생겼다. 이러한 언어적 간극은 남북의 정상적인 교류를 위해 해결해야 할 과제이다. 국어학과 언어학 분야에서는 오랫동안 북한의 조선어와 남한의 우리말에 관한 연구가 이루어져 왔다. 이들에서는 분단 이후 각 언어의 변천 과정을 국어사적 관점과 사회언어학적 관점에서 살펴보고, 현재 양 언어의 차이점을 밝히고 그 차이를 극복하는 방안을 제시하고 있다. 특히 겨레말큰사전 남북공동편찬사업회를 중심으로 한 여러 연구는 남북 언어 통일을 준비하는 기반이 되고 있다[1].

남북한 언어 차이에 대한 연구들은 대부분 한정된 남북한의 언어자료를 활용하고 있어 대용량 자료에 대한 관찰은 부족한 실정이다. 본 논문에서는 대용량의 남북한 자료로부터 남북한의 언어 차이를 단어 중심으로 관찰하고, 남북한의 단어 대응 방향을 제시하고자 한다. 이를 위하여 <표준국어대사전>[2], 통일부의 북한정보포털의 남북한 언어 비교[3], <우리말큰사전>[4]으로부터 북한의 어휘와 남한 어휘의 대응에 대한 기초 정보를 추출한다. 이와 함께 네이버의 북한 관련 뉴스[5]와 북한의 <로동신문>[6], <우리민족끼리>[7]에서 자동으로 수집된 대용량 자료에 한국어 형태소 분석기를 적용하여 어휘를 추출하고, 사전에서 얻은 어휘 대응 자료를 활용하여 남한과 북한 어휘의 대응 특성을 살펴본다.

본 논문은 다음과 같이 구성된다. 2장에서는 남북한 어휘 대응에 대한 기존 연구를 살펴보고, 3장에서는 각종 사전에 기반한 남북한 어휘 대응의 특성을 살펴본다. 4장에서는 온라인으로 수집한 남북한 매체의 뉴스 기사에서 추출한 남북한 어휘의 대응 특성을 살펴보고, 사전으로 대응어를 찾을 수 없는 어휘에 대한 자동 대응 방안을 제시한다. 5장에서는 결론을 맺고 향후 연구 방안을 제시한다.

Ⅱ. 기존 연구

국어학 분야에서는 남북한 언어 차이에 관한 다양한 연구가 진행되었다. 본 장에서는 이러한 논문 중에서 다양한 예제를 소개하고 있는 연구들을 중심으로 기존 연구들을 살펴본다.

[8]에서는 남북한 언어의 차이를 맞춤법과 어휘 중심으로 분석하였다. 맞춤법 비교에서는 (1) 사전 배열 순서와 자모 명칭의 차이, (2) 두음법칙 적용의 차이, (3) 어문규범에서 북한어는 사이시옷이 쓰이지 않는 차이, (4) 띄어쓰기 차이가 있음을 제시하였다. 어휘의 차이에서는 크게 (가) 형태 차이, (나) 같은 형태의 다른 의미 차이, (다) 신조어(ex. 가두녀성(㊫ 전업주부), 평양속도(㊫ 빠른 속도))를 큰 분류로 분석하였으며, 세분류로 형태 차이에서는 (가-1) 방언이 문화어(북한의 표준어)가 된 경우(ex. 망탕(㊫ 마구). 번지다(㊫ 거르다)) (가-2) 외래어 말 다듬기(ex. 가루소젖(㊫ 분유), 큰물(㊫ 홍수)) (가-3) 어문 규정 차이에 의한 차이(ex. 녀성, 눈섭, 이발(이빨), 아빠트, 뻐스) (가-4) 표현 차이(바쁘다(㊫ 어렵다), 면목이 있다(㊫ 안면이 있다)를, 어휘에서 같은 형태이지만 다른 의미로 (나-1) 어감 차이(ex. 보채다(㊫ 나서다), 소행), (나-2) 의미 추가(ex. 세포(㊫ 당 조직), 아저씨(㊫ 형부), 어버이(㊫ 김일성))를 제시하였다. 또한, 우리나라와 비슷한 분단 상황을 거친 두 나라인 독일과 베트남의 경우를 분석하여, 교류가 지속되었던 동서독은 언어에 큰 차이가 없는 것에 반해, 베트남은 역사적 이유로 큰 차이는 없었으나 전문용어에는 언어 차이 문제가 아직까지도 연구 과제로 남아 있음을 밝혔다.

언어 규범을 중점적으로 살피고 있는 [9]는 남한에서는 '하여'를 여 불규칙으로 보지만 북한에서는 '여'를 규칙으로 취급하고, ㅂ 불규칙을 '고마와', '가까와', '고마왔다' 로 표기하는 등의 예제를 상세히 제시하고 있다. 또한 북한에서는 부사나 명사화 접미사의 특성으로 '도리여', '드디여', '태여나다', '헤여지다', '헤염치다'로 표기하며, '넙적다리'(㊫ 넓적다리), '넙적바지'(㊫ 넓적바지), '널다랗다'(㊫ 널따랗다), '널직하다'(㊫ 널찍하다), '일찌기'(㊫ 일찍이), '금이'(㊫ 금니), '사랑이'(㊫ 사랑니) 등에서도 많은 어휘 차이가 발생하였음을 보인다.

<겨레말큰사전>과 <표준국어대사전>의 북한어를 대상으로 한 연구들도 진행되었다. [10]에서는 겨레말큰사전 편집위원인 저자가 편집 과정에서 발견한 우리 표준어와 북한 문화어의 차이와 통합 방안을 제시하고 있다. [11]에서는 겨레말큰사전의 편찬 경과와 북한어의 정의, 남북한 언어 차이의 배경, 사전편찬 방향과 문제점에 대해서 서술하고 있다. 이 연구에서는 각기 다른 어문 규범, 특히 어두 ㄹ이나 ㄴ의 표기문제, 사이시옷 문제 등이 사전편찬 과정에서 논란이 될 것으로 언급하고 있다. 또한 널리 알려진 북한어 '얼음보숭이'(㊫ 아이스크림)는 실제 잘 쓰이지 않으며, 아이스크림의 상품 이름인 '에스키모'가 더 많이 쓰인다는 등의 예시를 제시하고 있다. [12]는 주로 <표준국어대사전>의 북한어 기술에 대한 내용을 주로 다루고 있으며, 같은 형태도 의미적 차이를 가지는 경우가 있음을 지적하였다. 예를 들어 숙지는 '충분히 잘 앎'(熟知)의 의미로 쓰이지만 북한어에서만 쓰이는 의미인 '여행하거나 행군하는 도중에 묵어 가는 곳'(宿地)이 표준국어대사전에서는 첫번째 의미로 등재된 것의 문제점을 기술하고 있다. [13]에서는 <표준국어대사전>과 <조선말대사전>의 비교 자료를 제시한다.

[14]에서는 북한 방언의 상당수가 문화어에 포함되어 '남새'(㊫ 채소), '게사니'(㊫ 거위), '가마치'(㊫ 누룽지), '눅다'(㊫ 헐하다)처럼 남한의 표준어와 달라진 어휘가 있는가 하면, 한자어 어휘가 토박이말로 다듬어져서 달라진 어휘도 '흙깔이'(㊫ 객토), '바래기'(㊫ 표백), '왼쪽공격수'(㊫ 좌익수)처럼 상당하다는 점을 지적하였다. 또한, 우리말에는 광복 이후 수많은 외래어가 쓰여 차이가 발생하였으며, 어휘 다듬기 활동을 정부기관이 중심이 되어 적극적으로 추진한 북한과 달리, 남한에서는 정부 기관뿐만 아니라 민간단체에서도 전개되었다는 점도 지적하고 있다.

최근에는 온라인 데이터에 기반한 북한어 연구가 등장하고 있다. [15]에서는 KBS의 시사프로그램 '남북의 창'의 남북 총 42개 방송스크립트에 KLIWC를 사용하여 심리학적 해석을 중심으로 언어차이를 분석하였다. [16]에서는 OCR로 문자인식을 수행한 북한 <로동신문>과 남한의 <조선일보>와 <동아일보>의 각 천여 개의 기사를 AntConc로 분석하여, '-하-'와 결합되는 복합어에 대한 분석을 어휘의 수와 비율 차이 등의 수치 분석을 시행하였다.

이러한 기존 연구들은 남북한의 어휘 차이를 체계적으로 관찰하여 언어 차이 극복을 위해 의미있는 결과들을 제시하고 있으나, 주로 정적인 성격의 소규모 데이터나 사전에 의존하거나 온라인 데이터를 사용하는 경우에도 그 크기가 크지 않다. 본 연구에서는 대용량의 북한 문서에 대한 분석을 통해 남한 어휘과 북한 어휘의 차이에 대해 관찰하고자 한다. 다음에서는 남북한 사전에 기반한 어휘 대응과 대용량 뉴스 기사에서의 어휘 대응에 대해서 살펴본다.

Ⅲ. 남북한 사전의 어휘 대응

본 장에서는 북한어-남한어 대응 사전과 남한어 사전의 수집 방법과 수집된 사전의 분석 결과를 보인다.

3-1 북한어 어휘 사전 수집

북한어 어휘 수집에서는 온라인에서 수집이 가능한 <표준국어대사전>(이하 <표준>으로 약칭)[2], 통일부의 북한정보포털 남북한 언어비교[3]의 두 개의 기존 자료를 사용하였다.

<표준>은 Fig. 1[1]과 같이 남한어와 함께 북한어를 포함한다. 북한어로 총 61,002 표제어를 포함하고 있으며, 어휘의 품사별 4개수는 Table 1과 같다. <표준>에서 같은 품사를 가지면서 다른 표제어로 나뉘어 등록된 동음이의어는 처리상의 편의를 위해 하나의 다의어로 합병하였으며, 결과로 총 60,674개의 어휘를 대상으로 대응을 진행하였다.

가까운갈래 ***ga-kka-un-gal-lae***
명사 북한어 생물 '근친1(3. 혈연이 가까운 관계)'(近親)의 북한어.
Noun. North Korean, Life, North Korean word for 'geun-chin1(3. a near relation)'
.....
게사니 ***ge-sa-ni***
1. 명사 방언 '거위1(오릿과의 새)'의 방언(강원, 경기).
Noun. Dialect. Dialect(gang-won, gyeong-gi) of goose
2. 명사 북한어 동물 '거위1(오릿과의 새)'의 북한어.
Noun. North Korean Animal, North Korean word of goose
....
분각 分刻 ***bun-gag***
명사 북한어 매우 짧은 시간. *Noun. North Korean. no time*

그림 1. <표준국어대사전>[2]의 북한어 일부
Fig. 1. Examples of North Korean in PyoJun Dictionary[2]

1) Some of tables and figures inserted in this paper are written in both Korean and English due to the specialty of the paper that analyzes Korean language itself.

표 1. <표준국어대사전>의 품사별 북한어 개수
Table 1. Word Count per POS of North Korean part in [2]

POS	Count	POS	Count	POS	Count
Noun	44988	Dependent noun	83	Ending	22
Verb	8939	Interjection	82	Pronoun	17
Adverb	4298	Determiner	54	Numeral	8
Adjective	2485	Affix	22	Postposition	4

Fig.2는 통일부에서 제공하는 북한정보포털의 남북한 언어비교(이하 <통일부>)이다. 남북한 언어비교 자료는 품사 정보 없이 남한말과 북한말의 1025개 쌍을 제공한다.

그림 2. 통일부의 북한정보포털[3]의 남북한 언어비교
Fig. 2. Comparison between South and North Korean Languages in [3]

하나의 북한 어휘에 여러 남한 어휘가 대응될 수가 있어 중복을 제외하여 수집하여 965개의 북한 어휘를 얻었으며 수동으로 품사를 부착하였다.

3-2 남한어 어휘 사전 수집

3-1에서 획득한 북한어 사전의 어휘와 남한어 어휘의 대응 관계를 확인하기 위해 <우리말큰사전>(이하 <우리말>)[4]의 어휘를 추출하였다. <우리말>은 406,959개의 표제어로 구성되어 있으며 품사별 단어수는 Table 2와 같다. 7,974개의 단어는 적합한 품사가 표시되어 있지 않았으며, 자동사, 타동사/불완전 타동사와 같이 세부 품사가 기술된 경우는 최상위 품사 동사(Verb)로 표시하였다.

표 2. <우리말큰사전>의 품사별 단어 개수
Table 2. Word Count per POS of [4]

POS	Count	POS	Count	POS	Count
Noun	304351	Determiner	1111	Postposition	396
Verb	56003	Interjection	954	Numeral	339
Adjective	16064	Dependent Noun	808	Stem	274
Adverb	15817	Postfix	447	Prefix	267
Ending	1708	Pronoun	439	Copula	7

표 3. 사전의 남북한 어휘 대응 결과
Table 3. Word Mapping of South&North Korean Dictionary

Source	Mapping Type				Count	Ratio
PyoJun	its own South Korean Word	same South Korean Word with Urimal-Keun-Sajeon	Same Chinese	①	8134	13.4%
			Different Chinese	②	168	0.3%
			No Chinese	③	7632	12.6%
		same South Korean Word with Tongil-Bu		④	69	0.1%
		no match to Urimal or Tongil-Bu		⑤	9029	14.9%
	no South Korean Word	Same Word Form with Urimal-Keun-Sajeon	Same Chinese	⑥	996	1.6%
			Different Chinese	⑦	1350	2.2%
			No Chinese	⑧	3373	5.6%
		Different Word form but Same Chinese with Urimal-Keun-Sajeon		⑨	32	0.1%
		Only Same with Tongil-Bu		⑩	72	0.1%
	No Mapping			⑪	29819	49.1%
	TOTAL				60674	100.0%
Tongil-Bu	Duplicate with PyoJun			⑫	622	64.5%
	Duplicate with Urimal-Keun-Sajeon			⑬	87	9.0%
	No Mapping			⑭	256	26.5%
	TOTAL				965	100.0%
TOTAL North Korea Word without duplication					61017	-

3-3 사전 정보에 기반한 남북한 어휘 대응

추출된 각 사전을 사용하여 각 북한 어휘에 대한 남한 어휘의 대응 관계를 추출하였다. Table 3은 추출 결과를 요약한 것이다. 표에서 PyoJun은 <표준>의 60,674개의 어휘에 대한 대응 결과(①~⑪)를, Tongil-Bu는 <통일부>의 어휘에 대한 대응 결과(⑫~⑭)를 보인다.

아래에서는 1) <표준>의 북한 단어에 대응 남한어가 정의된 경우, 2) 대응 남한어가 정의되지 않은 경우, 3) <통일부>의 단어비교 각각으로 나누어 분석한다.

1) <표준국어대사전> 북한어에 대응 남한어가 정의된 경우

Table 3에서 ①~⑤는 <표준>의 북한어가 대응되는 남한어를 가지는 경우이다. Fig.1의 '가까운갈래'에서는 '근친'을, '게사니'에서는 '거위'를 대응되는 남한 어휘로 뽑는 예에 해당한다.

표에서 ①~③은 대응된 남한어가 <우리말>(Urimal-Keun-Sajeon)에 존재하는 경우다. [12]의 '숙지'(熟知/宿地)의 예에서처럼 다의어나 동형이의어의 경우 남한어 단어 대응에 의미적 오류가 발생할 수 있다. Fig.1의 '가까운갈래'에 대한 '근친(近親)'에서처럼, 대응된 남한 어휘에 한자 표현이 있으면 <표준>과 <우리말>의 한자 표현까지 같은지의 여부를 검사하여 대응 오류를 방지할 수 있다. Table 3의 ①은 대응 남한어가 동일한 한자 표기로 <우리말>에 등재된 경우를, ②는 한자 표기가 일치되지 않는 경우를, ③은 한자 표기를 가지지 않는 경우이다. ④는 대응된 남한어가 <우리말>에는 없으나 통일부 사전의 대응 남한어와 동일한 경우를, ⑤는 대응된 우리말이 <우리말>과 <통일부> 사전 모두에 존재하지 않은 경우를 나타낸다.

Table 3에서 <표준>의 북한 표제어 중 13.4%가 동일한 한자 표현을 가진 <우리말>의 표제어에 대응(①)되었고, 다른 한자어 표현을 가진 경우(②)가 0.3%, 한자어가 없는 경우(③)가 12.6%로 나타났다.

Table 4는 각 경우에 대해 임의로 추출한 결과에 대한 수동 분석 결과를 보인다. 표에서 Comprehensible(이해가능) 열은 북한어가 순우리말을 사용하는 경우로, ①의 북한어 '등불끄기'(㊟ 소등)가 해당된다.

표 4. 표 3 ①~⑤의 남북한 어휘 대응 수동 분석 결과
Table 4. Analysis of Word Mapping of ①~⑤ inTable 3

	Correct Word Mapping of North-South			Incorrect Mapping
	Compre-hensible	Different Notation	Different Word	
①	45.83%	29.17%	25.00%	-
②	-	-	30.77%	69.23%
③	36.67%	23.33%	40.00%	-
④	50.00%	40.00%	10.00%	-
⑤	36.84%	26.32%	36.84%	-

Different Notation(음운현상과 표기법의 차이) 열은 ①의 '라한(羅漢)'(남 나한), ③의 '뼈쪼각'(남 '뼛조각')이 해당된다. Different Word(다른 단어)열은 ②의 '강작약'(남 백작약), ③의 '장달음'(남 줄달음)이 해당한다. ②의 <표준>에 북한어에 대응되는 남한어가 있으나 그 한자 표현이 <우리말>에 일치하는 것이 없는 경우에서는 '탑형'의 한자로 <표준>에는 '塔型'이 <우리말>에는 '塔形'으로 표기된 경우와, '백작약'의 한자어가 <표준>에는 '白芍藥'이지만 <우리말>에는 '白灼藥'으로 표기되는 등 일치시킬 수 없는 경우가 30.77%에 해당하였으며, 69.23%는 다른 의미의 단어가 대응된 경우였다. 그 예로는 젖제화(남 유화(乳化))에 대해 <우리말>에는 柳花, 榴花, 類化, 油畵, 遺畵 등이 '유화'의 각 의미의 한자로 등재되어 대응시킬 수 없는 경우에 해당하였다. 하지만 ②의 대부분의 경우는 '젖제화'와 유사하게 '작은혀'(남 소설(小舌)) , '알쓸이철'(남 산란기(産卵期))와 같이 한글표현을 사용한 것으로, 남한어 사용자도 문맥 내에서는 의미를 이해할 수 있는 경우에 해당하였다.

Table 3의 ④는 <표준>의 북한어에 대응된 남한어가 <우리말>에는 없는 경우이다. Table 4의 분석에서는 한글표현을 사용한 경우('물스키'(남 수상스키)), 음운현상의 차이('리산가족'(남 이산가족))가 대부분으로 나타났으며, 다른 표현('타프춤'(남 탭댄스))의 경우도 일부 있었다. 대부분이 복합어를 표제어로 취급하여 <우리말>의 단어에 대응되지 않은 것으로 나타났다.

Table 3의 ⑤는 <표준>의 북한어와 대응된 남한어가 <통일부>에도 나타나지 않고 대응된 남한어가 <우리말>에도 등재되지 않은 경우로 그 비율은 14.9%였다. Table 4의 분석에서는 '빗맞추다'(남 '빗맞히다'), '거리감'(남 '원근감'), '막농사군'(남 '막농군')과 같이 음운 현상이나 한글표현, 다른 표현으로 <우리말>에 등재되지 않아도 충분히 이해가능한 우리말인 것으로 분석되었다.

2) <표준국어대사전> 북한어에 대응 남한어가 없는 경우

Table 3의 ⑥~⑧은 <표준>에는 북한 단어에 대응되는 남한 단어가 없지만, 북한 단어의 형태와 동일한 단어가 <우리말>에 표제어로 등재된 경우다. 이를 한자까지 동일한 경우(⑥), 한자가 다른 경우 (⑦), 한자 표기가 없는 순우리말(⑧)으로 나눈다.

Table 5는 Table 3의 ⑥~⑧의 각 경우에 대한 수동 분석 결과를 보인다. 전체에서 53.33%는 북한어가 남한어와 동일한 경우(Same Word)에 해당하였으며, 13.33%는 북한어가 <우리말>에서도 북한 지역의 방언으로 명시된 경우(Dialect)였다.

표 5. 표 3 ⑥~⑧의 남북한 어휘 대응 수동 분석 결과
Table 5. Analysis of Word Mapping of ⑥~⑧ inTable 3

	Correct Word Mappingof North-South		Incorrect Mapping
	Same Word	Dialect	
⑥	71.43%		28.57%
⑦			100.00%
⑧	45.00%	20.00%	35.00%
TOTAL	53.33%	13.33%	33.33%

표 6. 표 3 ⑨~⑩의 남북한 어휘 대응 수동 분석 결과
Table 6. Analysis of Word Mapping of ⑨~⑩ inTable 3

	Correct Word Mapping of North-South		
	Comprehensible	Different Notation	Different Word
⑨		90.91%	9.09%
⑩	40.00%		60.00%

표 7. 표 3 ⑪의 남북한 어휘 대응 수동 분석 결과
Table 7. Analysis of Word Mapping of ⑪ in Table 3

	Incorrect Mapping of North-South	
	Comprehensible	Different Word
⑪	48.33%	51.67%

분석 결과에서는 북한어가 남한에서는 다른 의미로 쓰이는 경우가 33.33%에 해당하였다. '수표(手票)'의 경우 동일한 한자 표기를 가짐에도 불구하고 북한에서는 'sign'의 의미로 사용되었고, 동사 '미우다'가 북한에서는 '냉대하다'의 의미로 쓰이지만 <우리말>에서는 '메우다'의 경상도 방언으로 등재된 경우에 해당하였다. 한자 표기가 다른 ⑦의 경우 대부분 다른 의미인 것으로 나타났다. 북한어 '음질(陰質)'은 우리말큰사전에서는 陰疾, 音質으로만 사용되었다. 하지만, <로동신문>과 <우리민족끼리>의 북한 문서에서도 음질이 모두 音質의 의미로 사용되어 표준국어대사전에 기반한 북한어 연구의 한계가 있음을 알 수 있었다.

Table 3의 ⑨와 ⑩는 <표준>의 북한어에 대응되는 남한 단어가 없는 경우로, ⑨는 <표준>의 북한 단어와 형태는 다르지만 한자 표기는 동일한 남한어가 존재하는 경우를, ⑩은 한글이나 한자 표현이 모두 다르지만 <통일부>에 대응 남한어가 있는 경우이다.

해당 경우가 100개 미만으로 많지 않았다. Table 6은 Table 3의 ⑨와 ⑩에 대한 분석 결과를 보인다. ⑨는 예측한 대로 대부분 두음법칙 등에 의한 음운 현상으로 나타났으며, '均徵'을 <표준>의 북한어에서는 '균징'으로, 우리말큰사전에서는 '균치'로 나타낸 경우를 포함한다. ⑩에서는 이해 가능(Comprehensible)('튀긴고기떡'(남 어묵)), 다른 단어(Different Word)('증견자'(남 '증인')) 등의 경우로 나타났다.

Table 7은 Table 3의 ⑪의 <표준>의 북한어에 대응되는 남한어를 자동으로 추출하지 못한 경우에 대한 분석을 보인다. 결과에서는 '콧구멍이 위로 뚫려 보기 흉하게 생긴 코'를 '천장코'로 표현한 것과 같이 우리나라에서도 충분히 이해할 수 있는 경우(Comprehensible)가 48.33%, '비기게 된 씨름'을 '빅씨름'으로 표현하는 완전히 다른 단어 사용(Different Word)가 51.67%로 분석되었다.

3) 통일부 북한정보포털의 남북한 언어 대응

Table 8은 Table 3의 <통일부>의 북한어-남한어 대응 ⑫~⑭에 대한 분석 결과를 보인다. ⑫는 <통일부>의 북한어가 <표준>에 북한어로 등재된 경우에 해당하며 ①~⑧과 ⑩의 경우와 중복된다.

표 8. 표 3 ⑫~⑭의 남북한 어휘 대응 수동 분석 결과
Table 8. Analysis of Word Mapping of ⑥~⑧ inTable 3

	Correct Word Mapping of North-South					Incorrect Word
	Same Word	Compre-hensible	Different Notation	Dialect	Different Word	
⑫	8.16%	51.02%	15.31%	4.08%	18.37%	3.06%
⑬		40.00%		30.00%	30.00%	
⑭		80.00%	20.00%			

⑬은 <통일부>의 북한어가 <표준>의 북한어로는 등재되지 않았으나 그 형태가 <우리말>에 표제어로 등록된 경우이다. 이 중 40%는 이해가능(Comprehensible)한 경우(‘색날기’(남 탈색)), 30%는 북한 방언(Dialect)으로 <우리말>에 표기된 경우, 20%는 다른 단어(Different Word)를 쓴 경우(‘팔팔아’(남 앵무새)나 ‘탈등’(남 연등))에 해당하였다. ⑭는 이외의 경우로 이해 가능한 경우(‘비닐온실’(남 비닐하우스))나 외래어에 대한 음운현상의 차이(‘에꽈도르’(남 ‘에콰도르’)로 대응된 경우에 해당한다.

4) 대응실패의 경우에 대한 의미 가능 여부 분석

Table 4~Table 8의 분석은 <표준>의 북한어에 대한 <우리말>과 <통일부>를 이용하여 자동으로 남한어를 대응시킨 경우, 대응된 남한어가 북한어와 같은 의미에 해당하는지에 대한 평가이다. 평가 결과에서 이해 가능(Comprehensible)하거나 음운현상이나 표기법 차이(Different Notation)에 해당하는 경우는 남한어 사용자도 북한 문장을 충분히 이해할 수 있다. 이에 반해 북한어가 방언(Dialect)이거나 완전히 새로운 단어(Different Word)인 경우, 남한어와 의미가 다른 경우(Different Sense)는 그 의미를 이해하기 어렵다.

Table 9는 북한어를 남한어 사용자가 이해가능한가의 입장으로 평가한 결과를 보인다. <표준>의 북한어 중 45.25%, <통일부>의 북한어 중 30%, 전체적으로 43.98%의 단어가 남한어 사용자에게는 어떤 의미인지 이해하기 어려운 것으로 분석되었다. 하지만, 이 수치는 사전에 등재된 단어 기준이므로 현실에서 사용되는 단어에서의 비율과는 차이가 있을 것으로 예상된다.

표 9. 북한어를 이해할 수 있는지 여부에 대한 수동 분석 결과
Table 9. Analysis of Understandability of North Korean words as South Korean natives

	Understandable	Misunderstandable
PyoJun	54.75%	45.25%
North-Portal	70.00%	30.00%
TOTAL	56.02%	43.98%

Ⅳ. 남북한 뉴스 매체에서의 어휘 대응

본 장에서는 남북한 뉴스 매체에서 추출한 남북한 어휘의 대응에 대한 분석을 통해, 실제 사용되는 언어 현상에서의 어휘 대응에 대해서 살펴본다.

4-1 남북한 매체 내에서의 기사 자동 수집과 분석

실제 사용되는 언어 표현에서의 남북한 단어 대응을 비교하고 분석하기 위해 남북한 매체로부터 기사들을 자동 수집하였다[2]. 북한 매체에서의 자료 수집은 <로동신문>(이하 <로동>)[6]과 <우리민족끼리>(이하 <민족>)[7]의 온라인 사이트를 대상으로 한다. <로동>은 온라인에 최초 기사가 게재된 2016년 1월부터 2019년 1월 27일까지, <민족>은 2015년 1월부터 2019년 1월 27일까지의 기사를 수집하였다. 북한 문서의 비교 집합인 남한 문서로는 네이버의 뉴스를 수집하였으며, 북한 매체와 유사한 유형의 기사를 수집하기 위해 정치 분야의 북한 뉴스를 대상으로 수집하였다. 북한 매체와 기간을 맞추기 위해 2015년 1월부터 2019년 1월 27일까지 수집하였다.

수집된 기사는 XML 형식으로 저장하였으며 그 구조는 아래와 같다.

```
<ARTICLE>
<URL> URL </URL>
<CATEGORY> category </CATEGORY>
<DATE> published date </DATE>
<TITLE> news title </TITLE>
<CONTENTS>
news contents
</CONTENTS>
<WRITER> reporter name </WRITER>
</ARTICLE>
```

Table 10은 각 매체의 기사수와 어절 수를 보인다. 어절 수는 제목과 기사 본문 내용을 대상으로 한다. <로동>(RODONG)은 약 32만개의 기사에 약 천만 어절로 구성되어 있었고, <민족>(UriMinjok)은 약 18만개의 기사에 약 6백만 어절로 구성되어 있다. 네이버 뉴스는 여러 언론사의 기사를 포함하여 북한 매체에 비해 기사수와 어절 수 모두 큰 값을 나타났다.

표 10. 수집된 남북한 매체 기사의 수, 어절수
Table 10. Statistics of Extracted Norh/South News Articles

	Rodong	Minjok	Naver
Article	32,408	18,066	178,008
Eojeol	10,842,712	5,992,295	20,460,849

2) 국내에서는 북한 매체가 접근되지 않아 해외에서 자료 수집을 시행하였으며, 학술 목적의 자료 수집 및 사용, 출판에 대하여 통일부를 통해 문제가 없음을 확인받았다.

얻어진 남북한 매체의 어휘 대응을 위해 형태소 분석을 통해 내용어를 추출하였다. 형태소 분석에는 KoNLPy[17]에서 제공하는 hannanum, komoran, kkma의 세 개의 형태소 분석기를 사용하였다. Table 11은 각 매체에 대해 각 형태소 분석기를 적용하여 얻은 명사, 동사, 형용사, 관형사, 부사, 감탄사의 총 개수를 보인다. 북한 문서의 경우 형태소 분석 결과 미등록어로 분석되는 비율이 높아 미등록어도 포함한다. 표 11에서 Word(Uniq)는 중복되어 나타나는 단어를 1개로 취급한 결과다.

북한 문서에 대한 형태소 분석을 위해서 고유한 문서 특징과 남한어와는 다른 언어규범으로 형태소 분석에 추가적인 처리가 필요했다. 전처리로는 북한 문서에서 전각으로 표현된 숫자나 도량형 표현을 국내 형태소 분석기에 적합하게 변환하는 작업 등을 수행하였다. 북한 문장에서 자주 발생하는 '되였다', '하시였다' 등의 활용[9]을 처리하기 위해 각 형태소 분석기의 사전을 수정하였다.

표 11. 남북한 기사의 각 형태소 분석기 별 내용어 어휘 수
Table 11. Word statistics of Norh/South News Articles with various POS tagger and user dictionary

			Rodong	Minjok	Naver
Word	hannanum	w	8,817,029	4,922,073	17,259,560
		wo	8,877,334	4,941,050	17,266,569
	komoran	w	13,987,175	7,619,463	22,742,934
		wo	14,162,970	7,637,605	22,357,993
	kkma	w	14,967,459	8,280,532	25,806,855
		wo	15,408,583	8,448,361	25,862,909
Word (Uniq)	hannanum	w	379,456	269,886	611,756
		wo	395,242	280,214	609,397
	komoran	w	57,470	49,676	77,845
		wo	63,498	53,473	80,058
	kkma	w	57,236	50,015	69,445
		wo	62,068	54,095	70,261

또한, 3장에서 추출한 표준국어대사전에 등재된 북한어 중에서 우리말큰사전에 포함되지 않은 5만여 단어를 각 형태소 분석기의 사용자 사전에 추가하였다. hananum의 경우 사용자 사전의 용량 한계가 있어 기존 사전 단어로 분할되거나 긴 단어를 제외한 2만여 단어만 사용자 사전에 추가하였다. komoran과 kkma는 5만여 단어를 모두 사용자 사전에 추가하였다.

Table 11에서 w행은 사용자 사전을 포함하여 형태소 분석을 한 결과, wo행은 사용자 사전을 사용하지 않은 결과이다. 결과에서는 hannanum 형태소 분석기가 다른 분석기에 비해 전체 내용어 수(Word)는 절반 가량이지만 유일단어수(Word(Uniq))수는 5~6배로 나타났다. 분석에서 kkma와 komoran은 '조국통일'을 '조국+통일'으로 분석하지만 hannanum은 한 단어로 분석하는 등의 복합어를 선호하는 경향이 강한 것으로 나타나, 본 연구의 단어 대응에는 적합하지 않다고 판단하였다.

4-2 북한 기사 내 단어의 남한어 대응 분석

실제 사용되는 현실 언어 문서인 북한 기사의 단어의 양상을 분석하기 위해, <로동>과 <민족>의 기사의 내용어를 추출하고 각각이 <표준>의 북한 단어, <우리말>의 표제어, 네이버의 북한 뉴스에 나타나는 단어와의 일치하는 정도를 분석하였다.

Table 12는 각 매체별 단어의 대응 비율을 보인다. ALL은 <로동>(Rodong)의 단어와 <민족>(Minjok)의 단어를 통합하여 계산한 경우 결과를 보인다. 표에서 Word는 전체 단어에 대한 대응 비율을, Word(Uniq)는 중복을 제외한 단일 단어의 대응 비율을 나타낸다.

표 12. 북한 매체별 단어 대응 비율
Table 12. Word Mapping Ratio per North Korean News

			Rodong	Minjok	ALL
Word	ko-moran	Naver News	98.95%	99.18%	99.03%
		Urimal Dic	87.18%	87.56%	87.32%
		Pyojun Dic	21.16%	19.71%	20.65%
		no match	0.36%	0.33%	0.35%
	kkma	Naver News	99.03%	99.23%	99.10%
		Urimal Dic	89.62%	89.15%	89.45%
		Pyojun Dic	17.27%	16.96%	17.16%
		no match	0.23%	0.24%	0.23%
Word (Uniq)	ko-moran	Naver News	64.81%	73.47%	60.84%
		Urimal Dic	50.23%	56.78%	46.54%
		Pyojun Dic	17.72%	14.98%	15.89%
		no match	21.04%	16.09%	25.22%
	kkma	Naver News	64.26%	71.66%	59.90%
		Urimal Dic	61.47%	64.76%	57.42%
		Pyojun Dic	14.71%	11.98%	13.22%
		no match	16.05%	14.16%	20.43%

대응에서는 사용자 사전을 이용한 komoran과 kkma의 형태소 분석만 사용하였다. 각 행에서 Pyojun Dic는 각 매체의 단어가 <표준>의 북한어 표제어에 해당하는 비율, Urimal Dic은 <우리말>의 남한어 표제어에 해당하는 비율, Naver News는 네이버 뉴스의 단어에 해당하는 비율, no match는 세 곳 모두에 발생하지 않는 단어의 비율을 보인다.

매체별 결과에서는 북한의 공식 대표 일간신문인 <로동>에 비해 인터넷 선전과 선동 중심인 <민족>이 일상적인 단어와 남한 관련 내용을 많이 포함하여 <우리말>과 네이버뉴스와의 일치도가 더 높은 것으로 나타났다.

각 사전과 남한 뉴스의 대응비율에서는 <표준>의 북한어 사전에 등재된 단어와의 대응 비율이 17~20.%인 것에 비하여, <우리말>은 87~90%, 네이버뉴스는 99%의 대응 비율을 보였다. 이로부터 <표준>의 북한어가 남한어와 차이나는 단어 중심으로 작성되었다는 사실과 함께, 실제 문서에서는 남북한 언어가 큰 유사성을 가진다는 점을 확인할 수 있었다.

형태소 분석기별 결과에서는 ALL에 대해서 komoran 형태소 분석기에서는 17,754개의 단어가 대응을 찾지 못했고, kkma 형태소 분석기에서는 14,436개의 단어가 대응을 찾지 못했다. 두 형태소 분석기 모두에서 대응되지 않은 단어 중 73% 가량이 미등록어에 해당하였다. Table 13은 각 형태소 분석기별 미대응 단어의 빈도 상위 20개의 목록을 보인다. 결과에서는, "줴치/동사"가 두 형태소 분석기의 사용자 사전에 모두 등록되어 있음에도 불구하고, "줴쳐댔다"를 kkma에서는 "줴치/VV+어/ECS 대/VXV+었/EPT+다/EFN"으로 분석하는 것에 비해 komoran는 미등록어로 분석하는 등의 결과를 보여, 북한어에 대한 형태소 분석에서는 kkma가 우수한 결과를 내는 것으로 판단되었다. 하지만, kkma의 형태소 분석에서도 "알다싶이"(㊟ 알다시피), "지니였던"(㊟ 지니었던), "놓으시였던"(㊟ 놓으셨던), "말하군 하였다"(㊟ 말하곤 했다) 등의 분석에서 오류가 발생하여 추가 연구나 남북 협력이 필요한 것으로 나타났다.

표 13. 각 형태소 분석기별 빈도 최상위 20위 북한 내용어 목록
Table 13. Top 20 frequent unmatched North Korean contents words with each POS tagger

	kkma	komoran
1	적폐청산/UN *jeog-pe-cheong-san*	무들/NNP *mu-deul*
2	싶이/UN *sipi*	<ㅌ./NA *<t.*
3	스꼬/UN *seu-kko*	101/NNP
4	지니이/VV *ji-ni-i*	오끼나와/NNP *Okinawa*
5	놓으/UN *noh-eu*	줴쳐댔다./NA *jwe-chyeo-daess-da.*
6	하군/UN *ha-gun*	어머님께/NNP *eo-meo-nim-kke*
7	포옹력/NNG *po-ong-lyeog*	숭엄히/NA *sung-eom-hi*
8	웨쳤/UN *we-chyeoss*	두산천/NNP *du-san-cheon*
9	도꾜/UN *Tokyo*	민주꽁고주체사상연구/NA *min-ju-kkong-go-ju-che-sa-sang-yeon-gu*
10	찾으시였/UN *chaj-eu-si-yeoss*	부럼없어라>의/NA *bu-leom-eobs-eo-la>ui*
11	베아링/NNG *be-a-ling*	ㄷ>의/NA *d>ui*
12	웨치면/UN *we-chi-myeon*	웨쳤다./NA *we-chyeoss-da.*
13	꾸드스/UN *kku-deu-seu*	적폐청산을/NA *jeog-pe-cheong-san-eul*
14	웨친/UN *we-chin*	웨치면서/NA *we-chi-myeon-seo*
15	웃으/UN *us-eu*	적페를/NA *jeog-pe-leul*
16	갈론/NNG *gal-lon*	웨침이/NA *we-chim-i*
17	때려부시/VV *ttae-lyeo-bu-si*	차광수/NNP *cha-gwang-su*
18	일에/UN *il-e*	발아래/NNG *bal-a-lae*
19	발아래/NNG *bal-a-lae*	재일본조선민주녀성동맹/NNP *jae-il-bon-jo-seon-min-ju-nyeo-seong-dong-maeng*
20	보살피심속/UN *bo-sal-pi-sim-sog*	<<ㅌ./NA *<<t*

두 사전이나 네이버 뉴스에서 찾을 수 없는 단어(no match)는 중복 허용(Word)의 경우에는 0.3% 수준인 것에 비해, 유일 단어(Word(Uniq))에서는 20% 수준으로 나타나, no match의 단어들이 희소성이 높은 단어로 나타났다. 결과 분석에서는 kkma의 ALL에서 유일 단어 중 0.01%에 해당하는 빈도 상위 10개 단어('하다', '있다', '위하다', '되다', '대하다', '것', '수', '우리', '인민', '조선')가 전체 단어의 7.37%를 차지하였으며, 빈도 상위 1%의 707개의 단어가 전체 단어의 56.19%를 차지하여, 북한 매체에 대한 우리말 큰사전과 네이버 뉴스와의 높은 단어 대응이 이러한 고빈도 단어에 의한 것임을 알 수 있었다.

Ⅳ. 결 론

본 논문에서는 다양한 사전과 뉴스 매체를 사용하여 남북한 어휘 대응을 분석하였다. 사전으로는 <표준국어대사전>의 북한어, <우리말큰사전>의 표제어, 통일부 북한정보포털의 언어 비교를 사용한 분석에서는 사전의 북한어 중 약 44%는 남한어 사용자가 이해하기 어려운 단어로 나타났다. 뉴스 매체에 대한 분석에서는 <로동신문>과 <우리민족끼리>, 네이버 북한 뉴스를 사용하였으며, 북한 매체에서 사용되는 단어 중 99%의 단어가 네이버뉴스에 발생하는 것으로 분석되어 남북한 언어의 유사성을 확인할 수 있었다. 유일단어에 대한 분석에서는 40% 이상의 북한 매체의 단어가 남한 매체나 사전에 등장하지 않아 남북한 교류를 위해 해결해야 하는 대상으로 나타났다. 추가적으로 우리나라에서 개발된 형태소 분석기 hannanum, komoran, kkma 중에서는 kkma가 가장 좋은 분석 결과를 나타냈으나, 북한어 특징을 반영하기 위해서는 미흡한 것으로 나타났다.

향후 연구로는 남북한 언어 규범의 음운현상 차이 등으로 인한 어휘 차이를 자동으로 대응하고, 수집된 대용량 남북한 문서를 기반으로 미대응 단어를 자동으로 대응하는 등의 연구를 진행할 예정이다.

감사의 글

이 연구는 금오공과대학교 학술연구비로 지원되었음 (2019-104-103)

참고문헌

[1] Gyeoremal-keunsajeon, Available: http://www.gyeoremal. or.kr/
[2] Pyojun-Gugeo-Daesajeon, National Institute of Korean Language, Available: https://stdict.korean.go.kr/
[3] Comparison between South and North Korean Languages,

MINISTRY OF UNIFICATION, https://nkinfo.unikorea.go.kr/nkp/term/skNkLangCompare.do
[4] Urimal-Keun-Sajeon, Hangeul-Haghoe, Available: http://semanticweb.kaist.ac.kr/service/urimal/
[5] News Articles of North Korea, NAVER, Available: https://news.naver.com/main/list.nhn?mode=LS2D&mid=shm&sid1=100&sid2=268
[6] Rodong News, Available: www.rodong.rep.kp
[7] Uriminzokkiri, Available: www.uriminzokkiri.com/
[8] Jeon, Sutae, “Language Differences between South and North Korea and Scheme to Overcome”, Korean Language Culture School, No. 5, pp.69-96, National Academy of the Korean Language, Dec., 2000.
[9] Kim, Taeggu, “Comparison of Norms of South and Norh Korea and Seeking Unification Plan”, Humanities Research, Vol. 9, pp.27-79, Anyang Univ. Humanities Research Center, Sep., 2001.
[10] Han, Yong-Un, “The way for South and North standard language integration”, Korean thought and culture, No. 40, pp.301-322, Oct., 2007.
[11] Hong, Yun-pyo, "The Direction of Dictionary Compilation in Gyeoremal-keunsajeon(Unabridged and Unified Korean Dictionary)", Journal of Korealex, Vol. 9, pp. 23-52, Apr., 2007.
[12] Han, Seung Kyu, "Problems of the Description of Language of Northern Korean in Korean Dictionary, Pyojun-Gugeo-Daesajeon", Journal of Korealex, Vol. 18, pp.227-254, Oct., 2011.
[13] Han, Yong-Un, “Special Issue: Liberation, 70 years of division and unification: Inter-Korean vocabulary after division”, Forum for Korean Contemporary History. Vol. 6, pp52-70, Dec., 2015.
[14] Kwon, Jaeil, “Unification of Korean and North Korean vocabularies”, Sae-Gug-eo Saenghwal, Vol. 25, No. 4, Dec., 2015.
[15] Chang H. Lee1, Kyungil Kim and Jongmin Park, “Preliminary Analysis of Language Styles between South and North Korean Broadcastings", Journal of Korea Academy Industrial Cooperation Society, Vol. 11, No. 9, pp.3311-3317, Sep., 2010.
[16] Park, Myeong Su, “A study on the vocabulary contrast between North and South korea using the newspaper corpus”, Ph.D Thesis, Yonsei Univ., Feb., 2019.
[17] KoNLPy:Python Korean NLP, https://konlpy.org/ko/latest/

이현아(Hyunah Lee)

1996년 : 연세대학교 컴퓨터과학과 (학사)
1998년 : KAIST 전산학과 (석사)
2004년 : KAIST 전산학과 (박사)
2000년~2004년 : ㈜다음소프트 언어처리연구소
2004년~현 재 : 금오공과대학교 컴퓨터소프트웨어공학과 교수

※관심분야 : 자연언어처리, 텍스트데이터마이닝, 정보검색 등

송지혜(Ji-hye Song)

1997년 : 경북대학교 국어국문학과 (학사)
1999년 : 경북대학교 국어국문학과 (문학석사)
2009년 : 경북대학교 국어국문학과 (문학박사)
2005년~2011년 : 경북대학교 한국어문화원 책임연구원
2011년~2014년 : 경북대학교 기초교육원 초빙교수
2014년~현 재 : 금오공과대학교 교양교직과정부 교수

※관심분야 : 국어 어휘론, 국어사, 의미론 등