DOI | https://DOI.org/10.21481/jbunka..82.201908.005.

텍스트 마이닝을 활용한 일본어능력시험 내용 연구
— JLPT N3 문자 · 어휘를 중심으로—

이 유 희*
(e-mail : manyyh@hanmail.net)

<목 차>

キーワード：テキストマイニング(Text Mining), 自然言語処理(Natural Language Processing), 日本語能力試験3級(JLPT N3), 頻度数(Frequency), 形態素(Morpheme)

1. 들어가기

2012년 세계 경제 포럼에서 빅데이터 기술이 10대 기술 중 첫 번째로 선정된 이래, 4차산업 혁명의 미래시대 패러다임으로 빅데이터가 화두가 되기 시작하였다. 디지털 경제가 확산되고, 빅데이터 환경이 도래되면서 빅데이터 핵심기술 및 인재 양성은 전 세계적인 이슈로 부상되었다.

인공지능에 의한 혁신적인 사회변화의 기대와 기회 속에서 그동안 빅데이터 분석 연구는 이공계열을 중심으로 많은 연구소들이 세워지고, 다양한 연구 활동이 진행되어 성과를 창출해 왔다. 이러한 연구들이 점차로 인문, 사회과학 분야로 확산됨에 따라 인문사회 분야에 있어서도 빅데이터 및 텍스트 마이닝(Text Mining)을 활용한 학제간의 연구가 지속적으로 실시되고 있다. 그러나 아직까지 일본어를 비롯한 언어학 분야에 있어서는 이러한 기법을 활용한 텍

* 대전대학교, 강사, 일본언어문화

스트 분석과 교육 연구는 거의 시도되고 있지 않는 실정이다. 이러한 문제의식과 더불어 인공시대에 있어서 테크놀로지의 인문학적 활용의 필요성에 입각하여 본 연구에서는 빅데이터-자연어처리(NLP:Natural Language Processing)연구에 포커스를 두었다. 본 연구는 인공지능시대에 있어 빅데이터-자연어처리를 어떻게 일본어 교육 연구에 활용할 수 있고, 응용 가능한지를 목표로 두고자 한다. 이를 위해 본 연구에 있어서는 다음과 같은 목적을 두고자 한다. 첫째, 텍스트 마이닝 분석을 활용하여 JLPT N3 기출 문자·어휘 전체를 대상으로 빈출 키워드를 추출하여 시험 문제의 패턴과 경향을 분석한다. 둘째, 문제1부터 5까지 주요 상위 빈출 핵심어를 추출하여 각 파트별로 주요 키워드를 알아보고, 이에 대한 결과를 시각적으로 파악해 보도록 하겠다. 셋째, 전체 기출 문자·어휘를 형태소로 분류하고, 품사별 빈도수 및 빈도 비율을 측정하여 효과적인 교육 및 학습 방법을 제시하도록 하겠다.

2. 선행연구 및 연구방법

2.1. 선행연구

국내 텍스트 마이닝의 연구는 경제, 정책, 선거, 미디어, 주가 예측, 문화, 관광, 항공 산업, 국방, 재난, 산업공학 등, 사회 전반에 걸친 다양한 분야에서의 분석이 이루어지고 있다. 이러한 텍스트 마이닝 이슈를 다룬 선행연구는 문헌정보학, 경영학 등의 분야에서 활발히 진행되다가 점차로 관광학 분야 등, 다양한 영역으로 확대, 발전되어 왔다(박자현, 2013; 강범일, 2013; 김현정, 2015; 최정원, 2015; 박진균, 2017). 이와 같은 연구들은 다양한 관점을 통해 유용한 정보를 수집하여 실생활에 활용되고 있으며 미래 과제의 의미 있는 논의를 제기하고 있다.

상기의 분야 외에 텍스트 마이닝은 최근에는 더욱 연구 영역이 확대되어 인문학-교육학 분야에서도 다양한 분석이 시도되고 있다(권충훈, 2018; 조주연, 2018). 기존의 일본어, 영어 등의 언어학 분야의 텍스트 분석에서는 주로 코퍼스 언어학 연구와 이를 위한 프로그램이 활용되어 왔다[1]. 이 중에서 일본어

1) 영어필드에서는 김연주(2016) 등이 코퍼스 AntConc3.4.1w를 사용하여 텍스트 마이닝을 분석을 실시하였다. 일본어 필드에서는 민광준(2013)이 코퍼스를 활용하여 일반 고등학교 일본어 교과서의 가타카나의 실태를 조사하였고, 박선주(2018)는 '추나곤'(中納言) 일본 코퍼스를 활용하여 일본어 한자 어휘를 분석하였다.

능력시험(JLPT)을 대상으로 한 연구는 2010년부터 주를 이루고 있는데, 분석의 내용과 범위가 협소하며 종류와 양도 지극히 적다고 할 수 있다. 이러한 가운데, 鈴木美恵등(2016)은 국제교류기금(2006)『일본어능력시험출제기준』을 활용하여 JLPT의 외래어를 대상으로 '악센트'에 초점을 두어 연구를 실시하였다. 永野亜季(2013)는『공식문제집』(1회분, 2012년)과 국내 다락원, 시사, 동양문고의 모의시험[2] N1의 어휘를 조사하여 일본 국어 교육 어휘-「교과서 코퍼스 어휘표」[3]와 비교하고, 한국에서 출판된 문제집과 공식 문제집과의 차이를 분석하였는데, 동 연구는 어휘 자체의 비교보다는 '품사별 비교'에 한정되어 있다. 김유영(2015)은 구 일본어능력시험기출 독해 문제(1급~4급)를 대상으로 일본어 텍스트의 '가독성 레벨'에 초점을 두어 분석을 실시하였다. 조은영(2018)은 JLPT N1의『공식문제집』에 나타나는 부사와 한국에서 출판된 모의문제집 및 어휘표에 나타나는 '부사', JLPT N1 대책용 어휘집에 나타난 '부사'를 비교하여 그 차이점을 분석하였다. 이 외에 이도열(2012)는 JLPT N1과 EJU간 득점에 대해 비교 작업을 통해 '척도 분석'에 포커스를 두어 분석을 실시하였다.

상기와 같이 JLPT의 분석 내용으로는 가독성 연구(교육학), 악센트(음성학), 품사비교, 부사(언어학), 득점 척도 분석 등이 이루어졌다. 연구 대상으로는 N1 공식문제집 1회분, N1 모의문제집, 구 능력시험 기출문제 독해(1급~4급), N1 대책용 어휘집이 주로 활용되었다. 이처럼 위에서 언급한 바와 같이 능력시험 연구는 분석 대상, 내용, 범위, 방법 등이 모두 한정적이라 할 수 있다. 이공계 및 인문학 분야에서 R, Python 등과 같은 인공지능 프로그램을 활용한 텍스트 마이닝 기법이 활발히 활용되고 있는데 반해 일본어 등의 언어학 분야에서는 거의 시도되고 있지 않다. 이러한 근거에 입각하여 새로운 방법론과 어프로치를 바탕으로 N3부터 향후 N2, 1 단계별로 진행되는 연구를 통해 새로운 결과를 도출하기 위한 시도가 필요하다 할 수 있다. 텍스트 마이닝 기법이 많은 분야에서 활용되어 현안에 대처하고, 미래 지향적 문제 해결 방안을 모색해 오고 있는 만큼, 본 논문을 통해 이러한 기법을 활용해 봄으로써 연구의 폭을 넓히고, 다양한 지견을 확대할 수 있을 것이라 기대할 수 있을 것이다.

---

2) 이치우(2010)『新일본어 능력시험 N1 한권으로 끝내기-부록모의시험1회』, 신JLPT연구모임(2010)『신JLPT 신일본어능력시험 한권으로 합격하기 N1-부록모의시험1회』, 마츠모토세츠코・가네니와쿠미코・오기와라치카코(2010)『新일본어능력시험 실전 모의고사 N1-모의시험1회』

3) 국립국어연구소에서 제작한 어휘표. 2005년에 초, 중, 고등학교 전학년, 전과목의 교과서 1종씩을 채택하였고, Mecab엔진 등을 이용한 형태소분석 사전 UniDic을 사용.
(https://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html)

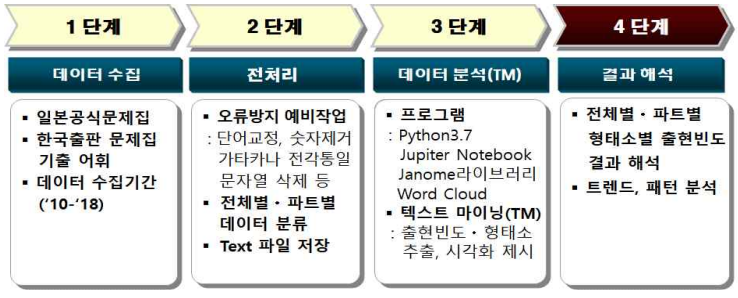
2.2. 연구 방법

2.2.1. 텍스트 마이닝 분석

빅데이터는 대용량의 정형 데이터(구조화된 데이터, Structured Data) 및 비정형 데이터(비구조화된 데이터, Unstructured Data) 등으로 분류된다. 정형 데이터(SD)는 정형화 된 수치 데이터를 의미하고, 비정형 데이터(UD)는 수치 데이터 이외의 텍스트나 영상, 이미지, 소리 등을 의미한다. 빅데이터 분석에 중요한 도구는 데이터 마이닝(데이터 분석)이다. 데이터 마이닝은 방대한 양의 데이터 분석을 통해 의미 있는 패턴과 규칙 (Berry and Linoff 1997;2004), 트렌드 및 데이터간의 상호관계를 추출하는 기술이다(Gartner Group,1994 ; Gartner.com; Fayyad 1996). 비정형 데이터 분석으로는 텍스트 마이닝, 소셜 네크워크 분석, 오피니언 마이닝, 군집분석(Cluster Analysis) 등이 있다. 텍스트로 이루어진 데이터를 분석하는 방법이 바로 텍스트 마이닝인데, Feldman과 Dagan(1995)은 텍스트 마이닝 정의와 함께 그 과정으로 데이터의 수집-형태소 분석-특징 추출-트렌드 분석 등을 제시하였다.

본 연구에서는 이와 같은 텍스트 마이닝 분석을 활용하여 데이터 수집, 전처리, 텍스트분석-빈도분석 · 형태소 분석, 결과해석 등 4단계로 설정하였다. 구체적으로는 하기와 같은 절차와 진행 방법(그림2-1 참조), 사용 프로그램, 연구 범위 및 대상을 구분하여 분석을 실시하였다.

<그림 2-1> 연구절차



2.2.2. 데이터 수집

본 연구에 있어서는 N3의 문자 · 어휘를 중심으로 개정 이후 출판된 공식 문제집(国際交流基金 · 日本国際教育支援協会, 2012; 2018)과 기출 문제를 반영하고 있는 한국 내 출판된 문자 · 어휘 문제집(이치우, 2018; JLPT연구모임, 2018)을 분석 대상으로 설정하였다. N3 시험은 상반기(7월), 하반기(12월)를 포

함하여 연 2회로 실시되고 있는데, 2010년부터 2018년까지 총 18건의 자료를 수집하였다.

2.2.3. 데이터 전처리

첫 번째로 총 18건의 자료 중, 기출 문자·어휘를 수집하여 전체별, 파트별로 각각 분류하여 데이터 파일로 저장, 관리하였다. 두 번째로 전처리 과정은 텍스트의 체계적 분석을 위한 오류방지 예비 작업으로 교정 및 제거 작업이 필요한 과정이다. Janome로 형태소 분류 시에, 단어 분리와 루비 및 기호 등 제거 작업을 거치게 되므로 전처리에서는 불필요한 문자열을 먼저 삭제하고, 오탈자 확인과 연도·숫자, 기호 및 한국어, 그리고 형태소 분류 시 불필요한 'する' 동사 등의 단어를 제거하였다. 세번째로는 이것을 Python에서 불러와 사용할 수 있도록 파트별, 전체별 text파일로 나누어 별도로 저장하였다.

2.2.4. 데이터 분석

본 연구에서는 키워드 빈도수 및 형태소, 시각화 분석을 위해 Python3.7 프로그램과 Jupiter Notebook 도구, 분석툴·라이브러리 Janome(부수적으로 Mecab), Word Cloud를 활용하였다. 2014년 이후부터 SPSS프로그램을 대신하여 R이 통계용 언어로서 통계학자들을 중심으로 논문에서 사용빈도가 높아지고 있는 추세인데, 본 논문에서는 전 세계 범용적인 멀티프로그래밍 언어인 오픈소스 Python[4]을 사용하였다. Python의 실행 환경은 Eclipse, Visual Studio, PyCharm 등과 같이 여러 가지가 존재하지만, Jupiter Notebook툴은 프로그래밍 코드를 넣어 실행하기 편리하고, 결과 보존이나 자료작성, 공유 등 학습자들의 효율성을 높여 주는데 유용하기에 본 논문에 활용을 시도해 보았다. 인스톨은 Python언어와 이 툴이 모두 사용가능한 Anaconda3을 이용하여 실행하였다.

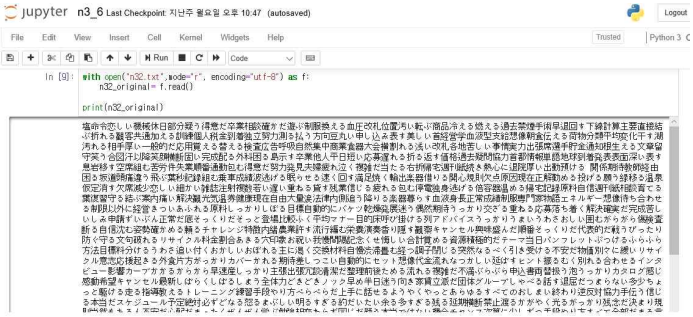
일본어 텍스트의 빈도수와 형태소 구분을 위해 Python의 형태소 분석엔진 Janome(Mecab)를 사용하였다. 두 엔진 모두 해석 결과의 정도가 좋지만, 형태소 분석의 속도에 있어서는 Janome쪽이 10배정도 느리다는 단점이 있으므로,

4) 이것은 C언어로 구현된 라이브러리를 그대로 간단히 사용 가능하고, Java언어의 특징을 모두 지니고 있으며, 짧은 코드로 읽고 쓰기 쉽고, 간결한 문법에다 손쉬운 데이터 파싱으로 알고리즘 개발에 사용하기 편리한 동시에 방대한 라이브러리를 제공하고 있는 장점이 있다. 또 web 프로그래밍, 수치연산 프로그래밍, 데이터 베이스 프로그램, 어플리케이션 등, 다양한 분야의 개발이 가능한 특징을 지니고 있다.

대량의 문서를 빠른 속도로 처리하기 위해서는 Mecab쪽을 사용하는 것이 좋다(mocobeta.github.io/janome).

본 연구에서는 우선 Python에서 간단히 pip 코멘드로 Janome를 인스톨하여 사용하였다. 먼저 자연어처리 분석을 위해 저장한 text파일을 Jupyter Notebook으로 아래와 같이 불러오도록 한다.

<그림2-2> 텍스트 데이터 임포트



불러 온 데이터를 pickle로 저장하여 사용할 수 있는데, pickle 모듈을 이용하면 모든 데이터 객체를 그대로 유지한 채 저장하고 읽기가 가능하다. 이 때 dump 메소드를 이용하여 문자열이나 값을 한꺼번에 전달할 수 있다. 소설이나 문장이 있는 텍스트의 경우에는 루비나 구두점, 기호 등을 삭제하고, seperator를 사용하여 문장을 list에 넣어 분할하고, 이 list를 pickle로 바이너리 파일에 저장한다.

다음은 빈출 키워드 추출 및 형태소 분석을 위해 pip으로 Janome를 인스톨한 이후, Janome tokenizer를 임포트 하고, 저장한 pickle의 바이너리 파일을 불러온다. 단어별로 분할하기 위해 'wakati=True'라는 인수를 적용하여 print해 주면 각 단어가 분할되어 나타나게 된다. 이 분할 된 단어들을 활용해 단어 출현 빈도수를 추출할 수 있는데, words라는 빈 list를 만들어 주고, 이 words list에 모든 단어가 저장되도록 이전에 분할한 list를 words list에 결합한다. 그리고 이 words 안의 모든 단어를 카운트하도록 collections를 사용하여 아래와 같이 상위 빈도수를 측정한다.

<그림2-3> 단어 빈도수 측정

단어 빈도수 추출 후, 이러한 빈도 키워드를 데이터 시각화 작업툴인 일본 wordcloud를 이용하여 결과를 시각적으로 나타내 보도록 하였다. 이에 대한 결과는 다음 장에서 자세히 논해 보도록 하겠다.

마지막으로 형태소 분석은 att에 인수를 지정하는 방법으로 가능한데, att=part_of_speech로 입력하고, word count를 출력해 주면 형태소별 측정이 가능하게 된다. 이를 위해서 먼저, Python 라이브러리 collections의 Counter를 이용하여 tokenize 메소드 인수 wakati를 True로 하고, 문자열을 나누어 단어 출현 회수를 카운트 한다. 그 다음 part_of_speech.split를 이용하여 품사별로 간단히 카운트 해 보기로 하겠다. 또 Python에서는 TokenCountFilter의 인수 att=part_of_speech를 사용하여 형태소를 분류하게 되면, 품사1,품사2,품사3 등으로 더 구체적으로 형태소를 세분화하여 추출 할 수 있다. 이에 대한 구체적인 결과에 대해서도 다음 장의 분석 결과 파트에서 다루어 보기로 하겠다.

3. 분석 결과

3.1. 빈도 분석 결과

3.1.1. 전체 결과

전체 기출 문자 · 어휘를 대상으로 출현 빈도를 분석해 본 결과, 핵심 키워드는 131개이며, 중복 출현 횟수는 최대 5회로 나타났다. 상위 5회~2회 출현한 핵심 키워드를 요약 · 정리해 살펴보면 아래 <표3-1>과 같다.

<표 3-1> 빈출 핵심어 분석 (상위 5회~2회, 2010-2018년)

2010-2018

번호	단어	빈도수	품사	번호	단어	빈도수	품사	번호	단어	빈도수	품사	번호	단어	빈도수	품사
1	すぎる	5	동사	34	努力	2	명사	**67**	**しゃべる**	**2**	**동사**	100	少し	2	부사
2	**的**	**4**	**명사/접미사**	35	血液	2	명사	**68**	**話す**	**2**	**동사**	101	さっき	2	부사
3	**想像**	**3**	**명사**	36	分類	2	명사	**69**	**教える**	**2**	**동사**	102	正常(だ)	2	명사(형・동)
4	**期待**	**3**	**명사**	37	平均	2	명사	70	転ぶ	2	동사	103	同じ(だ)	2	연체사(형・동)
5	出張	3	명사	38	検査	2	명사	71	回す	2	동사	**104**	**複雑だ**	**2**	**형・동**
6	規則	3	명사	39	横断	2	명사	72	結ぶ	2	동사	**105**	**盛んだ**	**2**	**형・동**
7	疲れる	3	동사	40	完成	2	명사	73	表す	2	동사	**106**	**上手だ**	**2**	**형・동**
8	断る	3	동사	41	応募	2	명사	74	覚える	2	동사	107	不安だ	2	형・동
9	怒る	3	동사	42	協力	2	명사	75	困る	2	동사	108	清潔だ	2	형・동
10	そっくりだ	3	형・동	43	順番	2	명사	76	組む	2	동사	109	短気だ	2	형・동
11	**得意だ**	**3**	**형・동**	44	通勤	2	명사	77	包む	2	동사	110	大変だ	2	형・동
12	**きつい**	**3**	**형용사**	45	経由	2	명사	78	預ける	2	동사	111	本当だ	2	형・동
13	**うっかり**	**3**	**부사**	46	記録	2	명사	79	逃げる	2	동사	112	うわさ	2	명사(1자)
14	**卒業**	**2**	**명사**	47	成績	2	명사	80	守る	2	동사	113	空	2	명사(1자)
15	**相談**	**2**	**명사**	48	楽器	2	명사	81	信じる	2	동사	114	業	2	명사(1자)
16	**制服**	**2**	**명사**	49	欠点	2	명사	82	落ち着く	2	동사	115	わけ	2	명사(1자)
17	**改札**	**2**	**명사**	50	現在	2	명사	83	確かめる	2	동사	116	ところ	2	명사(1자)
18	**経営**	**2**	**명사**	51	減少	2	명사	84	経つ	2	동사	**117**	**週刊紙**	**2**	**명사(3자)**
19	**帰宅**	**2**	**명사**	52	自信	2	명사	85	引き受ける	2	동사	**118**	**やり方**	**2**	**명사(혼독)**
20	**原料**	**2**	**명사**	53	解決	2	명사	86	かれる	2	동사	119	決まり	2	명사(동사연용형)
21	**制限**	**2**	**명사**	54	内緒	2	명사	87	どなる	2	동사	**120**	**厚い**	**2**	**형용사**
22	**目標**	**2**	**명사**	55	渋滞	2	명사	88	余る	2	동사	**121**	**悲しい**	**2**	**형용사**
23	**発展**	**2**	**명사**	56	整理	2	명사	89	残る	2	동사	122	おかしい(な)	2	형용사(연체사)
24	**指導**	**2**	**명사**	57	予定	2	명사	90	くたびれる	2	동사	123	苦しい	2	형용사
25	**手段**	**2**	**명사**	58	回収	2	명사	91	たまる	2	동사	124	緩い	2	형용사
26	**延期**	**2**	**명사**	59	理由	2	명사	92	そっと	2	부사	125	まぶしい	2	형용사
27	**活動**	**2**	**명사**	60	暗記	2	명사	**93**	**しっかり**	**2**	**부사**	126	キャンセル	2	외래어
28	**絶対(に)**	**2**	**명사(부사)**	61	性格	2	명사	94	がっかり	2	부사	127	リサイクル	2	외래어
29	位置	2	명사	62	募集	2	명사	95	どきどき	2	부사	128	スケジュール	2	외래어
30	過去	2	명사	63	修理	2	명사	96	すべて(の)	2	부사	129	カーブ	2	외래어
31	早退	2	명사	**64**	**疑う**	**2**	**동사**	97	必ず	2	부사	130	身につける	2	관용구
32	共通	2	명사	**65**	**重ねる**	**2**	**동사**	98	約	2	부사	131	明る	2	형용사어간
33	到着	2	명사	**66**	**迷う**	**2**	**동사**	99	だいたい	2	부사				

이와 같은 데이터 테이블 표기 방법 외에, 핵심어를 간단히 편리하게 알아볼 수 있도록 Word Cloud로 시각해 보면 다음과 같이 추출되는 것을 볼 수 있다. Python에서 Word Cloud를 설치하여 시각화 할 수 있는데, 여기에서는 신속한 처리가 가능하고 폰트 및 색깔, 배열 등 다양한 선택과 사용이 편리한 wordcloudjp 엔진을 사용하였다.

<그림3-1> 상위 핵심어의 Word Cloud 시각화

위 표에서는 붉은 색과 점선으로 강조된 단어들을 볼 수 있다. 이것들은 총 35개로 2018년(7월 전기 또는 12월 후기)에 재등장 한 것들인데, 2018년을 기준으로 1년에 기출문제가 다시 반복되어 재출제 된 확률이 50%가 된다는 결

과를 나타내고 있다[5]. 그만큼 기출어휘가 중요하며 우선순위를 두고 학습해야 한다고 할 수 있다.

먼저 첫 번째로 핵심 키워드별로 자세히 살펴보도록 하겠다. 상기에서도 언급하였듯이 N3에서는 2회 이상 등장한 핵심어는 총 131개로 나타났다. 이 중에서 가장 많이 출현한 핵심 키워드는 상위 5회를 기록한 'すぎる' 동사로 나타났다. 'すぎる'는 N4~N1의 문법문제에서도 자주 등장하고 있는 동사이다[6]. N3 문자·어휘에서는 동사 연용형(ます형), 형용사(い형용사)·형용동사(な형용사)의 어간과 결합되어 '너무~하다'의 '의미'를 묻는 문제와 '복합동사' 및 '동의어'를 묻는 형태로 빈출되고 있다. 문법형태로는 '明るすぎる'(너무 밝다, 2회), '多すぎる'(너무 많다, 1회), 복합동사로는 '通りすぎる'(지나가다, 1회), 동의어 문제로는 문제4에서 '経つ'와 같은 의미를 고르는 문제로 등장하였다. 이렇게 'すぎる' 동사는 N3의 문자·어휘와 문법 양쪽의 문제에 최적화되고, 패턴화 되고 있는 사실을 파악 할 수 있다.

4회 출현된 '的'는 접미사의 형태로 3자 명사와 결합되어 출제되고 있는데, '一般的', '自動的', '代表的', '積極的', 등의 형태가 등장하였다.

다음으로 3회 등장한 어휘들은 '想像', '期待', '出張', '規則', '疲れる', '断る', '怒る', 'そっくりだ', '得意だ', 'きつい', 'うっかり' 등으로 나타났다. 동사 '疲れる'는 문제2 표기에서 한자 고르기 문제 1회, 문제4에서 유의어 'くたびれる'를 고르는 문제로 2번이나 등장하였는데, 추후에도 같은 패턴으로 출제될 수 있으므로 반드시 유의어와 같이 암기·학습할 수 있도록 방향을 제시해 주는 것이 효과적이라 할 수 있다.

'断る'는 문제3 문맥규정의 괄호 넣는 문제(1회), 문제5의 용법 문제(2회)로 출제되었는데, 읽기와 의미가 모두 중요시되는 단어이다. 참고로 1급에서는 개정이전(1990~2009)에 "きっぱり断られた"의 형태로 부사의 의미를 묻는 문제로도 출현되었다.

'怒る'는 문제4 유의표현에서 수동형태인 'どなられる=怒られる'(1회), '短気だ=すぐ怒る'(2회)와 유사의미를 찾는 문제로 연결되어 등장하고 있는데, 유의표현에 있어서는 각각의 의미가 무엇인지 숙지하고, 동종 어휘군으로 분류하여 암

5) N3 문자·어휘 문제는 총 35개가 출제되는데, 기출문제 재 출제 확률은 6개월에 25%, 1년에 50%가 된다.

6) *大きすぎる。(N4, 2011) *味が濃くなりすぎてしまいました。(N3, 2010) *やりすぎるのもよくない。(N3, 2013) *いつも作りすぎてしまう。(N3, 2013) *ほんの一部にすぎない。(N2, 2010) *せっけんを使いすぎずに、さっと洗うのが肌にはよい。(N2, 2011) *人口54人の小さな村にすぎなかった。(N2, 2013) *問題点を指摘しようとしたにすぎず、(N1, 2010) *つい仕事に夢中になりすぎる。(N1, 2016)

기하는 것이 어휘력 향상에 도움이 될 것이다.

3회 출현한 문제 가운데, ‘そっくり’와 개정이전에도 등장하였던 ‘うっかり’의 단어를 볼 수 있다. ‘そっくり’는 부사(몽땅의 의미)가 아닌 ‘꼭 닮다’, ‘붕어빵이다’라는 형용동사의 그 의미파악을 묻는 문제로 패턴화 되어 빈출되고 있다. 또 이 두 단어 외에도 ‘しっかり’와 개정이전 기출문제였던 ‘がっかり’가 2회 출현하였는데, 이른바, ‘り’로 끝나는 어휘군에 기출문제의 경향성이 놓여 있다는 것을 알 수 있다.

형용동사 ‘得意だ’는 문제1(읽기)에서 2회, 문제4에서 유의어 ‘上手だ’를 선택하는 문제로 3회 등장하였다. 형용사 ‘きつい’(힘들다) 또한 문제4에서 유의어 ‘大変だ’를 고르는 문제로 2회 출현되어 출제패턴이 정형화 되어가고 있는데, 개정이전에 2급, 1급에서도 등장한 ‘(바지 등이)꼭 끼다’의 뜻으로 문제3에서도 등장하였으므로 ‘緩い’의 반대어로 학습하는 것이 효율성을 높일 수 있다.

다음으로는 2회에 걸쳐 등장한 문자 · 어휘를 하기와 같이 자세히 살펴보도록 하겠다. 먼저 동사로는 ‘疑う’, ‘重ねる’, ‘迷う’, ‘しゃべる’, ‘話す’, ‘教える’, ‘転ぶ’, ‘回す’, ‘結ぶ’, ‘表す’, ‘覚える’, ‘困る’, ‘包む’, ‘預ける’, ‘逃げる’, ‘守る’, ‘信じる’, ‘落ち着く’, ‘確かめる’, ‘経つ’, ‘引き受ける’, ‘かれる’, ‘どなる’, ‘余る’, ‘残る’, ‘くたびれる’, ‘たまる’ 등이 있다.

이중에서 의미를 명확히 숙지하고 있는가에 대한 문제 유형의 형태로는 ‘重ねる’(문제3), ‘落ち着く’((마음,기분 등이)가라앉다, 안정되다/문제3, 5), ‘確かめる’(문제3, 4), ‘経つ’((시간,세월 등이)지나다, 경과하다/문제3, 4), ‘かれる’(문제3, 5), ‘どなる’(문제4, 5), ‘たまる’((일,돈 등이)쌓이다, 모이다), 등이 있다.

또한 한자(문제2)를 묻는 문제로는 ‘重ねる’, ‘結ぶ’, ‘困る’, ‘包む’, ‘預ける’, ‘逃げる’(문제2, 2회), ‘信じる’, ‘守る’ 등이 등장하였다. 그리고 페어 형태의 패턴화 된 어휘는 ‘疑う=本当ではない’, ‘しゃべる=(よく)話す’, ‘覚える=暗記する’, ‘くたびれる=疲れる’(문제4), ‘どなる=怒る’, ‘余る=残る’, ‘たまる＝残る’, 등이 문제 4, 5에서 나타났다.

N3에서의 복합동사는 특히 ‘受ける’와 결합된 단어가 종종 출현되고 있는데, 대표적으로 ‘引き受ける’(2회), ‘受け入れる’(1회), ‘受け付ける’(1회) 등이다. ‘引き受ける’는 문제3, 문제5에 2회 등장한 만큼, 그 의미를 명확히 묻고 있으며, 항상 ‘受け入れる’와 혼동을 유도하는 경향적 특징이 두드러지는 단어라 할 수 있다. ‘受け入れる’는 N2에서도 출제(2011) 된 바가 있다. 그리고 위에서 언급한 ‘通りすぎる’ 외에 ‘締め切る’, ‘話しかける’, ‘落ち着く’, ‘見送る’, ‘似合う’, ‘区切る’ 등도 1회씩 등장하였는데, 추후에도 반복될 여지가 충분히 있으므로 복합동사에서는 이 단어들을 위주로 제시하여 학습토록 하는 것이 좋을 것이다.

다음으로 부사에 대해 살펴보면, ‘どきどき’, ‘すべて’, ‘本当’, ‘必ず’, ‘だいたい’, ‘少し’, ‘さっき’, ‘そっと’ 등이 2회에 걸쳐 등장하였다. 부사문제는 주로 문제3, 4, 5의 문형에서 묻기 때문에 명확한 의미를 알아두어야 한다. 특히 또 ‘どきどき’외에도 ‘そろそろ’, ‘いちいち’, ‘いよいよ’ 및 개정이후에 등장한 ‘からから’(바싹 바싹), ‘がらがら’, ‘ふらふら’, ‘ぶらぶら’와 같이 ‘단어2개가 중복되는 부사’에 기출 문제의 포인트가 있다는 점이 중요하다. 이러한 부사들은 모두 2급과 1급에서도 빈출되고 있으므로 ‘り’부사와 마찬가지로 중요성이 상당히 높다고 할 수 있다[7].

형용동사의 출현 형태를 보면, ‘複雑だ’, ‘盛んだ’, ‘上手だ’, ‘正常(だ)’, ‘同じ(だ)’, ‘得意だ’, ‘不安だ’, ‘清潔だ’, ‘短気だ’, ‘大変だ’, ‘本当だ’ 등이 2회 등장하였다. ‘複雑だ’와 ‘正常(だ)’는 문제2와 3에서 한자 고르기, 괄호 넣기로 출제되었다. 이 단어들은 각각 ‘複-復-腹’, ‘常-情’(1,2급은 ‘常-裳-賞-償’)과 같이 혼란을 유발하기도 하는데, 두 단어 모두 다소 까다로운 한자어를 바르게 선별해야 하는 문제적 특성을 지니고 있다고 할 수 있다.

‘盛んだ’와 ‘清潔だ’는 문제3과 5에서 나온 만큼 그 의미 파악에 포커스가 놓여 있다. 또 문제4의 같은 의미 고르기 문제로는 상기에서 지적한 ‘上手だ=得意だ’, ‘短気だ=すぐ怒る’, ‘きつい=大変だ’, 등이 있다. 또 ‘同じ(だ)’는 ‘同じところ’와 같이 연체사로 등장하기도 하였는데, 연체사 형태로는 ‘小さな’, ‘大きな’, ‘おかしな=変な’도 자주 등장하고 있으므로 N3에서 이와 같은 모든 연체사를 기본적으로 숙지해 둘 필요가 있다.

형용사는 ‘厚い’, ‘恋しい’, ‘おかしい’, ‘苦しい’, ‘緩い’, ‘まぶしい’, 등이 2회 등장하였다. ‘厚い’와 ‘恋しい’는 문제1과 2에서 한자읽기와 고르기 문제로 출현하였다. 특히 ‘厚い’는 청해에서도 ‘暑い’, ‘熱い’ 등과 혼동을 유발하므로 의미파악을 간과해서는 안 되는 중요 단어라 할 수 있다.

‘苦しい’는 문제1(한자읽기)과 문제3(문맥규정)에 등장하였는데, N4 · 5의 苦(にが)い(쓰다)와 항상 읽기 혼동을 유발하므로 주의 할 필요가 있다. 참고로 ‘苦手だ’는 N3, N2 문제로 출제(2003)된 만큼, ‘苦い · 苦しい · 苦手だ’의 형태로 연관어를 학습하는 것이 효과적이라 할 수 있다. 이외에 ‘まぶしい’는 문제4에 ‘明るすぎ

7) ‘り’ 부사 가운데, ‘がっかり’는 2급에서도 2회(개정이전), 1급에서 1회(2011) 등장하였고, ‘うっかり’는 2급과 1급에서 각각 1회씩(개정이전) 등장하였다. ‘はっきり’는 2급에서 1회(2014), ‘はっきりしない’, ‘はっきりした’ 등의 형태로는 2회(개정이전), 1급에서는 ‘はっきりする’ 형태로 1회(2011) 출제되었다. 또 ‘2개 중복되는 단어’ 가운데, 2급에서는 ‘いちいち’, ‘いよいよ’, ‘ぶらぶら’(개정이전)가 출현하였고, 개정이후에는 이중에 ‘ぶらぶら’(2011)가 1회 등장하였다. 1급에서는 ‘ふらふら’, ‘ぶらぶら’ (개정이전)가 1회씩 출제 된 바가 있다.

る'와 유의어 문제로 등장하였는데, 한국어 '부시다'의 어간을 취하고 있으므로 학습자들이 쉽게 받아들일 수 있는 단어이다.

1자 명사로는 'うわさ'(噂), '空', '葉', 'わけ'(訳), 'ところ'(所) 등이 2회 출현되었다. 이중에 'うわさ'는 문제3 문맥규정에서 의미를 묻는 문제로 다뤄지고 있다. 이중에 특히 '空(から)'는 '읽기', '의미', '용법' 문제로 출현하고 있는 중요성이 높은 단어라 할 수 있다. 같은 의미인 'がらがら'(텅빔)도 문제3에서 등장하고 있는데, 기출단어 'からから'(바싹바싹)과 혼동을 유발하므로 의미와 형태에 주의를 기울여야 한다.

'わけ'는 문제4의 유의어 '理由'를 찾는 문제로 등장하였다. 'ところ'도 N5-4에서 출제되는 '곳'의 뜻이 아닌 '同じところ'=共通点, '悪いところ'처럼 '점'으로 해석되는 문제가 중요하게 다루어지고 있다.

다음으로 3자명사와 훈독명사의 형태로는 '週刊紙', 'やり方', '決まり'등이 2회 출현되었다. 이중에 'やり方'는 문제4에서 '手段'의 유의어 선택 문제로, '決まり' 또한 '規則'를 고르는 동의어 문제로 계속해서 출제되고 있다.

기타, 외래어로는 'スケジュール＝予定', 'カーブ＝曲がる' 등이 유의어 형태로 등장하였고, 'キャンセル'은 문제3 괄호 넣기에 출현하였다. 이것의 연어로 볼 수 있는 '取り消す'는 N1에서는 동의어 문제로 출현(2017)되고 있다. 'リサイクル'는 'ゴミ'+'減る', '減らす', '増える', '減少する', '増加する'와 같은 관련 명사와 증감의 동사가 같이 출현하는 경향이 강하기 때문에 연관어에도 중요성을 두어야 할 필요가 있다[8]. 외래어는 가장 쉬운 파트 중에 하나이지만, 청해에서 알아듣기 어려운 'マナー'(2017), 'ユーモア'(2010) 등은 반복적인 연습을 통해 리스닝에 익숙해져야 한다. 이와 같은 외래어에 있어서는 연관어 학습-청해로 연계되는 교육과 방향제시가 기본적으로 요구된다고 할 수 있다.

마지막으로 관용구를 살펴보면, '身につける'는 문제 5(용법)에서 올바른 의미를 묻는 문제로 2회 등장하였다. 관용구는 '気に入る'(2012)가 개정이후 1회 재등장한 것을 제외하고는 대체적으로 개정이전보다 출제문제가 적은 편이다[9]. 그러나 개정이전에 나온 '気が長い', '気をつける'와 2급에서도 개정이전에 출제된 '気をつける', '気に入る', '気が長い' 등은 학습에 필수적인 파생어들이라 할 수 있

8) 2015년12월 문자어휘 기출 문제의 경우, 「缶やペットボトルなども捨てないでリサイクルすれば、ごみを減らすことができる」처럼 リサイクル은 보통 뒤에 ゴミ와 증감의 동사들이 오는데, 缶, ペットボトル,燃える(2017기출)ゴミ, 燃えないゴミ등과 같은 리사이클 연관어도 함께 숙지하는 것이 효율적이다.

9) 개정이전에는 'お世話になる', '顔が広い', '口がかたい', '気が長い', '気をつける', '気に入る' 등이 문제로 출제되었는데, 개정이후에는 이중에 '気に入る'(2012)가 재등장하였다.

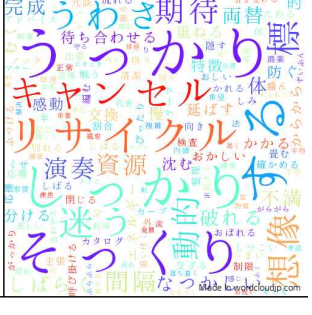
다. 또 '気がつく', '気がする', '気にする', '気になる' 등과 같은 '気'관련의 기본적 표현들도 그룹화하여 단계별로 명시해 두는 것이 학습효과를 제고하는 하나의 방법이라 할 수 있을 것이다.

나머지 본 절에서 다루지 않은 2자 한자명사에 대해서는 하기에서 파트별로 자세히 논해 보도록 하겠다.

3.1.2. 파트별 결과(문제1~문제5)

본 절에서는 문제1~5까지 파트별로 가장 많이 출현한 키워드를 Python으로 추출해 보고, Word Cloud로 핵심어가 한눈에 보일 수 있도록 시각적으로 나타내 보았다.

<표3-2> 파트별 핵심어와 Word Cloud 시각화

파트구분	핵심어	빈도수	시각화
문제1 (읽기)	得意 改札 卒業 到着 過去 努力 表す	2 2 2 2 2 2 2	
문제2 (표기)	帰宅 週刊紙 現在 楽器 成績 記録 逃げる	2 2 2 2 2 2 2	
문제3 (문맥규정)	うっかり 想像 目標 期待 自動的に 迷う しっかり そっくり キャンセル リサイクル うわさ	3 2 2 2 2 2 2 2 2 2 2	

문제4 (유의어)	すぎる (すぐ)怒る=短気だ 理由 =わけ 予定 =スケジュール 規則 =決まり 手段=やり方 指導する=教える しゃべる=話す くたびれる=疲れる 残る 上手だ(に) 本当だ 同じ(だ) 大変だ=きつい すべて 少し さっき 絶対 =必ず 約 =だいたい まぶしい =明るすぎる ところ	4 3=2 2=2 2=2 2=2 2=2 2=2 2=2 2=2 2 2 2 2 2=2 2 2 2 2=2 2=2 2=2 2	
문제5 (용법)	活動 空 性格 募集 修理 断る 身につける	2 2 2 2 2 2 2	

상기 표의 밑줄 표시 단어들은 2018년에 재출제 된 것들을 의미하는데, 문제수가 많은 문제3에서의 빈도가 약간 높다고 할 수 있다. 문제1(읽기)에서 가장 두각을 보인 단어는 '得意', '改札', '卒業', '過去', '到着', '努力', '表す'(2회), 등으로 나나났다. '過去(かこ)'는 '去'의 읽기가 '去年(きょねん)'과 다른 한자어 발음 특성 때문에 개정이전에는 N2에서도 출현되었고, 개정이후 부터는 N3에서 출제되고 있다. 한자읽기는 이렇게 같은 한자지만 읽기가 다른 경우, 장음 · 단음의 문제(到着(とうちゃく)'), 탁점 여부의 문제(努力(どりょく)), 훈독 · 음독발음 문제(相手(あいて), 台所(だいどころ), 合図(あいず), 内側(うちがわ), 笑顔(えがお)....) 등과 같이 출제문제가 패턴화 되어 있으며, 한자 2자명사가 중점적으로 등장하므로 한자읽기에 철저히 대비해야 할 파트이다[10].

10) 참고로 문제1외에 다른 파트에서도 등장한 2회 출제된 명사이자, 읽기에 주의해야할 단어는 位置(문4), 早退(문5), 共通(문4), 血液型(문2), 想像(문3), 分類(문5), 検査(문3), 出張(문3), 横断(문4), 完成(문3), 応募(문3), 協力(문4), 順番(문3), 通勤(문4), 得意だ(문4), ~的(문3), 覚える(문4), 回す(문2), 結ぶ(문2), 困る(문2), 組む(문2), 包む(문2), 苦しい(문3) 외에 자 · 타동사 折る · 折れる(문1) 등이 있다.(밑줄부분은 장음 및 탁점 유무 등에 유의)

문제2는 한자 찾기 파트로 문제5와 마찬가지로 한자에 익숙하지 않은 한국 학습자들이 가장 어려워하는 부분이다. 위의 표에서 확인되는 바와 같이 주요 단어로는 2회 출현된 '帰宅', '週刊紙', '現在', '楽器', '成績', '記録', '逃げる', 등이 있다. 이 단어들은 각각 '宅-沢', '刊-間/紙-誌', '在-左', '楽-薬', '績-積-適', '録-緑-縁', '逃-兆-桃'와 같이 거의 고정적으로 혼동을 유발하고 있으므로 반복적인 학습을 통해 숙지하여 정확한 한자를 고르도록 하는 것이 포인트라 할 수 있다[11].

문제3에서는 'うっかり'(3회), '想像', '目標', '期待', '自動的', '迷う', 'しっかり', 'そっくり', 'キャンセル', 'リサイクル', 'うわさ'(2회), 등이 두각을 나타냈다. 앞에서도 이미 지적한 바와 같이 3회에 걸쳐 등장한 'うっかり', 그리고 'しっかり', 'そっくり' 등, 주로 'り'형태의 부사가 두드러지고 있는 것을 알 수 있다. '想像'는 문제3(2회) 외에 문제1(1회)에서도 등장하였다. 문제1에서는 'そうぞう'의 'う'문자를 삭제하여 읽기 발음에 혼란을 초래하므로 장음의 여부에 주의를 기울여야 한다. '期待'도 문제3(2회)외에 문제2(1회)에서 출현하였는데, 문제2에서는 '待-持'의 형태로 등장하는 기출문제의 특성을 지닌다.

문제4에서의 가장 주요한 특징으로는 상기에서도 언급한 'すぎる'의 단어가 4회로 집중적으로 등장한 점인데, 주로 "너무~하다"는 뜻의 문법과 유의어 등을 고르는 문제가 꾸준히 출현되었다는 점이다. 이 외에도 유의어 문제는 '手段=やり方', '指導する=教える', 'しゃべる=話す', '短気だ=すぐ怒る', '理由=わけ', '予定=スケジュール', '規則=決まり', 'くたびれる=疲れる', '残る=余る', '大変だ=きつい', '絶対=必ず', '約=だいたい', 'まぶしい=明るすぎる' 등이 2회 등장하였다. 기타, 의미파악의 여부와 어휘력이 중요한 문제4에서 1회 출제된 단어 중에는 '内緒=話さない' 등도 있다. 이러한 것들은 모두 JLPT 기출 문제의 패턴화 특징을 알 수 있는 어휘들이므로 연관어로 학습하여 어휘력을 향상시키는 방법이 효과적이라 할 것이다.

문제 5에서는 올바른 활용을 묻는 문제로 '空', '活動', '性格', '募集', '修理', '断る', '身につける' 등이 2회 출현되었다. 가장 특징적인 단어로는 위에서도 지적하였듯이 '空'를 들 수가 있다. 개정이전에는 문제1에서 등장하였는데, 단어적 특성상, 읽는 방법과 의미(텅빔)을 정확히 파악해야 하고, 1급이라면 空っぽ까지 연관해서 학습해 볼 수도 있다.

11) 참고로 문제2와 다른 파트에서 등장한 2회 출제된 명사이자, 한자에 주의해야 할 단어로는 '期待'(문3), '経由'(문5), '規則'(문4), '欠点'(문4), '減少'(문5), '自信'(문3), '解決'(문3), '正常'(문3) 등이 있다. (밑줄부분은 한자 고르기 문제에서는 틀린 한자로 치환되어 혼동을 유도하므로 유의)

‘募集’는 개정이전에 문제1 한자읽기에서 등장한 바와 같이 읽는 방법, 한자 구분(募-墓-幕), 의미까지 모두 연관해서 학습해 두어야 할 것이다. 기타, 문제5에서 1회 출제된 단어 중에는 ‘渋滞’와 ‘滞在’ 등이 있는데, 모두 그 의미에 집중할 필요가 있으며, ‘渋滞’는 장음 발음과 한자 ‘滞’를 ‘帯’로 혼동하지 않도록 유념하여 학습토록 하는 것이 중요한 포인트이다.

3.2. 형태소 분석 결과

상기에서 언급한 part_of_speech.split를 이용하여 전체 문자 · 어휘를 품사별로 간단히 카운트 해 보고, 또 att=part_of_speech를 사용하여 품사1,품사2,품사3 등으로 구체적으로 형태소를 세분화하여 추출 해 보았다. 그리고 ‘part_of_speech.startswith’(또는 POSKeepFilter)를 사용하여 품사별로 개수를 출력 해 보았다. 이에 대한 결과를 N3 품사 구분 용도에 맞도록 종합적으로 재정리를 해 보면 아래와 같다.

<표3-3> 품사별 빈도 순위

순위	품사	빈도수	비율(%)
1	명사	367	49.7
2	동사	199	26.9
3	형용사	45	6.1
4	형용동사	39	5.3
5	부사	39	5.3
6	외래어	29	3.9
7	접미사	15	2.0
8	연체사	4	0.5
9	접두사	2	0.3
계		739	100

품사별로는 명사(49.7%) -동사(26.9%) -형용사(6.1%) -형용동사(5.3%) -부사(5.3%)- 외래어(3.9%)- 접미사(2.0%) -연체사(0.5%)-접두사(0.3%) 순으로 카운트 되었다.

N3의 중요도를 크게 4가지 나누어 살펴보면, 첫 번째로 한자 2자 명사(한자 1자 명사 포함)가 약 49%로 대략 과반수에 근접한 비율을 보이고 있듯이 상기에서 언급한 131개의 빈출 핵심어와 마찬가지로 가장 포커스를 두고 학습해야 할 파트이다. 그리고 그 다음 두 번째로 강조되어야 품사는 약 27%의 빈도 비율을 보이고 있는 동사이다. 세 번째로는 약 17%의 형용사, 형용동사, 부사

파트이며, 네 번째로는 기타 약 7%의 외래어, 접미사, 접두사, 연체사 파트이다.

접미사와 접두사는 출현 빈도수는 적은편이며 비교적 기초적인 단어들이 등장하고 있는 것을 알 수 있다. '的', '紙', '型', '(乗車, 入場..)券', '書', '家', '料', '学', '差', '産', '点', 등의 접미사 중에서는 상기에서 언급한 '的'외에 '紙'는 3자명사 '週刊紙'로 2회 등장하였다. '型'는 血液型(1회)처럼 변칙 발음을, 또 '券'은 '巻'과 같이 한자오용을 주의해야 하며, '書'는 '申込書'(1회)처럼 읽기와 의미 묻는 문제가 중요한 포인트이므로 주의를 기울여야 한다. 기타, 비교적 출제빈도가 아주 적은 접두사는 '全'과 '翌'이 각각 '全人口'(1회), '翌年'(1회) 등으로 등장하였다.

마지막으로 연체사는 '同じ', 'おかしな', 'あらゆる'등이 각각 '同じところ=共通点', 'おかしな=変な', 'あらゆる=全ての'와 같이 대부분 유사 의미를 묻는 문제로 출제되었다. 상기에서도 지적한 이러한 연체사 활용형태를 먼저 파악하고, 기출어휘 모두 연어패턴으로 숙지함으로써 학습효과와 능률을 제고시킬 수 있을 것으로 판단된다.

4. 나가기

본 연구에서는 텍스트마이닝 분석을 기반으로 Python3.7 프로그램, Jupiter Notebook 툴, Janome 형태소 분석엔진을 활용하여 개정이후 N3 기출 문자·어휘의 상위 빈출 핵심어를 추출하고, 문제패턴과 경향을 분석하였다.

N3에서 2회 이상 출현된 핵심어는 총 131개로 나타났다. 이 중에서 가장 많이 출제된 단어는 상위 5회를 기록한 'すぎる' 동사이다. 그 다음으로는 4회 출제된 핵심어로 '的', 3회는 '想像', '期待', '出張', '規則', '疲れる', '断る', '怒る', 'そっくりだ', '得意だ', 'きつい', 'うっかり', 등으로 나타났다.

파트별로는 문제1 읽기에서 두각을 나타낸 단어는 '得意', '改札', '卒業', '過去', '到着', '努力', '表す'(2회), 등으로 나나났다. 문제1에서는 읽기 변칙 한자, 장음·단음, 탁점 여부, 훈독·음독발음 문제가 빈출되고 있는데, 이러한 문제에 해당되는 이 단어들이 반복해서 등장하고 있음을 알 수 있다.

문제2 표기에서는 '帰宅', '週刊紙', '現在', '楽器', '成績', '記録', '逃げる'(2회), 등이 출제되었는데, 각각의 한자들은 모두 유사 한자로 바뀌어 혼동을 유도하

는 특징적인 단어들이라 할 수 있다.

문제 3 문맥규정에서는 ‘想像’, ‘目標’, ‘期待’, ‘自動的’, ‘迷う’, ‘うっかり’, ‘しっかり’, ‘そっくり’, ‘キャンセル’, ‘リサイクル’, ‘うわさ’, 등이 출제되었다. ‘うっかり’가 3회에 걸쳐 등장한 바와 같이 주로 ‘り’형태의 부사가 두드러지고 있으며, ‘캔슬’, ‘리사이클’의 가타카나가 출현되고 있는 경향을 파악할 수 있다.

문제4 유의어 파트의 특징으로는 ‘すぎる’의 단어가 4회로 집중적으로 등장하였고, 그 문법과 유의어 등을 고르는 문제가 꾸준히 출현되었다는 점이다. 이외에도 유의어 문제는 ‘手段=やり方’, ‘指導する=教える’, ‘しゃべる=話す’, ‘短気だ=すぐ怒る’, ‘理由=わけ’, ‘予定=スケジュール’, ‘規則=決まり’, ‘くたびれる=疲れる’, ‘残る=余る’, ‘大変だ=きつい’, ‘絶対=必ず’, ‘約=だいたい’, ‘まぶしい=明るすぎる’ 등이 2회 등장하였다. 이러한 단어들은 페어로 등장하여 반복적으로 패턴화 되고 있기 때문에 향후에도 출현 가능성이 높을 것으로 추정되며, 연관어로 학습하여 어휘력을 향상시키는 방법이 효과적이라 할 수 있다.

문제 5에서는 정확한 활용을 묻는 문제로 ‘空’, ‘活動’, ‘性格’, ‘募集’, ‘修理’, ‘断る’, ‘身につける’ 등이 2회 출현하였는데, 1자명사의 가장 특징적인 단어로는 ‘的’ 이외에도 ‘空’를 들 수가 있다.

마지막으로 품사별로는 명사(49.7%) -동사(26.9%) -형용사(6.1%) -형용동사(5.3%)-부사(5.3%)- 외래어(3.9%)- 접미사(2.0%) -연체사(0.5%)-접두사(0.3%) 순으로 빈도 비율이 집계 되었다. 결과적으로 ① 131개의 빈출어휘, ② 한자 2자 명사, ③ 동사, ④ 형용사 · 형용동사 · 부사, ⑤기타, 외래어 · 접미사 · 접두사 · 연체사 순으로 그 중요성을 두고 학습하는 것을 제안해 볼 수 있다.

본 연구에서는 지면과 시간의 제약 상, 개정이전의 문자 · 어휘 내용은 다루지 못한 관계로 분석 데이터의 양이 적은 한계가 있다. 텍스트 마이닝은 다량의 데이터를 이용하는데 유용하고, 키워드 추출과 패턴 파악을 통해 미래 예측을 가능케 하는 강점이 있다. 본 연구는 이러한 연구의 첫 스텝이자 시도로 향후에는 개정이전과 이후의 N2, N1 문자어휘를 중심으로 보다 많은 양의 분석을 단계적으로 시도해 보고, 추후 저작권 문제가 해결된다면 텍스트 마이닝에 훨씬 더 효과적인 독해파트 내용을 다루어 지금까지 파악되지 못했던 결과를 밝혀 보도록 하겠다. 이러한 부분들에 관해서는 향후의 연구과제로 삼기로 하겠다.

본 연구의 의의 및 기대효과는 다음과 같다. 일본어 분석에 있어 지금까지 거의 시도되고 있지 않는 텍스트 마이닝 기법을 활용해 봄으로써 연구의 폭을 확대하고, 키워드 도출과 전체별, 파트별, 형태소별 및 시각화 분석을 통해 보

다 새로운 결과를 창출하는데 본 연구가 다소 일조 할 수 있을 것으로 기대된다. 또 비교적 정확한 데이터 결과 및 시각적 자료를 제시함으로써 일본어 학습자들의 시험대비와 현 교육기관에서 보다 효과적이고 객관적인 티칭자료 활용될 수 있을 것으로 기대를 할 수 있다. 나아가서는 저작권 문제 해결 이후, 본 연구 결과를 바탕으로 레벨별 · 분야별 디지털 교재를 제작하여 향후의 디지털 시대의 학습자들을 위한 교육 자료로 활용 할 수 있을 것으로 기대한다.

【참고문헌】

강범일 · 송민 · 조화순(2013) 「토픽 모델링을 이용한 신문 자료의 오피니언 마이닝에 대한 연구」 『한국문헌정보학회지』 47(4), pp.315-334. (DOI:https://doi.org/10.4275/KSLIS.2013.47.4.315)

김연주 · 이건수(2016) 「텍스트 마이닝을 활용한 고등학교 영어 교과서 어휘 목록 개선 방안에 대한 제안」 『언어학』 24(4), 대한언어학회, pp.281-301. (DOI:https://doi.org/10.24303/lakdoi.2016.24.4.281)

김유영(2015)「일본어 텍스트의 가독성 레벨 분석-旧일본어능력시험 기출문제 데이터에 대한 통계적 검증을 기반으로-」 『일본학보』 103, pp.21-40.

김현정 · 조남옥 · 신경식(2015)「항공산업 미래유망분야 선정을 위한 텍스트 마이닝 기반의 트렌드 분석」 『지능정보연구』 21(1), pp.65-82.

권충훈(2018)「텍스트 마이닝과 언어네트워크 분석을 활용한 중등교사임용 교육학시험 내용 분석」 『교육혁신연구』 28(3), 부산대학교 교육발전연구소, pp.1-25. (DOI:https://doi.org/10.21024/pnuedi.28.3.201809.1)

민광준(2013)「코퍼스를 이용한 고등학교 일본어 교과서의 가타카나어 분석」 『일본어학연구』 38, 한국일본어학회, pp.73-89.

박선주(2018)「코퍼스를 이용한 일본어 어휘학습의 제안-'的'가 붙는 어기의 요소를 중심으로」 『일본어 교육』 85, pp.1-16.

박자현 · 송민(2013)「토픽모델링을 활용한 국내 문헌정보학 연구동향 분석」 『情報管理学会誌』 30(1) 한국정보관리학회, pp.7-32. (DOI:https://doi.org/10.3743/KOSIM.2013.30.1.007)

박진균 · 김택윤 · 송민(2017),「텍스트마이닝을 이용한 윤동주 연구의 개체계량학적 분석」 『한국비블리아학회지』 28(1), pp.191-207. (DOI:https://doi.org/10.14699/kbiblia.2017.28.1.191)

이도열(2012)「일본어 능력평가 시험간의 비교척도 고찰:JLPT N1레벨과 EJU일본어 득점에 대한 비교 척도 분석을 중심으로」 『일본어문학』 52, pp.55-77.

조주연 · 김현숙 · 조민제(2018)「텍스트 마이닝을 활용한 청소년 문제의 이슈변화 분석 -2008-2018년 인터넷 신문기사를 중심으로-」 『교육혁신연구』 28(4), 부산대학교 교육발전연구소, pp.461-482. (DOI:https://doi.org/10.21024/pnuedi.28.4.201812.461)

최정원·한호선·이미영·안준모(2015)「텍스트마이닝 방법론을 활용한 기업 부도 예측 연구」 『生産性論集』 29(1),pp.201-228. (DOI:https://doi.org/10.15843/kpapr.29.1.201503.201)

押尾和美外(2008)「新しい日本語能力試験のための語彙表作成に向けて」国際交流基金日本語教育紀要(4), pp.71-86.

(DOI:https://doi.org/10.20649/00000047)

鈴木美恵 · 최영숙(2016)「日本語における外来語のアクセントの分析：JLPT語彙の外来語を中心に」『日本語文学』37, pp.69-94. (DOI:https://doi.org/10.21792/trijpn.2016..73.004)

永野亜季(2013)「新JLPT問題集レベルN1における語彙調査-公式問題集と韓国出版の問題集を比較して」『일본어문학』60,pp.87-108.(DOI:https://doi.org/10.21792/trijpn.2013..60.005)

조은영(2018)「日本語能力試験対策用の教材に見られる副詞の出現傾向について」『日本語文学』76, pp.185-203. (DOI:https://doi.org/10.18704/kjjll.2018.03.76.5)

Feldman,R.,&Dagan,I.(1995). Knowledge Discovery in Textual Databases. *KDD, Vol.95,* pp.112-117.

Gartner Group(1994). Data Mining:The next generation of business intellegence? *ATG Research Note T-517-246*, Gartner Group Inc., Stamform, CT.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes and Jeffrey Dean Google.(2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics, vol. 5*, pp.339 - 351.

Michael J.A.Berry and Gordon S. Linoff(2004). *Data Mining Techniques(Second Edition).* Wiley Publishing, Inc, p.7.

U.Fayyad, G.Piatetsky-Shapiro, and P.Smyth.(1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, Vol.17*, Menlo Park, Calif.:AAAI Press, pp.37-54.

【용례출전】

이치우(2018)『JLPT콕콕찍어주마N3문자 · 어휘』,『JLPT콕콕찍어주마N3문법』, 다락원

_____『JLPT콕콕찍어주마N2문자 · 어휘』,『JLPT콕콕찍어주마N2문법』, 다락원

_____『JLPT콕콕찍어주마N1문자 · 어휘』,『JLPT콕콕찍어주마N1문법』, 다락원

JLPT연구모임(2018)『딱한권 JLPTN3』, 시사일본어사

_____『딱한권 JLPTN2』, 시사일본어사

_____『딱한권 JLPTN1』, 시사일본어사

国際交流基金 · 日本国際教育支援協会(2012)『日本語能力試験JLPT公式問題集N3』,凡人社

国際交流基金 · 日本国際教育支援協会(2018)『日本語能力試験JLPT公式問題集第二集N3』,凡人社

논문 투고 일자 : 2019. 06. 30.
논문 심사 일자 : 2019. 08. 02.
게재 확정 일자 : 2019. 08. 05.

<要旨>

テキストマイニングを活用した日本語能力試験の内容研究
－JLPT N3文字・語彙を中心に－

李有姫

本研究では、テキストマイニング分析に基づき、Python3.7、Jupiter Notebook、Janome(Mecab)エンジンを利用して、N3既出文字・語彙(2010-2018)の頻出キーワードを抽出し、問題のパターンと傾向を分析することにより、効果的な学習の方向を提案した。N3で2回以上頻出した単語は計131個だった。上位キーワードは「すぎる」(5回)、「的」(4回)、「疲れる」、「断る」、「怒る」、「出張」、「規則」、「そっくりだ」(3回)などである。パート別には、問題1で最も多く出現した単語は、「得意」、「改札」、「卒業」、「過去」、「到着」、「努力」、「表す」(2回)などとなった。問題2では、「帰宅」、「週刊紙」、「現在」、「楽器」、「成績」、「記録」、「逃げる」(2回)、問題3では「うっかり」(3回)、「想像」、「目標」、「期待」、「自動的」、「迷う」、「しっかり」、「そっくり」、「うわさ」、「キャンセル」、「リサイクル」(2回)などとなった。問題4では、「すぎる」(4回)、「手段=やり方」、「指導する=教える」、「しゃべる=話す」、「(すぐ)怒る=短気だ」、「理由=わけ」、「規則=決まり」、「くたびれる=疲れる」、「大変だ=きつい」(2回)、問題5では、「空」、「活動」、「性格」、「募集」、「修理」、「断る」、「身につける」(2回)などが登場した。これらのキーワードは、各パートの問題の類型に適した特性を帯びていると同時に、パターン化されており、今後も再出題される可能性が高い語彙と推測される。

品詞別には、名詞(49.7%)-動詞(26.9%)-形容詞(6.1%)-形容動詞(5.3%)-副詞(5.3%)-外来語(3.9%)-接尾辞(2.0%) -連体詞(0.5%)-接頭辞(0.3%)の順で頻度の割合が集計された。結果的には、①131頻出語彙、②漢字2字名詞、③動詞、④形容詞・形容動詞・副詞、⑤その他、外来語・接尾辞・接頭辞・連体詞の順で、その重要性を置き、学習することを提案したい。

A Study on the Contents of the Japanese Language Proficiency Test Using Text Mining Analysis
－ Focusing on JLPT N3 Vocabulary －

Lee, Yu-Hee

The objectives of this study were to : (1) extract keywords of the previous N3 test characters and vocabulary (2010-2018), and (2) analyze exam patterns and trends by utilizing Python3.7 program, Jupiter Notebook tool, and Janome(Mecab) stemming engine, based on text mining analysis.

A total of 131 words were extracted twice or more frequently in N3. The top key word was 'すぎる'(5 times) and '的'(4times) and '疲れる', '断る', '怒る', '出張', '規則', 'そっくりだ'(3times) were used in most of the exam.

In terms of sections, in Question 1, the most frequently appearing words were '得意', '改札', '卒業', '過去', '到着', '努力', '表す'(2times). In Question 2, seven words '帰宅', '週刊紙', '現在', '楽器', '成績', '記録', '逃げる'(2times) appeared and 'うっかり'(3times), '想像', '目標', '期待', '自動的', '迷う', 'しっかり', 'そっくり', 'うわさ', 'キャンセル', 'リサイクル' (2times) appeared in Question 3. In Question 4, the word 'すぎる' appeared intensively and the similar words '(すぐ)怒る=短気だ', 'くたびれる=疲れる', '大変だ=きつい'(2times) were set. In Question 5, '空', '活動', '性格', '募集', '修理', '断る', '身につける' (2 times) appeared as frequent words. These keywords have characteristics that are appropriate for the question type of each part and are patterned, so it is presumed that they are likely to be asked in the future.

With respect to the parts of speech, the frequency ratio was counted by the order of Noun(49.7%)-Verb (26.9%)-Adjective(6.1%)-Adjective verb (5.3%)-Adverb(5.3%)-Katakana(3.9%)-Suffix(2.0%)- Prenominal adjective (0.5%)-Prefix(0.3%). As a result, it is proposed to study, in order of the importance, the following parts of speech : (1) 131 high-frequency vocabularies, (2) two Kanji nouns, (3) verbs, (4) adjectives, adjective verbs, adverbs, and (5) other Katakana, suffixes, prenominal adjectives, and prefixes.