국립국어원 2019-01-43

발간등록번호
11-1371028-000786-01

한국어 정보 처리를 위한 어휘 관계 기초 자료 구축

사업 책임자명

김 소 정

국립국어원

국립국어원 2019-01-43

발간등록번호
11-1371028-000786-01

한국어 정보 처리를 위한 어휘 관계 기초 자료 구축

사업 책임자명

김 소 정

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 '한국어 정보 처리를 위한 어휘 관계 기초 자료 구축'에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2019년 11월 ~ 2020년 2월

2020년 02월 20일

사업 책임자: 김소정 소장(나라지식정보)

사업 기관 나라지식컨소시엄

사업 책임자 김소정 소장

사업 참여자 윤택기, 김종수, 김동진, 전화자, 김혜연, 최운천, 김정호, 서민석, 안정균, 김장현, 권혁주, 전호섭, 김은경, 하지연, 김기형

사업 보조원 조경찬

자문 위원 김진해, 박승희

품질 보증 손영호

[국 문 초 록]

본 사업의 목적은 한국어 인공지능 서비스 개발과 국어사전 정보 고도화를 위해 한국어 분산 의미 모델을 개발하는 것으로서, 세밀한 의미 정보가 식별된 어휘별로 비슷한말, 반대말, 상위어, 하위어 등 어휘 간의 관련성에 대한 실제 사용 양상을 평가하고 이에 대한 통계 정보를 확보하는 것이다.

이를 위하여 『우리말샘』에 존재하는 약 33만 개의 어휘쌍을 분석하여, 어휘쌍의 중복을 제외하고 사용자 양상 평가의 대상이 되는 20만 어휘쌍을 도출하였다. 제시된 20만 어휘쌍을 200개의 비슷한말, 반대말, 상위어, 하위어 세트로 구분하여 총 5만 명의 응답자가 평가를 하였으며, 응답자별로 1,005개의 어휘쌍에 대한 평가를 수행하도록 하였다. 최종적으로 유효한 응답 40,730건을 도출하였으며, 각 어휘쌍별로 최소 200명의 응답을 확보하였다.

수집된 사용자 응답에 따라, 총 20만 개의 평균과 표준편차를 활용한 베타계수, 응답의 분포의 정보량 제공을 위한 엔트로피 계수를 도출하고, 어휘 자체가 가지는 난이도에 따른 난이도 계수를 산출하였다.

본 과제를 통해 제시된 한국어 분산 의미 모델이 제시하는 시사점은 다음과 같다. 첫째, 방대한 분량의 사용자 평가 결과는, 최신의 인공지능 기계 학습에서는 제공할 수 없는 어휘별로 세밀하고 감성적인 거리 측정이 언어 사용자의 평가를 통하여 가능하다. 둘째, 표제어와 관련어에 대한 응답에 있어 사용자는 어떤 어휘를 먼저 인지하는지가 후속하는 응답에 영향을 미칠 수 있다. 셋째, 본 설문은 어휘를 의미별로 세분화하여 사용자가 평가하는 측면에서 세계 최초의 시도이다. 본 결과를 통해 다의어 형태의 자연어 처리에 대한 개선된 서비스 개발 전략이 가능할 것으로 기대한다.

핵심어: 한국어 분산 의미 모델, 실증 연구, 관련어, 정보 이론

[Abstract]

We aim to develop a Korean distributional semantic model investigating usage of crowds as well as obtaining empirical information, which is expected to be a basis of various Korean AI service and advancement of Korean open dictionary system.

In specific, we analyzed about 330,000 vocabulary pairs in open dictionary and derived 200,000 vocabulary pairs excluding overlapping pairs. A total of 50,000 respondents evaluated the 200,000 vocabulary pairs fall in similar words, opposite words, upper words, and lower words. Finally, 40,730 valid responses were collected and at least 200 responses were obtained for each vocabulary pair.

Based on the collected user responses, we derived three fold coefficients - a beta coefficient using a total of 200,000 averages and standard deviations, an entropy coefficient for providing information amount of the distribution of responses, and a difficulty coefficient according to the difficulty of the vocabulary itself.

The implications of this project are as follows. First, the vast amount of user evaluation results clearly suggest that the collected information regarding the perceived distance collected from questionnaire is a valid and even can complement the state of art AI machine learning. Second, it disclose the pattern that the position of words – whether head or tail - may affect the response. Third, this survey is, as researchers aware, the world's first attempt in terms of evaluating users by segmenting vocabulary by meaning. Thus, results of this study are expected to establish various strategies for such word disambiguity processing.

key words: Korean distributional semantic model, empirical study, relational/associated words, information theory

요약보고서

본 사업의 목적은 한국어 인공지능 서비스 개발과 국어사전 정보 고도화를 위해 한국어 분산 의미 모델을 개발하는 것으로서, 세밀한 의미 정보가 식별된 어휘별로 비슷한말, 반대말, 상위어, 하위어 등 어휘 간의 관련성에 대한 실제 사용 양상을 평가하고 이에 대한 통계 정보를 확보하는 것이다.

이를 위하여 『우리말샘』에 존재하는 33만여 개의 어휘쌍을 분석하여, 어휘쌍의 중복을 제외하고 사용자 양상 평가의 대상이 되는 20만 어휘쌍을 도출하였다. 제시된 어휘쌍에 대한 품질 높은 사용자 양상 평가 결과를 수집하기 위해, 어휘 간 관련성에 대한 사용자 평가 및 수집 데이터를 공개한 캠브리지 대학의 Simlex-999 등 분산 의미 모델 수립 문헌을 참고하여, 응답자에 대한 제시 및 질의 방법, 설문 결과에 대한 평가 방법을 설계하였다. 실제 설문 수행 전 전문가 검토 및 1,000개 문항에 대한 200명 대상의 시험 공정(파일럿 테스트)를 수행하고 본 공정에 착수하였다. 본 공정에서는 20만 어휘쌍을 200개의 비슷한말, 반대말, 상위어, 하위어 세트로 구분하여 총 54,000명의 응답자가 평가를 하였으며, 응답자별로 1,005개의 어휘쌍에 대한 평가를 수행하도록 하였다. 최종적으로 유효한 응답 40,730건을 도출하였으며, 각 어휘쌍별로 최소 200명의 응답을 확보하였다. 응답에 대한 평가 시간은 약 3~4초 소요된 것으로 분석되며, 본 공정과 시험 공정(파일럿 테스트)의 상관 계수 평가는 약 0.85점으로 국제 학술지에서 기초 자료 연구에서 통용되는 일반적인 평가 점수를 상회하였다.

수집된 사용자 응답에 따라, 총 20만 개의 평균과 표준편차를 활용한 베타계수, 응답의 분포의 정보량 제공을 위한 엔트로피 계수를 도출하고, 어휘 자체가 가지는 난이도에 따른 난이도 계수를 산출하였다. 이와는 별도로 각 응답에 소요되었던 시간을 산정하였다. 상호 배타적인 정보를 가지는 이들 3가지 계수 및 응답 시간은 수준 높고 다양한 인공 지능 서비스 개발 및 평가에 활용될 수 있을 것으로 기대한다. 하지만 일반인의 관점에서는 이해하기에 어려움이 있을 수 있으므로 사용자의 직관적인 이해를 도울 수 있도록 이 중에서 베타계수를 대상으로 하여 어휘 간의 관계 및 품사를 기준으로 강/중/약 등급을 추가로 제시하였다.

한편, 본 연구와 함께 진행된 추가 공정에서는 어휘쌍이 아닌 의미 번호가 비식별된 개별 어휘를 대상으로 한국어 사용자가 받는 내면적인 느낌 또는 의미의 강도를 추출하였다. 이를 위해 Harvard VI-4 분류체계의 기본 모델이었던 Osgood(1952)의 방식을 검토하고 개발하여, 1,000개의 형용사와 부사 어휘 각각에 대해 5개 차원에 대한 질문을 실시하여 총 5,000개의 응답이 수집되도록 설계되었고, 총 30명의 유효한 응답을 확보하여 분석하였다.

본 과제를 통해 한국어 분산 의미 모델이 제시하는 시사점은 다음과 같다.

첫째, 방대한 분량의 사용자 평가 결과는, 최신의 인공지능 기계 학습에서는 제공할 수 없는

어휘별 주관적인 느낌에 대한 면밀한 거리 측정이 사용자 설문 평가를 통하여 가능하다는 것을 제시하였다(예: 가장 관련성이 높은 비슷한 말은, '어머니-엄마', 가장 반대된다고 느끼는 반대말은 '당선-낙선'이었다). 특히 본 사업에서 제시된 설문은 어휘 관계별 거리 측정을 위한 가장 합리적인 문항을 도출하였다. 해당 문항은 내용 검토, 동료 검토, 시험 공정(파일럿 테스트) 등의 과정을 통해 제시한 것이다.

둘째, 어휘가 갖는 구체성의 정도에 따라 사용자 응답이 달라질 수 있음을 시사한다. 예를 들어 비슷한말과 반대말에 대한 응답의 베타계수는 상위어와 하위어에 대한 계수보다 높다. 또한, 표제어가 상위어인 경우와 하위어인 경우의 응답에 대한 편차가 발견되었다. 이는 적은 의미를 먼저 인지하고자 하는 심리적 요인에 기여할 수 있으며, 이 같은 사용자 인지의 '방향성'이나 '편견'의 존재는 인공지능 서비스 개발에 적용될 수 있는 여지가 크다. 뿐만 아니라 20만 쌍이라는 풍부한 기초 데이터는 이와 관련된 초기 연구의 초석이 될 수 있을 것으로 기대한다.

셋째, 본 설문 결과는 어휘를 의미별로 세분화하여 사용자가 평가한다는 측면에서는, 본 연구 수행 기관이 파악하기로 세계 최초의 시도이다. 본 결과를 통해 다의어 처리에 대한 다양한 전략 수립도 가능할 것으로 기대한다.(예를 들어, '가깝다-멀다'의 경우 시간 등의 은유적 의미로 쓰일 때에 비해 기본의미로 쓰일 때 더 높은 강도와 바른 응답 시간을 보였다. 마찬가지로 '가늘다-굵다'에 대한 평가는 기본 의미적 용법에서는 관련성을 '강'으로, 파생 의미적 용법에서는 '중'으로 평가되었다.)

넷째, 추가 설문을 통해 분석된 감성-극성 분석(sentimental polarity analysis)은 어휘 관계 분석에 대한 확장 및 사전 분류체계의 제시는 물론, 한국어에 대한 감성 분석에 대해 더 다채로운 분석이 가능함을 제시하였다. 또한, 다양한 대체어 제시 알고리즘에 적용될 수 있으며(예를 들어, '예쁘다'는 문맥에 따라 '좋다'로 대체할 수 있다), 집단 간 어휘 수용 차이가 존재할 수 있음을 시사한다.

본 연구 결과와 제시된 기초 자료는 향후 『우리말샘』 사전의 고도화와 시각화에 적용될 수 있을 것으로 기대한다. 아울러 우리말을 활용한 다양하고 유용한 인공지능 서비스의 개발 및 관련 연구에 활용될 수 있기를 기대한다.

목 차

Ⅰ. 과업 개요

1. 사업명

○ 한국어 정보 처리를 위한 어휘 관계 기초 자료 구축

2. 사업 기간

○ 2019년 11월 20일 ~ 2020년 2월 20일

3. 사업비

○ 935,000,000원

4. 주관 기관

○ 국립국어원

5. 수행 기관

○ 나라지식정보 컨소시엄

6. 사업 목적 및 필요성

○ 한국어 분산 의미 모델 등 한국어 인공지능 개발과 국어사전 정보 고도화를 위한 단어별 상관관계의 통계 정보 확보

○ 국어사전 등에 반영되어 있는 어휘 관계 정보(동의어, 비슷한말, 반의어, 상하위어 등)의 실제 사용 양상 평가로 관련 정보 정밀화

7. 주요 사업 내용

○ 국립국어원에서 제공하는 어휘쌍(20만여 항목)을 대상으로 평가

○ 평가용 어휘쌍 제시 방법 및 질의 방법, 평가 방법(평가 척도 등) 설계

○ 관련 어휘쌍을 이용한 어휘 관계 평가 설문 조사(20만여 항목 중 50% 이상, 항목당 200명 이상 평가) 실시 및 한국어 분산 의미 모델 평가용 어휘 관계 척도(유사도) 제시

○ 국어사전 어휘 관계 보완용 전체 평가 항목의 어휘별 등급화

Ⅱ. 사업 추진 내용

1. 사업 추진 목표

○ 어휘 관계의 정밀화 수립
- 국어사전 등에 반영된 어휘 관계 정보(비슷한말/반대말/상위어/하위어)의 사용 양상 평가

○ 단어별 상관관계 통계 정보 확보
- 한국어 인공지능 개발과 국어사전 정보 고도화를 위한 단어별 상관관계 통계 정보 확보

2. 사업 범위

○ 평가 척도 설계
- 평가용 어휘쌍 제시 방법
  * 피평가자가 어휘 관계 평가로 주어지는 어휘쌍에 대한 정보를 획득할 수 있는 방안을 예문 또는 뜻풀이 등으로 제시
  * 피평가자가 평가 대상 어휘쌍을 인지하지 못할 경우 해당 어휘쌍을 평가에서 제외(단, 전체 20만여 항목 중 50% 이상 평가)
- 어휘 관계 정보별(동의어, 비슷한말, 반의어, 상하위어 등)로 해당 어휘 관계를 제대로 평가할 수 있는 설문 방법의 수립 및 설계
- 설문 대상으로 제시된 어휘쌍의 관계를 제대로 평가할 수 있는 답변안 설계
  * 평가시간 약 5초 이내의 5점 척도로 예시

○ 어휘 관계 평가 설문 조사

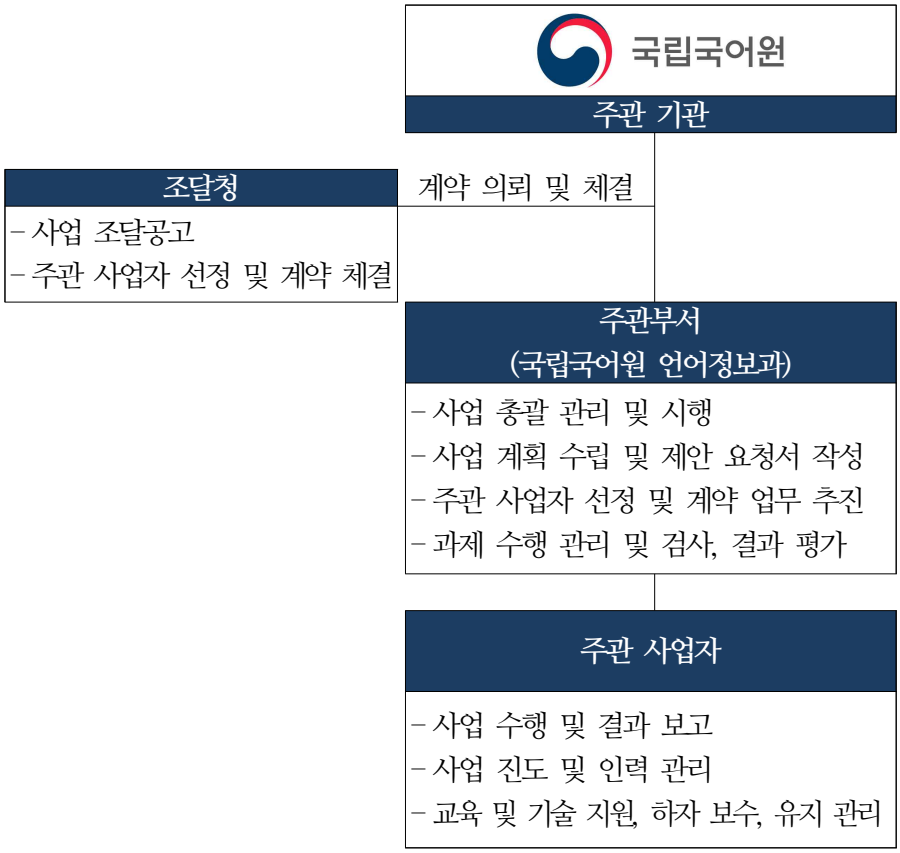
- 전체 대상 항목(20만여 항목 중 50% 이상, 항목당 200명 이상 평가)을 총 세 부문으로 나누어 어휘 관계 평가 설문 조사 실시
  * 비슷한말-반대말(7만여 항목) / 상위어(7만여 항목) / 하위어(6만여 항목)
  * 전체 20만 항목을 모두 평가하되, 평가된 최종 자료가 50%인 10만 항목 이상이어야 함

○ 어휘 관계 설문 조사 결과를 한국어 분산 의미 모델 평가용 어휘 관계 척도(유사도)로 제시
- 스피어만 상관계수(Spearman correlation coefficient) 등으로 산출
- 수집 자료의 파일 형식: 단어1, 단어2, 품사, 평가값(항목 간 탭으로 구분)
- 국어사전 관련 어휘 관계와 조사 결과를 비교하여 전체 평가 항목의 어휘별 등급화
  * 전체 평가 결과를 바탕으로 국어사전 내 어휘 관계와 비교한 결과(유사도)를 바탕으로 등급화·차등화 실시
  * 평과 결과에 따라 최소 4개 이상의 부문으로 그룹화

Ⅲ. 사업 추진 체계

1. 총괄 추진 체계

가. 사업 추진 체계

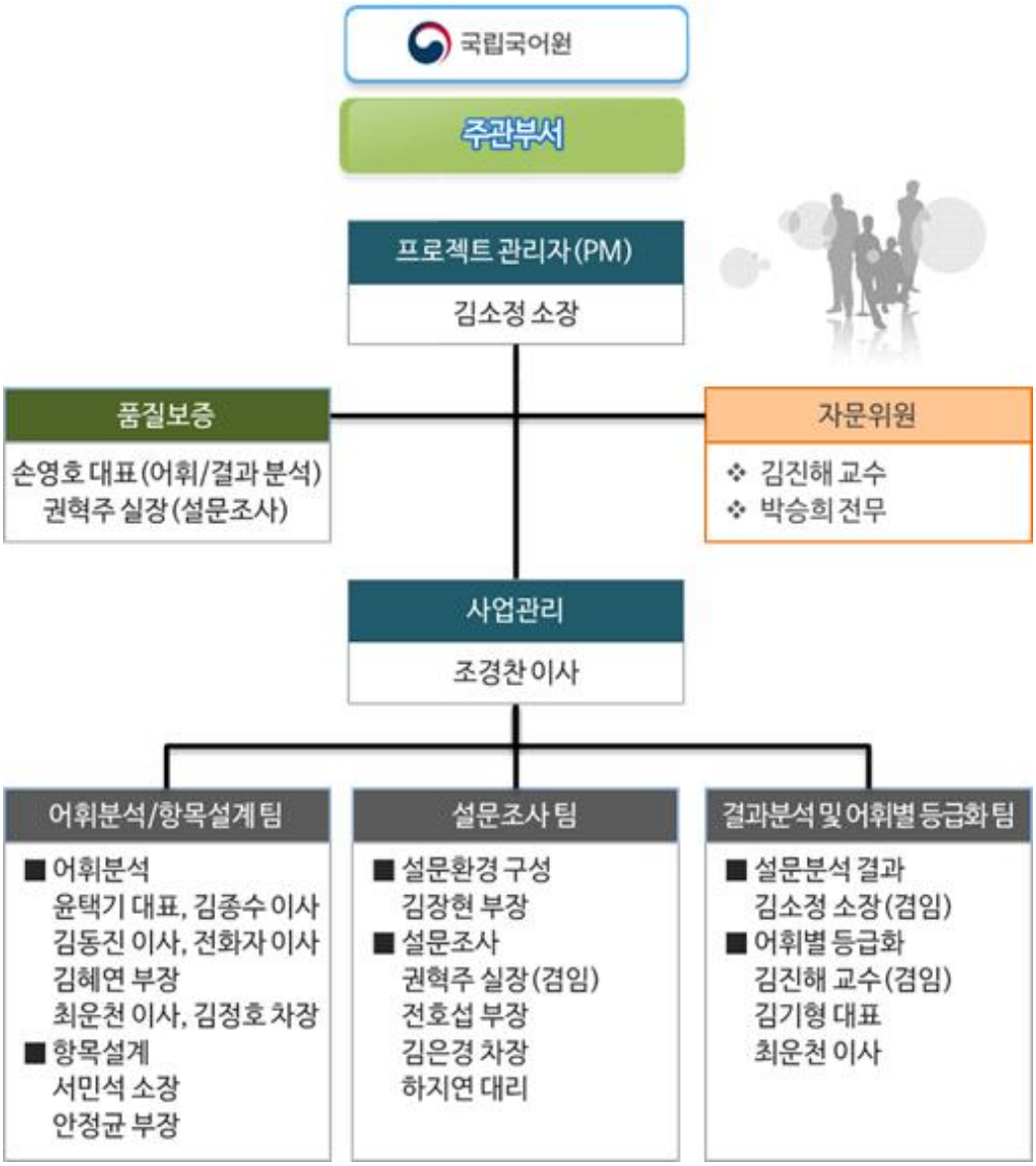


나. 기관별 역할 분담

구분	기관명	역할
주관 기관	국립국어원 (언어정보과)	• 사업 총괄 관리 및 시행 • 사업 계획 수립 및 제안 요청서 작성 • 주관 사업자 선정 및 계약 업무 추진 • 과제 수행 관리 및 검사, 결과 평가
주관 사업자	㈜나라지식정보	• 사업 수행 및 결과 보고 • 사업 진도 및 인력 관리 • 교육 및 기술 지원, 하자 보수, 유지 관리 • 기타 과제 수행에 필요한 사항 등

2. 사업 수행 추진 체계

가. 추진 체계도



나. 역할 및 책임

조직		역할 분담
사업 관리	용역 책임자	• 사업 책임 총괄 • 사업 추진 총괄 및 계약상의 사업 책임 총괄
	총괄PM	• 프로젝트 총괄 진행, 프로젝트 추진 활동 계획 조정 및 감독 • 각종 보고회의 준비 및 진행 등 체계적 의사소통 관리 활동 • 프로젝트 예산 관리 및 집행 • 투입 인력에 대한 계획 및 운영 관리 • 일정 및 진척 관리, 위험 요소 관리 • 세부 추진 실적 보고 및 납품
	사업관리	• 프로젝트 계획 및 운영 • 투입 인력 관리 및 사업 관리
자문 위원		• 항목 설계 및 설문 결과 분석 • 어휘별 등급화 자문
어휘 분석/ 항목 설계	어휘 분석팀	• 어휘 분석 방법론 수립 • 정교한 설문 조사를 위한 어휘별 난이도 부여 및 유형 분류 및 설문 분석 지원
	항목 설계팀	• 항목 설계 및 작성 • 시험 공정
설문 조사	설문환경 구성팀	• 웹 설문기획 및 제작
	설문 조사팀	• 설문 조사 진행 관리 및 조사 진행
결과 분석	설문 결과 분석팀	• 한국어 분산 의미 모델 평가용 어휘 관계 척도 수립 및 분석 및 평가 점수 산출
	어휘별 등급화팀	• 전체 평가 항목의 어휘별 등급화 • 통합 관리/설문 유형 분류 및 설문 분석 지원
품질 보증		• 어휘 분석/설계/결과 분석 • 설문 조사 및 운영

다. 참여 인력 총괄표

담당 업무		성명	소속	직위	비고
사업 관리	용역 책임자	손영호	㈜나라지식정보	대표이사	
	총괄PM	김소정	㈜나라지식정보	소장	
	사업 관리	조경찬	㈜나라지식정보	이사	
자문 위원		김진해	경희대학교	교수	
		박승희	㈜나라아이넷	전무	
어휘 분석/ 항목 설계	어휘 분석	윤택기	이지메타(주)	대표	
		김종수	이지메타(주)	이사	
		김동진	㈜나라지식정보	이사	
		전화자	㈜나라지식정보	이사	
		김혜연	㈜나라지식정보	부장	
		최운천	㈜낱말	이사	
		김정호	㈜낱말	차장	
	항목 설계	서민석	이지메타(주)	소장	
		안정균	이지메타(주)	부장	
설문 조사	환경 구성	김장현	㈜아이알씨	부장	
	설문 조사	권혁주	㈜아이알씨	실장	
		전호섭	㈜아이알씨	부장	
		김은경	㈜아이알씨	차장	
		하지연	㈜아이알씨	대리	
결과 분석	설문 결과 분석	김소정	㈜나라지식정보	소장	겸임
	어휘별 등급화	김진해	경희대학교	교수	겸임
		김기형	㈜낱말	대표	
		최운천	㈜낱말	이사	겸임
품질 보증		손영호	㈜나라지식정보	대표	겸임
		권혁주	㈜아이알씨	실장	겸임

3. 사업 수행 추진 경과

가. 사업 준비/보고

○ 제안 입찰(입찰 공고 번호: 20161033175-00) “한국어 정보 처리를 위한 어휘 관계 기초 자료 구축”
○ 계약 체결(계약 번호: 12198241200): 2019.11.20.
○ 착수계 제출: 2019.12.02.
○ 착수 보고회: 2019.12.19.
○ 중간 보고회: 2020.01.21.
○ 완료 보고회: 2020.02.18.
○ 정기 보고: 매주 금요일 주간 보고
○ 수시 보고: 안건 중심의 현황 보고(2019.12.02. ~ 2020.02.19.)

나. 사업 진행

○ 어휘 의미관계 자료 인수(30만 쌍): 2019.12.03.
○ 시험(파일럿) 공정: 2019.12.16. ~ 2019.12.20.
○ 어휘 분석: 2019.12.20. ~ 2020.01.16
○ 설문 항목 설계: 2019.12.23. ~ 2020.01.20.
○ 설문 조사: 2019.12.26. ~ 2020.01.23
○ 설문 결과 취합: 2019.12.30. ~ 2020.02.06.
○ 설문 결과 분석: 2020.01.13. ~ 2020.02.13
○ 추가 제안 설문 조사: 2020.02.05. ~ 2020.02.10.
○ 추가 제안 설문 결과 취합 및 분석: 2020.02.11. ~ 2020.02.17.
○ 어휘별 등급화 완료: 2020.02.18.

다. 사업 완료

○ 사업 준공: 2020.02.20.
○ 유지 보수: 2020.02.21. ~ 2021.02.20.

Ⅳ. 과업 실적

1. 과업 실적 현황

가. 공정 달성 및 주간 단위 실적

1) 공정 달성 현황

전체 공정률	150.5%	설계 단계	101.5%
		설문 단계	150.0%
		분석 단계	200.0%

공정 단계	업무 구분	과업 물량	누적 실적	진척률
설계 단계	어휘 분석	200,000	206,000	103.0%
	설문 항목 설계	200,000	200,000	100.0%
설문 단계	설문 조사	200,000	200,000	100.0%
	설문 결과 취합	100,000	200,000	200.0%
분석 단계	설문 결과 분석	100,000	200,000	200.0%

2) 주간 단위 실적 현황

공정단계	업무구분	과업 물량	1주차	2주차	3주차
설계 단계	어휘 분석	200,000	40,000	45,000	40,000
	설문 항목 설계	200,000	40,000	38,000	48,000
설문 단계	설문 조사	200,000	40,000	38,000	48,000
	설문 결과 취합	100,000		4,000	5,000
분석 단계	설문 결과 분석	100,000			

4주차	5주차	6주차	7주차	8주차	누계
81,000	-	-	-	-	206,000
40,000	34,000	-	-	-	200,000
40,000	34,000	-	-	-	200,000
32,000	20,000	89,000	50,000	-	200,000
10,000	31,000	-	85,006	73,994	200,000

나. 공정 작업 실적

1) 어휘 분석 실적 현황

구분	과업 물량	비슷한말-반대말	상위어	하위어	누계
		70,000쌍	70,000쌍	60,000쌍	
어휘 분석	200,000	72,000	72,000	62,000	206,000
		102.9%	102.9%	103.3%	103.0%

2) 설문 항목 설계 실적 현황

구분	과업 물량	비슷한말-반대말	상위어	하위어	누계
		70,000쌍	70,000쌍	60,000쌍	
설문 항목 설계	200,000	70,000	70,000	60,000	200,000
		100.0%	100.0%	100.0%	100.0%

3) 설문 조사 실적 현황

구분	과업 물량	비슷한말-반대말	상위어	하위어	누계
		70,000쌍	70,000쌍	60,000쌍	
설문 조사	200,000	70,000	70,000	60,000	200,000
		100.0%	100.0%	100.0%	100.0%

4) 설문 결과 취합 실적 현황

구분	과업 물량	비슷한말-반대말	상위어	하위어	누계
		70,000쌍	70,000쌍	60,000쌍	
설문 조사	100,000	70,000	70,000	60,000	200,000
		100.0%	100.0%	100.0%	200.0%

5) 설문 결과 분석 실적 현황

구분	과업 물량	비슷한말-반대말	상위어	하위어	누계
		70,000쌍	70,000쌍	60,000쌍	
설문 결과 분석	100,000	70,000	70,000	60,000	200,000
		100.0%	100.0%	100.0%	200.0%

2. 어휘 분석

어휘 분석 단계는 본 사업을 위해 국립국어원에서 제공한 자료를 분석/정리하고, 설문에 사용할 각종 어휘쌍을 준비하며, 필요한 데이터 세트를 정의하는 작업을 수행한다. 이 단계에서 처리할 중요한 작업은 다음과 같다.

- **국립국어원 제공 33만 어휘쌍 분석 및 정리**
- **설문에 사용할 20만 어휘쌍 선별**
- **선별된 20만 어휘쌍에 대한 예문이나 뜻풀이 추가**
- **선별된 20만 어휘쌍을 설문 조사를 위해 세트 분리**

가. 국립국어원 제공 33만 어휘쌍 분석 및 정리

1) 국립국어원 제공 33만 어휘쌍 분석

국립국어원에서 제공 받은 파일에는 총 329,696개의 어휘쌍이 있고 형태는 <표 1>과 같다(일부 항목만 제시). 품사, 원어, 전문 분야 정보 등 많은 정보를 포함하고 있지만, 본 사업에서 꼭 필요한 정보는 표제어와 관련어(관련 대응 표제어), 관련어 정보 3가지이다. 하나의 어휘쌍은 '표제어-의미 번호'로 구분된 표제어 부분과 '관련 대응 표제어-의미 번호'로 구분되는 관련어 부분으로 되어 있고, 두 단어의 관계를 나타내는 관련어 정보는 '관련어 정보-관련어 상세 정보'를 이용하여 파악한다.

일련 번호	표제어	의미 번호	품사	원어	전문 분야	관련어 정보	관련어 상세 정보	관련 대응 표제어	의미 번호	…
26554	안-벽	001	명사	안壁	『건설』	반대말		외벽	001	
54308	지상-층	001	명사	地上層	『건설』	반대말		지하-층	001	
35218	후-결제	001	명사	後決濟	『경영』	참고 어휘	대립어	선-결제	001	
24122	절하-하다	001	동사	切下하다	『경제』	반대말		절상-하다	001	
14282	반락	001	명사	反落	『경제』	반대말		반등	001	
25984	호재	002	명사	好材	『경제』	반대말		악재	001	

〈표 1〉 국립국어원에서 보내온 파일의 형태

본 사업의 목표인 비슷한말/반대말, 상위어, 하위어 3개 그룹에 대해 20만 어휘쌍을 구축하기 위해서

는 3개 관련어 그룹이 필요하지만, 비슷한말과 반대말은 그 성격이 달라서 설문 문항을 만들 때 구분해야 하기 때문에 둘로 나눈다. 그래서 비슷한말, 반대말, 상위어, 하위어 4가지로 구분한다. <표 2>의 참고 어휘 중 관련어 상세 정보가 대립어로 표시된 것들은 반대말에 포함시킨다. <표 2>는 4가지 관련어에 대한 어휘쌍의 수를 보여준다.

관련어 정보	어휘쌍 수	관련어 그룹	어휘쌍 수
반대말	5,334	반대말	13,390
참고어휘	8,056		
비슷한말	118,514	비슷한말	118,514
상위어	114,580	상위어	114,580
하위어	83,212	하위어	83,212
총합계	329,696		329,696

〈표 2〉 33만 어휘쌍의 관련어별 어휘쌍 수

향후 공정을 위해 원본 데이터 파일에 4개의 새로운 행을 추가하였다. 새로 추가된 4가지 요소(<표 3>)는 향후 데이터베이스 작업에 중요한 역할을 하는 것들이다. 33만 어휘쌍에 등장하는 표제어와 관련어는 한데 모아서 별도의 데이터베이스(DB)로 만들고 고유 번호(ID)를 부여하였다.

행 이름	의미
pair_ID	33만 어휘쌍의 일련번호
구분2	관련어의 4가지 그룹
표제어_word_ID	표제어의 고유 번호(ID)
관련어_word_ID	관련어의 고유 번호(ID)

〈표 3〉 새로 추가한 행의 이름과 의미

2) 33만 어휘쌍에 포함된 단어 정리

향후 공정의 효율적인 처리를 위해 33만 어휘쌍에 등장하는 단어들을 정리하고, 설문 문항 작성에 필요한 정보들을 추가한다. 표제어와 관련어에 나타나는 어휘를 모두 모아 중복을 제거한 후 고유 번호(word_ID)를 부가한다. 아울러 이해를 돕기 위해 간략한 단어 형태를 정의(word name)하고 별도의 테이블로 관리한다(<표 4>). 33만 어휘쌍에 등장하는 고유 단어는 모두 219,251개이다. ‘word_ID’는 가나다순으로 정렬한 후에 순차적으로 부여하였다. ‘word_name’은 ‘단어+의미 번호’ 형태로 단어 내에 포함된 기호를 모두 제거한 상태로 사용한다.

word name	word_ID
ㄱ001	1
ㄱㄴㄷ순001	2
ㄱㄴㄷ차례001	3
ㄱㄴ순001	4
ㄱㄴ차례001	5
…	…
ㅟ001	219247
ㅠ001	219248
ㅡ001	219249
ㅢ001	219250
ㅣ001	219251

〈표 4〉 단어별 고유 번호(ID)

3) 단어별 뜻풀이와 예문 추가

설문 조사에 제시할 예문과 뜻풀이는 국립국어원 누리집에서 『우리말샘』 사전(2019년 10월 9일자 엑셀 파일)을 다운받아 사용하였다. 다운받은 엑셀 파일을 하나로 정리하고, <표 4>의 'word name'과 동일한 형태로 표제어를 변환한 후 뜻풀이를 가져와 단어 파일에 추가하였다. 설문 조사에 사용할 목적이라 뜻풀이는 가능한 부가 정보를 제외하고, 1 ~ 2문장만 가져왔다. 이 과정에서 최신 『우리말샘』 사전에서 삭제된 단어가 발견되어(<표 5>) 그것이 포함된 어휘쌍은 제외하기로 했다.

word_ID	word name
13474	고모001
13478	고모부001
13480	고모부003
26340	글루코시다아제002
26811	금사003
32621	나흗날002
34099	남자배우001
42805	닷샛날002

77618	보리잎벌001
81394	부책001
82559	분장청회사기001
118915	아흐렛날002
127533	여드렛날002
129700	연우궁001
130641	열흘날002
147693	이렛날002
191988	큰언니002
191994	큰오빠002
196804	투탈001

〈표 5〉『우리말샘』에서 삭제된 단어들

『우리말샘』 사전에 예문이 있는 경우는 뜻풀이와 같은 방법으로 가져와서 별도의 행으로 추가하였다. 예문이 2개 이상 있는 경우는 첫 번째 나오는 예문만 사용하였다. 전체 단어 중 『우리말샘』에 예문이 있는 경우는 31%이다(<표 6> 참조).

구분	단어 수	비율
예문 있음	67,625	30.8%
예문 없음	151,627	69.2%
전체	219,252	

〈표 6〉『우리말샘』에서 찾은 단어별 예문 유무

<표 7>은 『우리말샘』 사전에서 가져온 정보를 토대로 구축한 테이블을 예로 보여준 것이다. 예문 행이 비어 있는 경우는 『우리말샘』에 해당 표제어에 대한 예문이 없는 경우이다. 예문을 추가할 때 해당 단어가 포함된 어절을 '{ }'로 묶어서 설문지 작성 시 활용하도록 한다.

word_ID	word name	예문	뜻풀이
14	가001		경계에 가까운 바깥쪽 부분.
15	가002		어떤 중심 되는 곳에서 가까운 부분.
16	가003	참기름을 따를 때 {가에} 흘리지 않도	그릇 따위의 아가리의 주변.

		록 조심해라.	
17	가004	{강가}.	'주변'의 뜻을 나타내는 말.
18	가005		서양 음악의 칠음 체계에서, 여섯 번째 음이름. 계이름 '라'와 같다.
19	가008	이 사람 말도 {가요}, 저 사람 말도 {가요} 하면 도대체 어떤 사람 말을 따라야 합니까?	옳거나 좋음.
20	가009	의원 여러분께서는 본 안건에 대해 {가인지} 부인지를 결정해 주시기 바랍니다.	회의 따위에서, 어떤 안건에 대하여 표결을 할 때 찬성하는 의사 표시.
21	가010	다른 과목들은 성적이 괜찮은 편인데, 체육만 {가를} 받았다.	성적이나 등급을 '수, 우, 미, 양, 가'의 다섯 단계로 나눌 때 가장 낮은 단계.
22	가011	연소자 관람 {가}.	어떤 행위가 허용되거나 가능함 또는 좋음을 이르는 말.
23	가012		죄인에게 씌우던 형틀. 두껍고 긴 널빤지의 한끝에 구멍을 뚫어 죄인의 목을 끼우고 비녀장을 질렀다.
24	가013	성구는 그 대목에서 묘하게 처절해지는 버릇이 있었다. 외할머니 교하댁의 집에 대한 소문난 집착을 {가를} 잇고자 하는 맹목적 집념과 동일시하려는 그 나름의 시각 때문이었다.	예전에, 같은 호적에 들어 있는 친족 집단을 이르던 말.
25	가가004	소위 이러한 {'가가'는} '난전'이라 해서 불법화되어 있었으며….	'가게'의 원말.
26	가가대소001	위원장은 무엇이 그렇게 우스운지, 세수를 마칠 생각은 않고, 들창코를 벌름거리며 {가가대소를} 해 댔다.	소리를 내어 크게 웃음.

〈표 7〉 단어별 뜻풀이와 예문 추가

4) 단어별 어휘 난이도 정보 추가

단어별 어휘 난이도는 정교한 설문 조사를 위해 사용한다. 또한 어휘별 등급화를 위한 기초 자료로 사용할 수 있다. 어휘 난이도 정보는 ㈜낱말의 어휘 난이도 1등급에서 7등급 체계를 사용한다 (https://natmal.com/views/dictionary/syn 및 <표 8> 참조).

어휘 난이도가 부여된 단어는 20만 정도이고, 난이도가 없는 단어는 등급 8로 하여 구분하였다.

word name	낱말_난이도(8등급)
가게001	1
가게002	1
가까이001	1
…	.
가겟집001	2
가겟집002	2
가격003	2
…	.
가감례001	7
가감승합제001	7
가감역001	7
…	.
히페리온003	8
힐로002	8
힝이002	8
힝힝하다002	8

〈표 8〉 ㈜낱말의 어휘 난이도 예시

㈜낱말의 난이도 정보를 이번 사업에 사용하기 위해서는 ㈜낱말의 국어사전과 『우리말샘』의 표제어를 의미 번호 차원에서 연결시켜야 한다. ㈜낱말의 국어사전은 『표준국어대사전』(2013년 10월)의 표제어를 사용하고 있다. 이번 사업을 위해 국립국어원에서 『표준국어대사전』(2019년)과 『우리말샘』 표제어 간 변환테이블을 제공하였다. ㈜낱말의 국어사전이 『표준국어대사전』 2013년 기준이고, 국립국어원에서 보내준 자료가 2019년판 『표준국어대사전』 기준이라 ㈜낱말의 국어사전을 먼저 『표준국어대사전』(2019)에 연결시킨 다음에 그 결과를 『우리말샘』과 연결시켰다.

㈜낱말 국어사전 표제어 ⇨ 『표준국어대사전』2019 표제어 ⇨ 『**우리말샘**』 표제어

여기서 고려해야 할 것은 『표준국어대사전』과 『우리말샘』이 표제어를 나누는 기준이 다르다는 것이다. 『표준국어대사전』은 먼저 어깨번호로 구분한 다음에 의미별로 구분하지만, 『우리말샘』은 바로 의미별로 구분한다. ㈜낱말의 난이도는 『표준국어대사전』의 어깨번호 수준으로 부여되어 있어 『우리말샘』과

1:1 대응이 되지 않는 경우도 있다. 즉, ㈜낱말의 단어 하나가 『우리말샘』에서는 2개 이상의 의미와 연결될 경우 『우리말샘』의 모든 의미에 동일한 난이도를 부여하였다. <표 9>에서 ㈜낱말 국어사전의 '가늘다'의 난이도는 1등급이다. 이 경우 『우리말샘』의 7개의 '가늘다' 모두 난이도 1등급이 된다. ㈜낱말 국어사전의 '소비04'의 난이도는 2등급이다. 이 경우 그에 대응하는 『우리말샘』의 '소비(006)'과 '소비(007)'은 모두 난이도 2등급이 된다. 나머지 '소비01, 소비02, 소비03, 소비05, 소비06'은 난이도 등급이 없기 때문에 대응하는 『우리말샘』의 '소비(001), 소비(002), 소비(003), 소비(004), 소비(005), 소비(008), 소비(009),'는 모두 난이도 등급 8을 부여한다.

『표준국어대사전』	『우리말샘』	『우리말샘』_뜻풀이
가늘다	가늘다001	물체의 지름이 보통의 경우에 미치지 못하고 짧다.
	가늘다002	소리의 울림이 보통에 미치지 못하고 약하다.
	가늘다003	물체의 굵기가 보통에 미치지 못하고 잘다.
	가늘다004	빛이나 연기 따위가 희미하고 약하다.
	가늘다005	표정이 얼굴에 나타날 듯 말 듯 약하다.
	가늘다006	사이가 좁고 촘촘하다.
	가늘다007	움직이는 정도가 아주 약하다.
소비01	소비001	'보습'의 방언(강원, 경북).
	소비002	비(妃)가남편인왕을상대하여자기를낮추어이르던일인칭대명사.
소비02	소비003	나이 어린 계집종.
	소비004	계집종이 상전을 상대하여 자기를 낮추어 이르던 일인칭 대명사.
소비03	소비005	일에 든 비용.
소비04	소비006	돈이나 물자, 시간, 노력 따위를 들이거나 써서 없앰.
	소비007	욕망을 충족하기 위하여 재화나 용역을 소모하는 일. 본래적 소비와 생산적 소비가 있다.
소비05	소비008	벌이 알을 낳고 먹이와 꿀을 저장하며 생활하는 집.
소비06	소비009	상소(上疏)에 대하여 임금이 내리던 대답.

〈표 9〉『표준국어대사전』과 『우리말샘』의 표제어 차이

본 사업에서는 ㈜낱말의 어휘 난이도를 토대로 난이도 등급을 상중하 3개로 구분하여 사용한다(<표 10>).

쉬운 단어가 '하'이고 어려운 단어가 '상'이 된다. 어휘 난이도 1~8등급 체계는 설문 세트를 구성할 때와 설문 세트 내 어휘쌍을 쉬운 어휘부터 어려운 어휘순으로 배열할 때 사용한다. 단, ㈜낱말의 어휘 난이도 비공개 원칙에 따라 본 사업의 중간 결과로는 '상중하'로 변경한 난이도 등급 형태로 제공한다.

구분	낱말난이도	난이도 등급
쉬운 단어	1	하
	2	
	3	중
	4	
	5	
	6	상
어려운 단어	7	
등급 외	8	

〈표 10〉 ㈜낱말의 어휘 난이도를 이용한 등급화

나. 설문에 사용할 20만 어휘쌍 선별

다음 과정으로 국립국어원에서 제공 받은 33만 어휘쌍으로 본 사업에서 목표로 하는 20만 어휘쌍을 선별하는 작업이 필요하다. 20만 어휘쌍은 비슷한말/반대말 7만, 상위어 6만, 하위어 7만을 목표로 한다. 33만 어휘쌍에는 표제어와 관련어가 상위어 관계인 경우와 그 반대인 하위어 관계인 경우도 모두 존재하는 어휘쌍이 다수 존재하여 둘 중 하나만 선별하기로 했다. 시험 공정(파일럿 테스트)를 거치면서 동일한 어휘쌍에 대해 설문자에게 상위어인지를 물어보는 것이 하위어를 물어보는 것보다 더 낫다는 내부 의견에 따라 국립국어원의 동의를 거쳐서 상위어 7만, 하위어 6만 어휘쌍을 선별하는 것으로 수정하였다.

1) 제외 대상 어휘쌍 선별

『우리말샘』 최신 버전에서 삭제된 단어 19개(<표 5> 참조)가 포함된 어휘쌍은 제외한다. 제외된 어휘쌍은 <표 11>과 같다.

pair_ID	표제어	의미 번호	관련어 정보	관련어 대응 표제어	의미 번호
18265	보리주	001	비슷한말	보리-잎벌	001
30576	금니	002	상위어	금사	003
30799	이금	005	상위어	금사	003
83014	칠궁	001	하위어	연우-궁	001
119797	글루코시데이스	001	비슷한말	글루코시다아제	002

138861	고숙-주	001	비슷한말	고모-부	003
176777	남-배우	001	비슷한말	남자 배우	001
192581	고모-님	001	상위어	고모	001
192585	고모부-님	001	상위어	고모-부	001
202885	대고모-님	001	상위어	고모	001
212789	백-고모	001	상위어	고모	001
224863	시-고모	001	상위어	고모	001
245250	중-고모	001	상위어	고모	001
261736	작은-언니	001	참고 어휘	큰-언니	002
261739	작은-오빠	001	참고 어휘	큰-오빠	002
262542	가루	001	하위어	금사	003
273058	날	001	하위어	여드렛-날	002
273066	날	001	하위어	열흘-날	002
273088	날	001	하위어	닷샛-날	002
273092	날	001	하위어	이렛-날	002
273109	날	001	하위어	나흗-날	002
273112	날	001	하위어	아흐렛-날	002
275074	누이	001	하위어	고모	001
277074	도분	001	하위어	투탈	001
289802	빚	001	하위어	부책	001
310495	자기	014	하위어	분장^청회^사기	001

〈표 11〉 삭제된 단어가 포함된 어휘쌍

의미 번호를 제외한 표제어와 관련어가 동일한 경우는 제외한다. 설문 문항에 의미 번호는 안 보이기 때문에 설문자들이 같은 단어에 대해 질문을 한다고 오해할 가능성이 매우 높기 때문이다. 예를 들어 보면, “주교의 비슷한말은 주교입니까?”라는 형태의 설문 문항을 제시한다면 설문자들이 어떻게 받아들일까? 먼저 나온 ‘주교’는 ‘주교(002)’이고, 나중에 나온 ‘주교’는 ‘주교(003)’이다. 표제어와 관련어가 같은 단어인 어휘쌍은 모두 1,449개이며 그 중 일부를 보이면 <표 12>와 같다.

pair_ID	표제어	의미 번호	관련어 정보	관련어대응표제어	의미 번호
39	주교	002	비슷한말	주교	003
239	가톨릭교-회	002	상위어	가톨릭교-회	001
249	고백	002	상위어	고백	001
372	성심	003	상위어	성심	002
573	가톨릭교-회	001	하위어	가톨릭교-회	002
2153	건평	002	상위어	건평	001
2461	배치-도	003	상위어	배치-도	001
2781	종탑	002	상위어	종탑	001
2784	주도	005	상위어	주도	001
2829	천장	002	상위어	천장	003
3070	건평	001	하위어	건평	002
3162	천장	003	하위어	천장	002
3604	채산	003	상위어	채산	002
3697	매방-선택	001	반대말	매방-선택	002
4813	이피에스	002	비슷한말	이피에스	001
4883	결제	002	상위어	결제	001
4935	공급	003	상위어	공급	002
5030	노동	003	상위어	노동	004
5115	매주	004	상위어	매주	003
5284	상품	006	상위어	상품	005

〈표 12〉 표제어와 관련어가 동일한 어휘쌍

2) 역방향 존재 어휘쌍의 경우 한쪽만 선별

『우리말샘』의 자료에서 역방향 존재라는 것은 『우리말샘』의 어휘쌍이 일반적으로 '표제어 → 관련어 대응 표제어'로 대응된다고 할 때, 그것의 역방향인 '관련어 대응 표제어 → 표제어'도 존재하는 경우를 말한다. 일반적으로 비슷한말이나 반대말 관계는 단어의 방향성이 의미를 갖지 않는다고 본다. 하지만 상위어, 하위어 관계는 방향성을 가지고 있다.

<표 13>에서 역방향이 존재하는 단어들을 살펴보자. 반대말 관계의 경우 '경합성 → 비경합성'과 그 역방향인 '비경합성 → 경합성' 어휘쌍이 모두 존재하고, 비슷한말 관계의 경우 '혼잣소리 → 혼잣말'과

그 역방향인 '혼잣말 → 혼잣소리'가 모두 존재한다. 상위어와 하위어는 개념상 반대 개념이며 서로 역방향으로 관계가 맺어져 있다. '독백 → 대사'는 상위어 관계로 '독백'의 상위어가 '대사'인데, 역방향은 '대사 → 독백'으로 하위어 관계이다. 즉, '대사'의 하위어가 '독백'이라는 의미이다.

역방향이 존재하는 어휘쌍의 경우는 주관 기관과 협의하여 역방향이 존재하는 경우는 둘 중 하나만 선택하는 것으로 결정하였다.

pair_ID	표제어	의미 번호	관계	관련어 대응 표제어	의미 번호
5927	경합-성	001	반대말	비-경합성	001
6026	비-경합성	001	반대말	경합-성	001
6041	선-결제	001	반대말	후-결제	001
3658	후-결제	001	반대말	선-결제	001
6119	자-은행	001	반대말	모-은행	001
6006	모-은행	001	반대말	자-은행	001
186593	혼잣-소리	001	비슷한말	혼잣-말	001
166035	혼잣-말	001	비슷한말	혼잣-소리	001
166036	혼잣-말	001	비슷한말	독백	001
177825	독백	001	비슷한말	혼잣-말	001
181665	승강-기	001	비슷한말	엘리베이터	001
154851	엘리베이터	001	비슷한말	승강-기	001
83310	독백	002	상위어	대사	023
83949	대사	023	하위어	독백	002
83311	독백	002	상위어	행위	001
326213	행위	001	하위어	독백	002
25644	독백-체	001	상위어	문체	002
26481	문체	002	하위어	독백-체	001

〈표 13〉 역방향 존재 어휘쌍

3) 관련어 그룹별 어휘쌍을 고려한 선별

역방향이 존재하는 어휘쌍의 수는 <표 14>과 같다. 서로 반대 개념인 상위어/하위어 관계인 경우 역방향이 존재하는 62,454개에 대해 상위어를 선별할 경우, 그에 대응하는 하위어는 제외해야 한다. 하위어를 선별할 경우에도 그에 대응하는 상위어를 제외해야 한다. 이때 목표로 하는 상위어 7만, 하위어 6

만을 충족하도록 선별해야 한다.

구분	어휘쌍 수
반대말	5,254
비슷한말	78
상위어	62,454
하위어	62,454
총합계	130,240

〈표 14〉 역방향이 존재하는 어휘쌍의 수

상위어/하위어 모두 양방향으로 어휘관계가 맺어져 있기 때문에 어느 쪽을 우선하여 설문하여도 큰 문제는 없다. 그러나 설문자의 입장에서 보다 정확히 답변할 가능성을 높이는 방법은 하위어를 묻는 것보다 상위어를 묻는 것이라는 논의 결과에 따라 상위어를 우선으로 하여 설문을 실시하였다. 이에 따라 최종적으로 선정된 어휘쌍은 <표 15>와 같다.

구분	어휘쌍 수
비슷한말/반대말	70,000
상위어	70,000
하위어	60,000
총합계	200,000

〈표 15〉 작업에 필요한 어휘쌍의 수

4) 어휘 난이도를 고려한 선별

<표 15>에서 정한 목표 어휘쌍을 위해 아래와 같은 원칙으로 선별하였다.

- 표제어와 관련어 모두 예문이 있는 경우를 우선한다.
- 어휘 난이도가 '하/중'인 경우를 우선한다.
- 표제어와 관련어 중 어휘 난이도가 다를 경우 더 낮은 어휘 난이도를 가진 것을 우선한다. 예를 들어, 표제어의 어휘 난이도가 '하'이고 관련어의 어휘 난이도가 '상'인 경우라면 이 어휘쌍을 설문 대상으로 선정한다.

위와 같은 방법으로 1차 선별을 실시한 다음, <표15>에서 정한 어휘 관계별 어휘쌍 수에 부합하도록 조정한다. 실제 작업 과정에서 상위어는 7만 개보다 적고, 하위어는 6만 개보다 많아서 수작업으로 상

위어와 하위어가 역방향 관계에 있는 어휘쌍의 선정 여부를 뒤바꾸어 <표 15>에서 제시한 어휘쌍 수에 부합하도록 하였다.

최종 선정 결과는 <표 16>에서 보인 것처럼 역방향이 존재하는 65,120개의 어휘쌍을 제외한 264,576개이다. 여기에 앞서 언급한 제외 대상 어휘쌍을 추가로 제외해야 한다.

구분	전체 어휘쌍	단방향 어휘쌍	역방향 선정	역방향 제외	선정된 어휘쌍
반대말	13,390	8,136	2,627	2,627	10,763
비슷한말	118,514	18,436	39	39	118,475
상위어	114,580	52,126	20,646	41,808	72,772
하위어	83,212	20,758	41,808	20,646	62,566
총합계	329,696	199,456	65,120	65,120	264,576

〈표 16〉 역방향이 존재하는 경우를 고려하여 최종 선정한 어휘쌍의 수

이 외에도 상위어/하위어 관계가 순환적으로 맺어진 오류들을 제외하였는데, 『우리말샘』에서 확인할 수 있는 순환 연결의 사례는 아래 표와 같다.

pair_ID	표제어	의미 번호	품사	관련어 대응 표제어	의미 번호	구분
가르침(001)〈지도(011)〈교육(001)〈가르침(001)						상위어
187982	가르침	001	명사	지도	011	
246040	지도	011	명사	교육	001	
194928	교육	001	명사	가르침	001	
가식(001) 〈 거짓(001) 〈 가식(001)						
188200	가식	001	명사	거짓	001	
190363	거짓	001	명사	가식	001	
가입(001)〈참가(001)〈가입(001)						
188308	가입	001	명사	참가	001	
247728	참가	001	명사	가입	001	
수월(005) 〉 누월(001) 〉 수월(005)						하위어
296156	수월	005	명사	누월	001	
275064	누월	001	명사	수월	005	
누승(001) 〉 제곱(001) 〉 누승(001)						
56573	누승	001	명사	제곱	001	

57257	제곱	001	명사	누승	001	
제곱(001) 〉 승멱(002) 〉 제곱(001)						
57260	제곱	001	명사	승멱	002	
57077	승멱	002	명사	제곱	001	

〈표 17〉 계층구조가 이상한 상위어/하위어의 예(일부)

다. 선별된 20만 어휘쌍에 대한 예문 추가

사업의 성격상 뜻풀이보다는 예문을 보여주는 것이 평가에 도움이 되므로 『우리말샘』 사전에 예문이 없는 경우 별도로 예문을 수집하여 추가하기로 하였다. 『우리말샘』의 표제어는 의미별로 구분되어 있기 때문에 개별 의미에 맞는 예문을 찾는 작업은 시간이 많이 걸리고 어렵다. 그 분야는 프로그램을 이용하여 해결할 수 없어 사람이 문장 하나하나를 보고 해당 단어의 의미(뜻풀이)에 일치하는 예문인지를 판단해야 한다. 어휘 난이도가 높은 경우는 잘 사용하지 않는 단어이기 때문에 예문을 찾기가 매우 어려워 새로 만드는 것이 나을 경우도 있다. 이렇게 만든 예문을 일반인이 보고 어떤 의미인지 한번에 알아보기란 어려울 수도 있다.

이렇듯 해당 어휘의 쓰임이 드문 경우에는 그에 적당한 예문을 찾는 것이 쉽지 않다. 만약 해당 예문을 임의로 작성하여 제시한다고 하여도 피평가자가 이미 익숙한 의미인 '가무005'를 먼저 생각하다 보니 '가무005'가 아닌 다른 의미로 쓰인 것이라는 사실을 설문 조사 순간에 알아차리기가 쉽지 않다는 현실적인 문제를 안고 있다(참고로 <표 18>의 '등장 회수'는 해당 단어가 33만 어휘쌍에 나타난 회수를 의미한다).

본 사업에서는 이런 문제가 있다는 것을 파악하고 최대한 예문을 찾아서 사용하고, 예문을 찾기 어려운 경우에는 뜻풀이로 대신하기로 한다. 매주 설문 조사에 사용할 어휘쌍을 선정하고, 그 어휘쌍에 포함된 표제어를 기준으로 예문이 없는 경우에 찾아서 보충하는 방식으로 사업을 진행한다.

word name	뜻풀이	예문 여부	예문	등장 회수
가무002	살림을 꾸려 나가면서 하여야 하는 여러 가지 일. 빨래, 밥하기, 청소 따위를 이른다.			5
가무003	자기 집이나 가까운 친척 집에 생기는 일이나 행사.			4
가무004	신라 때에, 피리에 맞추어 춤을 추며 노래하던 무악(舞樂). 내물왕 때 향인(鄕人)들이 지어 즐겼다고 하는데,			3

	가사와 악보는 전하지 않는다.			
가무005	노래와 춤을 아울러 이르는 말.	예문 있음	{가무에} 능한 기생.	6
가무006	노래하면서 춤을 춤.			1

〈표 18〉 단어 '가무'에 대한 뜻풀이와 예문

1) 다른 국어사전의 예문 이용

『우리말샘』에 예문이 존재하지 않는 경우 컨소시엄 내 자료 공유를 통해 얻은 A국어사전의 예문을 사용하였다. A국어사전의 예문을 사용하기 위해서는 『우리말샘』 표제어의 의미와 A국어사전의 의미가 같거나 비슷한 것을 찾아야 한다. 이 작업은 어깨번호나 의미 번호를 무시하고 단어 수준에서 『우리말샘』과 A국어사전의 표제어를 비교하여 동일한 표제어에 대해 『우리말샘』과 A국어사전의 뜻풀이를 비교하여 뜻풀이가 같거나 비슷한 경우 A국어사전의 예문을 사용하였다. 뜻풀이 비교는 두 사전의 단어별 뜻풀이 비교 프로그램을 개발하여 A국어사전의 여러 뜻풀이 중 『우리말샘』 표제어의 뜻풀이와 가장 유사한 한 개만 제시한 후 이 결과를 보고 국어 전문가들이 수작업으로 예문 사용 여부를 판단하게 하였다.

뜻풀이가 하나인 경우는 두 사전의 뜻풀이만 비교하기 때문에 비교적 작업이 쉽다. 하지만 A국어사전의 뜻풀이가 2개 이상일 때는 어느 뜻풀이가 『우리말샘』 표제어의 뜻풀이와 일치하는지를 비교하여 결정해야 하기 때문에 시간과 노력이 많이 든다. <표 19>에서 『우리말샘』의 '개성(007)'은 A사전의 어느 뜻풀이와도 일치하지 않아서 예문을 국어 전문가가 직접 추가한 경우이다. '개성(009)'의 경우에는 뜻풀이가 일치하는 것이 없어 모두 '사용 불가'로 표시하였다. '경기(002)'의 경우는 A사전에 여러 뜻풀이 중 프로그램이 제시한 2개의 뜻풀이 중 하나를 사용하고 의미가 다른 것은 '사용 불가'로 표시하였다. '경기(010)'의 경우는 해당되는 뜻풀이가 없어서 국어 전문가가 직접 예문을 추가한 경우이다.

그리하여 프로그램이 찾은 1만3천여 개의 예문을 국어 전문가가 비교 검토하여 8천여 개 정도를 추가하였다

표제어	단어	『우리말샘』_뜻풀이	A사전_뜻풀이	A사전_예문	설명
개성007	개성	경기도 서북부에 있는 시. 인삼의 명산지이며, 예로부터 보부상이 유명하다.	① 한 개인이 가지는 고유한 취향이나 특성.	{개성}에는 남대문, 만월대, 선죽교, 숭양 서원 따위의 명승지가 있다.	직접 추가
개성009	개성	지속적인 동기의 경향이나 환경적 자극에 대하여 비교적 일관성 있는 행동 성향 및	① 한 개인이 가지는 고유한 취향이나 특성.	현대는{개성}과 자아가 존중되는 세계이다	사용 불가

		반응을 일으키는 개인의 심리적 특성.			
개성009	개성	지속적인 동기의 경향이나 환경적 자극에 대하여 비교적 일관성 있는 행동 성향 및 반응을 일으키는 개인의 심리적 특성.	②각 개체의 특성.	사물의 {개성}을 강조하다.	사용 불가
경기002	경기	서울을 중심으로 한 가까운 주위의 지방.	②서울을 중심으로 하여 그 가까운 지역을 이르는 말.	{경기} 지역에 분포되어 있는 탈놀이를 산대 탈놀이라 한다.	사용
경기002	경기	서울을 중심으로 한 가까운 주위의 지방.	서울과 경기지방을 중심으로 불리는 민요. 장단은 주로 굿거리, 자진타령, 세마치등이 쓰인다.	서울 놀이마당에서는 '단오절 민속 예술 공연'을 주최하여 {경기}, 은율탈춤, 남사당 놀이 등 전통 민속놀이 공연을 펼친다.	사용 불가
경기010	경기	사고파는 사람 사이에 들어 흥정을 붙이는 일을 하는 사람.	서울과 경기지방을 중심으로 불리는 민요. 장단은 주로 굿거리, 자진타령, 세마치등이 쓰인다.	{경기}는 사고파는 사람 사이에 들어 흥정을 붙이는 일을 하는 사람.	직접 추가

〈표 19〉 다른 국어사전을 이용한 예문 추가

2) 20억 어절 말뭉치에서 찾은 예문 이용

『우리말샘』과 A국어사전에서 예문을 찾지 못한 경우는 ㈜낱말이 보유하고 있는 20억 어절 말뭉치 검색을 통해 예문을 찾았다. 말뭉치의 예문에는 긴 문장이 많은 편이라 해당 단어 중심으로 10~12어절만 나타나도록 예문 길이를 조정하였다. 이렇게 찾은 6만여 개의 예문을 국어 전문가가 『우리말샘』 뜻풀이와 비교하여 의미가 일치하는 예문만 찾아 추가하였다. 최종적으로 8천여 개의 예문이 추가되었다. 어휘 난이도가 낮아 누구나 아는 단어이거나 33만 어휘쌍에 등장한 횟수가 많은 단어가 예문이 없는 경우는 우선하여 직접 추가하였다(<표 20> 참조).

단어	표제어	예문	확인	등장 회수
곤충류	곤충류001	괭이갈매기는 잡식성으로 물고기 조개 {곤충류를} 잡아먹으며 때로는 배를 따라다니며 음식물	예문 사용	142
국화과	국화과001	꽃 피우는 것은 거의 모두 {국화과에} 딸린 식물들이다	예문 사용	102
단어	단어001	{단어와} {단어는} 띄어 써요	예문 사용	77
두해살이풀	두해살이풀001	가짓과의 한해살이풀 또는 {두해살이풀}	예문 사용	100
범죄인	범죄인001	사연도 있고 애꿎은 사람이 억울하게 {범죄인으로} 취급되기도 한다	예문 사용	78
볏과	볏과001	외떡잎식물인 {볏과}에는 개밀, 겨이삭, 쇠풀, 귀리 따위가 있다.	직접추가	65
불교인	불교인001	없었습니다 한국 비구니 및 여성 {불교인의} 위상이 제고되고 불교 발전의 중요한	예문 사용	91
애벌레	애벌레001	그동안 날도래 {애벌레를} 잡아먹으며 겨울을 났습니다	예문 사용	70
육십갑자	육십갑자001	천간과 12개의 지지로 이루어진 것이 {육십갑자다}	예문 사용	109
육십사괘	육십사괘001	하도와 낙서 선천팔괘 후천팔괘 음양오행 {육십사괘} 10간 28수를 통달해야만 비로소 접근할	예문 사용	127
육십화갑자	육십화갑자001	약 29.5일인 삭망월 주기에다 일진 {육십화갑자} 이십팔수 십이직을 조합해 점술적인 역주를	예문 사용	69
이십사시	이십사시001	{이십사시}는 하루를 스물넷으로 나누어 각각 이십사방위의 이름을 붙여 이르는 스물네 시이다.	직접추가	87
이십사절기	이십사절기001	매년 12삭의 크고 작은 것 {이십사절기의} 일시를 대통수로 추산하여 편찬하였다	예문 사용	69
전자기파	전자기파001	EMP란 Electro Magnetic Pulse의 약자로 {전자기파로} 인해 두근거리는 맥박처럼 짧은 시간에	예문 사용	68
포유동물	포유동물001	살고 있는 물 속에 사는 {포유동물은}	예문 사용	75
포유류	포유류001	곤충 어류 양서류 파충류 조류 {포유류} 등 갖가지 종류가 포함돼 있다	예문 사용	104

품사	품사001	유럽 언어에서 볼 수 있는 {품사} 중 하나이다	예문 사용	89
화합물	화합물001	지구상에서 수소는 대부분 다른 원소와의 {화합물} 형태로 존재한다	예문 사용	122
가본	가본001	겨울방학 때 못 {가본} 곳을 가는 것도 좋다	사용불가	5
보터	보터001	PC {보터의} 경우 각각의 소형 단말기에 두	사용불가	2
보털	보털001	공략하고자 산업별 정보를 모은 일명 {보털과} BBSBullei Board Sysem 개인 누리집	사용불가	2

〈표 20〉 '낱말 말뭉치'를 이용한 예문 추가 예

3) 뜻풀이 이용

『우리말샘』과 A국어사전에서 예문을 찾지 못한 경우, ㈜낱말이 보유하고 있는 20억 어절 말뭉치 검색을 통해서도 예문을 찾지 못한 경우는 『우리말샘』의 뜻풀이를 그대로 이용하였다. <표 21>에서 출처가 '뜻풀이 이용'으로 된 것이 『우리말샘』의 뜻풀이를 이용한 경우이다. '연기'라는 단어가 『우리말샘』에 모두 21가지의 의미를 가진 것으로 나온다. 그중 <표 21>의 11개의 의미가 33만 어휘쌍에 포함되어 있다. 일반인이 흔히 아는 '연기'는 '연기(009)', '연기(013)', '연기(014)' 정도일 것이다. 나머지 의미를 가진 '연기'에 대해서 예문이 있더라도 이해 못 할 가능성이 높고, '연기(009)', '연기(013)', '연기(014)' 중 하나로 착각할 가능성이 높다. 그리고 피평가자에게 뜻풀이를 바로 제시한다고 하여도 그 의미를 이해하기 어려운 경우가 많다.

표제어	예문	출처	뜻풀이
연기001	{연기}는 해마다 돌아오는 제삿날.	뜻풀이 이용	해마다 돌아오는 제삿날.
연기002	{연기}는 운수가 사나운 해.	뜻풀이 이용	운수가 사나운 해.
연기005	{연기}는 자세하게 적은 연보(年譜).	뜻풀이 이용	자세하게 적은 연보(年譜).
연기007	{연기}는 정하여지거나 경과한 햇수.	뜻풀이 이용	정하여지거나 경과한 햇수.
연기009	무기한 {연기}.	『우리말샘』	정해진 시기를 뒤로 미룸.
연기011	{연기}는 둘 이상의 것을 나란히 잇대어 적음.	뜻풀이 이용	둘 이상의 것을 나란히 잇대어 적음.
연기012	우리 방송국에서는 설날 특집으로 남녀 바둑 {연기}를 방송할 예정입니다.	A국어사전	바둑에서, 대국하는 쌍방이 복수로 편을 짜서 일정 수의 착수(着手)를 교대로 두어 나가는 바둑. 서로 착수에 대해서 의

			논을 할 수 없는 것이 상담기와 다르다.
연기013	굴뚝에서 {연기가} 나다.	『우리말샘』	무엇이 불에 탈 때에 생겨나는 흐릿한 기체나 기운.
연기014	{연기} 지도.	『우리말샘』	배우가 배역의 인물, 성격, 행동 따위를 표현해 내는 일.
연기015	괴로움이란 {연기}가 연쇄작용을 일으키면서 생기는 것이다	A국어사전	모든 현상이 생기(生起) 소멸 하는 법칙. 이에 따르면 모든 현상은 원인인 인(因)과 조건인 연(緣)이 상호 관계하여 성립하며, 인연이 없으면 결과도 없다.
연기016	{연기}는 중생의 지혜로 이해할 수 있는 정도로 설법하는 일.	뜻풀이 이용	중생의 지혜로 이해할 수 있는 정도로 설법하는 일.

〈표 21〉『우리말샘』의 뜻풀이를 이용한 예문 추가

『우리말샘』의 표제어는 의미 단위로 구분되어 있는데 어떤 경우는 예문만 가지고는 어느 의미인지 판단하기 어려운 경우도 있다. 예를 들어 '동시010'은 "주로 어린이를 독자로 예상하고 어린이의 정서를 읊은 시"이고 '동시011'은 "어린이가 지은 시"라고 되어 있다. 예문 "그는 동시를 읽고 있다"에서 '동시'는 두 가지 의미를 모두 내포하고 있다. 이런 문제로 인해 설문에 응답하는 사람이 『우리말샘』의 정확한 의미가 아닌 다른 의미의 단어로 생각하고 답을 할 가능성이 있다. 이것은 설문 결과 분석에 중요한 요소가 될 것이다.

<표 22>은 지금까지 과정을 거쳐서 추가된 예문의 수를 출처별로 보여주는 것이다. 333만 어휘쌍에 등장하는 단어 중 예문이 없는 경우가 예문이 있는 경우보다 훨씬 많은데, 예문이 없는 경우는 현재 그러한 의미로는 잘 사용되지 않는다는 것을 간접적으로 보여주는 것이라 할 수 있다. 이와 같이 일반인이 잘 알지 못하는 어려운 단어가 상당 부분 평가 자료로 포함되어 있기 때문에 20만 어휘쌍 중 50% 이상이 유효한 어휘쌍, 즉 10만 어휘쌍을 목표한 것이다.

예문출처	단어 수
『우리말샘』	67,630
A국어사전	7,032
낱말 말뭉치	6,178
뜻풀이 이용	38,010
직접 추가	396
총합계	219,246

〈표 22〉 예문이 추가된 경우 출처별 단어 수

라. 설문 조사를 위한 어휘쌍 세트 분리

1) 4개 관련어 그룹별

설문 조사에 사용할 설문 문항을 만들기 위한 필요 예문이 부족하기 때문에 전체 어휘쌍을 대상으로 세트 구분 작업을 할 수가 없다. 그래서 매주 예문을 추가하면서 대상 후보를 선정하였다. 1차 후보군은 『우리말샘』에 예문이 있어서 바로 설문 문항을 만들 수 있는 어휘쌍이고, 2차, 3차 후보군은 설문 조사 일정에 지장이 없는 선에서 예문을 채워넣어야 하는 어휘쌍의 수를 정한 것이다. 전체 329,696개 어휘쌍 중에 앞서 기술한 제외대상 어휘쌍을 빼고 남은 263,531개 어휘쌍이 후보가 된다. 이 후보 어휘쌍을 이용하여 평가용으로 사용할 수 있는 20만 어휘쌍을 생성하게 된다.

구분	1차후보	2차후보	3차후보	후보전체	제외	총합계
반대말	3,125	3,212	4,376	10,713	2,677	13,390
비슷한말	16,669	33,065	68,571	118,305	209	118,514
상위어	34,285	29,867	8,110	72,262	42,318	114,580
하위어	28,620	27,502	6,129	62,251	20,961	83,212
총합계	82,699	93,646	87,186	263,531	66,165	329,696

<표 23> 관련어 그룹별 후보 어휘쌍

2) 세트별로 어휘 난이도가 고르게 어휘쌍 배분

설문 응답자가 특정 평가 세트의 선택으로 실제 평가에서 피로도를 느끼지 않게 세트 간 어휘 난이도가 차이 나지 않도록 배분한다. 이를 위해 표제어와 관련어에 대한 어휘 난이도를 이용하여 어휘쌍의 난이도 오름차순으로 일련번호를 매긴다. 1천 개 어휘쌍이 한 개의 세트가 되도록 가능한 관련어 그룹별 세트 수를 정한 다음에 세트별로 한 개씩 차례로 배분한다.

3) 세트 내 어휘쌍 배열 순서를 어휘 난이도 순으로 배열

설문 응답자가 1천개의 어휘쌍에 대해 설문을 진행하는 과정에서 집중력을 높이기 위해 난이도가 낮은 것부터 높은 순서대로 어휘쌍을 배열한다. 처음에 너무 어려운 어휘가 등장하면 심리적으로 위축되어 설문을 포기할 가능성이 높기 때문이다.

4) 세트별 이름 부여 및 배열 순서 표시

세트별로 구분하고 세트 내 배열 순서를 난이도순으로 정했다면 별도의 행에 표시한다. 이 작업은 4차에 걸쳐서 진행하였다. <표 24>에 보이듯이 1차로 81개 세트를, 2차로 45개 세트를, 3차로 40개 세

트, 4차로 40개 세트를 만들어 총 206개 세트를 완성하였다. 비슷한말/반대말은 목표 70개 세트를 위해 72개 세트를 완성하였고, 상위어는 목표 70개 세트를 위해 72개 세트를, 하위어는 목표 60개 세트를 위해 62개 세트를 완성하였다.

구분	1차	2차	3차	4차	합계
반대말	3	4	3	0	10
비슷한말	16	20	11	15	62
상위어	34	9	15	14	72
하위어	28	12	11	11	62
총합계	81	45	40	40	206

〈표 24〉 관련어 그룹별 세트 수

3. 설문 항목 설계

가. 효율적인 설문 진행 상황 관리 및 결과 정보 추출을 위한 설계 방안

본 과제는 4만 명 이상으로 구성된 거대 인원의 패널에 대해 총 4천만 개(중복 포함)의 설문 문항을 제공하는 대량 설문 작업이 필수적이며, 제반 설문 진행과 결과 분석을 단기간에 완료하는 것을 목표로 하고 있다. 이를 위해 진행 기간 동안의 시행착오를 최소화할 수 있는 방안이 마련되어야 하며, 전체 설문이 완료되기 이전에 부분적인 결과 분석을 통해 진행상 발생할 수 있는 오류를 적시에 정정해야 할 것으로 예상되었다. 그리고 컨소시엄 업체들 간에 사용해 온 분석 도구와 데이터 처리 프로그램이 상이한 경우 발생할 수 있는 오류 발생의 가능성을 선제적으로 최소화해야 하였다. 따라서 데이터 생성과 전달의 체계를 결정하여 컨소시엄 내에서의 혼선을 줄였으며, 시험 공정(파일럿 테스트)과 본 공정의 설문 진행 중에도 분석과 개입이 용이하게 진행될 수 있도록 설계하였다.

1) 설문 세트의 고유 번호(ID) 부여

본 과제에서 진행될 설문은 총 20만 개의 어휘쌍을 200개의 세트로 분할함으로써 각 세트당 1,000개의 설문 문항을 수록하도록 기획되었다. 각 세트에 대해 최소 200명의 패널 응답이 확보되어야 유효한 결과를 얻을 수 있을 것으로 판단하였다. 각 설문 세트에 포함된 어휘쌍은 다른 세트에 포함되지 않도록 배분될 것이므로, 어휘쌍의 고유번호(pair_ID)를 기준으로 출제 문항을 선발하게 된다. 하지만 결과 분석은 물론 유효성 검증 등 진행상의 문제점이 발견되는 상황에는 각 세트를 기준으로 판단하는 것이 타당할 것으로 판단하였다. 따라서 각 세트별로 고유 번호(ID)를 부여하였으며, 그 규칙은 <표 25>와 같다.

세트_ID: 101 ~ 199	비슷한말
세트_ID: 201 ~ 299	반대말
세트_ID: 301 ~ 399	상위어
세트_ID: 401 ~ 499	하위어

〈표 25〉 세트 고유 번호(ID) 부여 규칙

세트명은 '설문 부문 + 부문별 세트 번호'로 구성되어 있고, 설문 부문은 비슷한말(1), 반대말(2), 상위어(3), 하위어(4)로 구분하였다. 설정한 세트명 부여 규칙에 따르면 비슷한말, 반대말, 상위어, 하위어 등 각 부문의 설문 세트를 99개(99,000개 문항)까지 수록할 수 있으므로 부문별 최대 7만 개 문항을 다루어야 하는 본 과제의 취지에 부합한다고 판단하였다. 또한, 시험 공정(파일럿 테스트)를 위한 설문 세트는 '000'으로 설정되었으며, 본 공정에서는 혼선을 방지하기 위해 세트명에 100, 200, 300, 400번은 사용하지 않았다.

이러한 세트 ID 부여 규칙은 숫자만으로 설문의 부문과 부문별 세트 번호를 알 수 있으므로 설문지 배포 상황을 용이하게 관리할 수 있고, 컨소시엄 업체들이 분석에 사용하는 컴퓨터 프로그램이나 분석 도구의 제한 조건(한글 파일명 제한, 파일명 길이 제한 등)에 따른 오류 발생을 줄일 수 있다는 장점을 갖는다.

본 과제의 설문 진행을 위해 비슷한말 총 60세트(101~160, 총 6만 어휘쌍), 반대말 총 10세트(201~210, 총 1만 어휘쌍), 상위어 총 70세트(301~370, 총 7만 어휘쌍), 하위어 총 60세트(401~460, 총 6만 어휘쌍)를 상기 세트명 부여 규칙에 입각하여 제작하였으며, 부문 총합 200세트(총 20만 어휘쌍)의 설문 세트가 마련되어 설문 공정으로 전달되었다.

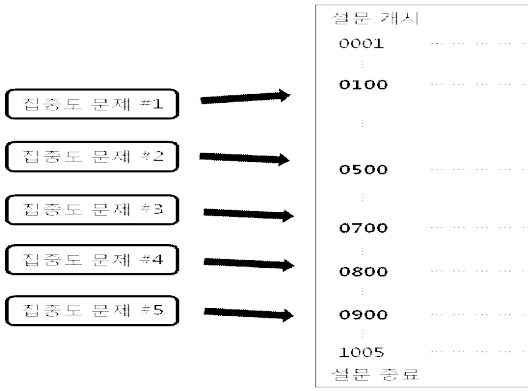
나. 설문 결과 신뢰성 확보 및 원활한 대량 설문 진행을 위한 설계 방안

1) 설문 응답 품질 감시를 위한 설계 방안

본 과제의 목표 달성을 위해서는 단기간에 대량 설문을 진행해야 하는데, 설문 결과에 대한 신뢰성 확보 방안은 중요하게 고려되어야 할 사안이라 할 수 있다. 각 설문 패널은 1,000개의 설문 문항을 담당하게 되어 응답을 진행하게 되는데, 본 과제에서 다루어야 할 어휘쌍의 60% 이상은 일상생활에서 접하기 힘든 난이도를 갖고 있어 설문 진행 중 휴식 시간을 갖더라도 설문 응답의 피로가 축적될 것으로 판단하였다.

또한, 임의로 선택된 질문을 지나치게 빠른 속도로 완료한 경우에는 기존 기법을 통해 구분하기 힘들 수가 있으므로, 단기간의 거대 인원을 모집하면서도 설문 패널의 성실성을 확보해야 하는 본 컨소시엄의 입장에서 부수적인 여과 기법을 마련해야 한다.

본 과제에서는 설문 응답의 패턴 분석을 통해 성실성 여부를 판단하여 불성실 응답을 제외하는 방법에 더하여, 설문 문항에 패널의 집중 여부를 파악할 수 있는 문항(이하 집중도 문항)을 삽입하여 패널의 집중 여부를 검증하는 방법을 병행하였다. 집중도 문제는 각 세트 당 5개씩 사용하되, 감시 지점의 효율적인 배치를 위해 설문 중반 이후의 과정에 집중적인 감시를 하는 것으로 결정하였고, 다음과 같이 배치하였다.



〈그림 1〉 설문 패널의 집중도 체크 지점

따라서 설문 문항 1,000개에 집중도 문항 5개를 더하여 각 설문 세트는 1,005개의 어휘쌍으로 구성되도록 하였다. 설문 패널의 집중도를 확인하기 위한 문항으로서, 시험 공정(파일럿 테스트)에 투입된 문항 중 어휘 등급이 가장 낮은 '하-하' 등급으로 구성된 어휘쌍을 선별하여 컨소시엄 및 자문 위원 회의를 통해 각 부문별로 5개씩 선발하였고, 선발된 집중도 확인용 문항은 다음과 같다.

구분	표제어	예문	관련어 대응 표제어	예문
비슷한말	투잡	요즘은 투잡을 뛰고 있다.	겹벌이	넉넉지 못한 생활비 때문에 겹벌이를 원하고 있었다.
	손빨래하다	속옷을 손빨래하다.	손세탁하다	담요를 손세탁하다.
	상대편	그는 항상 상대편의 눈을 보면서 말을 한다.	상대방	상대방의 입장에서 생각하다.
	스파이전	스파이전을 펼치다.	첩보전	두 나라는 치열한 첩보전을 벌였다.
	작별주	그들은 작별주나 한잔 하자고 거리로 나왔다.	이별주	이별주를 권하며 작별을 아쉬워 했다.
반대말	개교	우리 학교 개교 이래로 너 같은 수재는 처음이다.	폐교	이 마을에 있던 소규모 학교는 폐교가 결정되었다.

	단점	단점을 들추다.	장점	장점을 살리다.
	불합격	시험에서 불합격의 고배를 마시다.	합격	사법고시 합격을 축하합니다.
	앞말	그 문제는 앞말에서 이미 이야기된 바다.	뒷말	뒷말을 잇다.
	좌석	좌석에 앉다.	입석	좌석이 매진되고 남아있는 자리는 입석밖에 없었다.
상위어	해충	해충을 박멸하다.	벌레	벌레 한 마리
	꽃밭	철쭉꽃이 흐드러지게 피어 꽃밭의 물결을 이루었다.	밭	밭 한 뙈기
	과수밭	과수밭 관리가 제대로 되어야 과일의 알이 굵다.	밭	밭 한 뙈기
	금팔찌	굵은 홍보석이 달린 금팔찌 한 쌍을 마리아에게 주었다.	팔찌	팔찌를 끼다.
	늦여름	늦여름에 장맛비가 내렸다.	여름	무더운 여름
하위어	회사	제조 회사	경쟁사	자사 제품이 경쟁사보다 우수하다고 선전하였다.
	산업	새로운 산업에 종사하다.	관광산업	관광산업 육성
	영화	영화 관람	명화	명화를 방영하다.
	기류	환절기 기류를 타고 있는 하늘은 물기를 머금은 듯했다.	난기류	난기류에 휩쓸리다.
	계절	가을은 독서의 계절	가을	높디높은 가을 하늘

〈표 26〉 선발된 집중도 확인용 20개 문항(부문별 5개)

컨소시엄과 자문 위원의 내부 논의를 통해 선발된 각 부문별 5개의 어휘쌍은 객관적으로 일반 수준의 지식을 가진 패널들이 별다른 이견 없이 응답할 수 있을 것으로 판단되므로 무작위 응답 패턴으로 기존의 패널 응답 성실성 검토에서 감지할 수 없는 경우에 대한 대안으로 판단되었다.

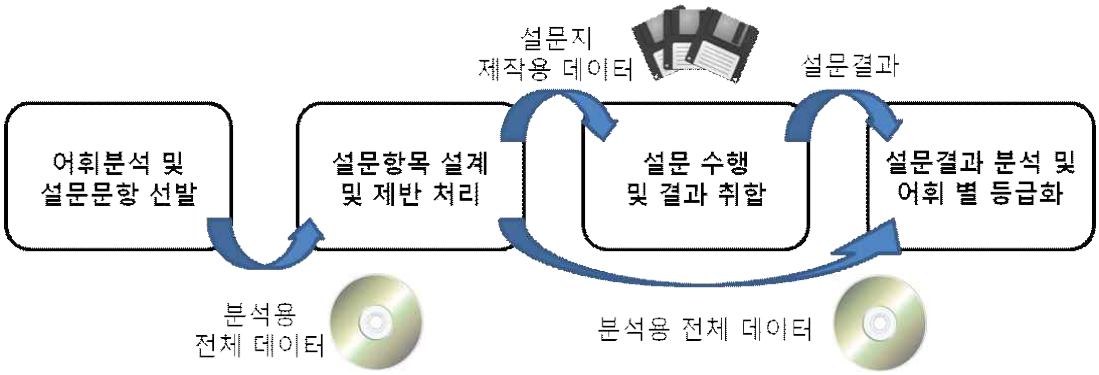
2) 설문 제작 단순화를 위한 설계 방안

본 과제의 설문 공정은 어휘 분석 및 어휘쌍 선별 기관(본 컨소시엄의 ㈜낱말)에서 전달되는 4만 건

의 데이터에 대해 설문 문항 설계 기관(본 컨소시엄의 이지메타(주))에서 설문 문항 데이터 확정 및 각종 처리(세트 분할 및 ID 부여, 집중도 문항 삽입 등)를 수행하고 그 결과를 설문 수행 기관(본 컨소시엄의 ㈜IRC)으로 전달하는 과정으로 구성되었다.

따라서 설문 수행 기관으로 전달되는 데이터는 매주 평균 40세트(4만 어휘쌍)의 설문지를 웹 형태로 제작하여 8천 명 이상의 패널에게 배포하는 작업에 가장 적합하도록 별도의 작업을 최소화할 수 있는 형태로 만들어져야 한다. 설문 결과 취합 후의 분석을 위해서는 설문에 사용되지는 않았으나 분석에 반드시 필요한 데이터가 유지되도록 관리되어야 한다.

이를 위해 본 과제에서는 어휘 등급, 품사, 의미 번호 등 모든 정보를 포함하여 분석 기관에서 사용할 데이터와 설문 담당 기관에서 사용할 데이터를 이원화하여 관리하는 방식으로 진행하였다.



〈그림 2〉 설문 항목 설계 공정에서의 데이터 이원화 과정

설문지 제작용 데이터의 가공 공정은 다음과 같이 진행되었다.

행번호	pair_id	표제어	의미번호	품사	어휘등급	예문	뜻풀이	대응어	…
1	…	…	…	…	…	…	…	…	…
⋮			(설문문항 데이터)						
40,000	…	…	…	…	…	…	…	…	…

정제 및 처리

행번호	pair_id	word1	의미번호	품사	어휘등급	제시문	예문	뜻풀이	word2	…
1	…	…	…	…	…	…	…	…	…	…
⋮				(분석용 데이터)						
40,000	…	…	…	…	…	…	…	…	…	…

〈표 27〉 설문 문항 데이터의 처리를 통한 분석용 데이터 가공

먼저 어휘 등급 분석을 통해 선발된 설문 문항용 어휘쌍 데이터를 전달 받아 정제 및 처리 공정을 진행하였다. 정제 공정은 다음과 같다.

① 어휘에 포함된 불필요한 기호 제거

국립국어원에서 전달 받은 어휘쌍의 '표제어' 및 '관련어 대응 표제어'에 포함된 '-'와 '^'는 설문과 분석에 모두 필요하지 않다고 판단하여 붙여 쓰는 것으로 통일하였다. 제거 작업의 결과는 'word1'과 'word2'로 표현하였으며, 원본인 '표제어'와 '관련어 대응 표제어'가 포함된 동일한 'pair_ID'에 저장되어 추후 원본 조회가 필요할 경우에도 차질이 없도록 하였다.

봄-꽃 → 봄꽃
가마^짓기 → 가마짓기

〈그림 3〉 어휘 기호 제거의 예

② 예문에 포함된 불필요한 기호 제거

『우리말샘』에 수록된 어휘 예문에 포함된 '<FL>'과 '</FL>', 그리고 '<I>'와 '</I>' 태그 기호를 제거하였다. 『우리말샘』에 수록된 어휘 중 일부 단어의 예문에 '?' 기호가 표시된 경우 제공 데이터에는 <FL>, </FL> 태그 기호가 표시되었는데, 설문과 분석에 필요하지 않을 것으로 판단하여 제거하였다.

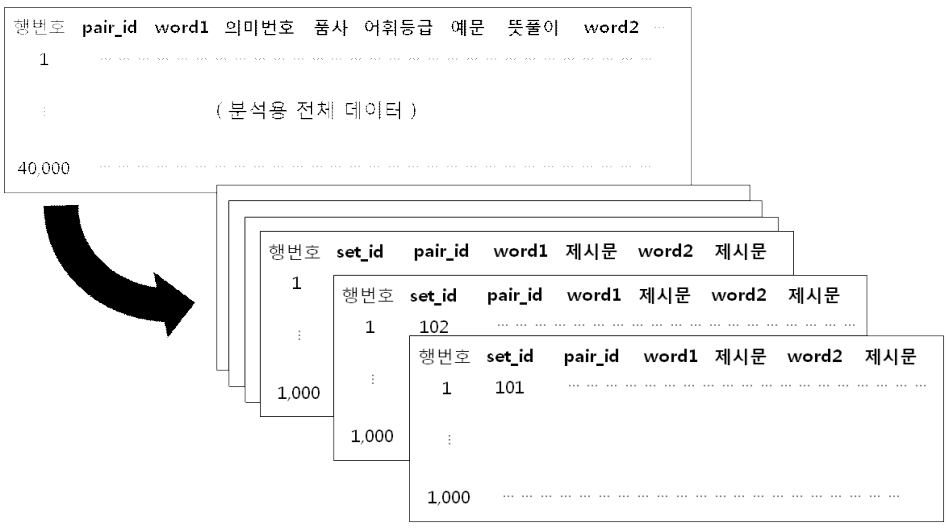
〈그림 4〉 〈FL〉, 〈/FL〉 태그 기호 제거의 예

또한 예문에 수식 또는 특수 외래어가 사용된 경우 제공 데이터에 <I>, </I> 태그가 표시되었는데 이 역시 설문과 분석에는 불필요한 것으로 판단하여 제거하였다.

〈그림 5〉 〈I〉, 〈/I〉 태그 기호 제거의 예

정제 및 처리 공정의 마지막 단계로서 설문 시 어휘에 대한 설명으로 제시될 예문 작업이 진행되었다. 분석용 데이터는 원본 데이터와 마찬가지로 예문과 뜻풀이 정보를 유지해야 하지만, 설문용 데이터에는 예문과 뜻풀이 중 하나만을 제공하여 설문지 제작 공정에 혼선이 없도록 하였다. 이를 위해 본 과제의 컨소시엄에서는 예문이 있는 경우 우선적으로 제공하고 예문 데이터가 없는 경우 뜻풀이를 제공하는 것으로 방침을 정하였고, 이에 설문용 데이터에는 각 단어에 대한 제시문으로 제공하기로 하였다. 제공 데이터에 예문과 뜻풀이가 모두 존재하지 않는 경우가 4건('혼분수', '헤서', '표준형', '수미산') 발견되었고, 『우리말샘』의 뜻풀이에 근거하여 작성된 예문을 적용하였다.

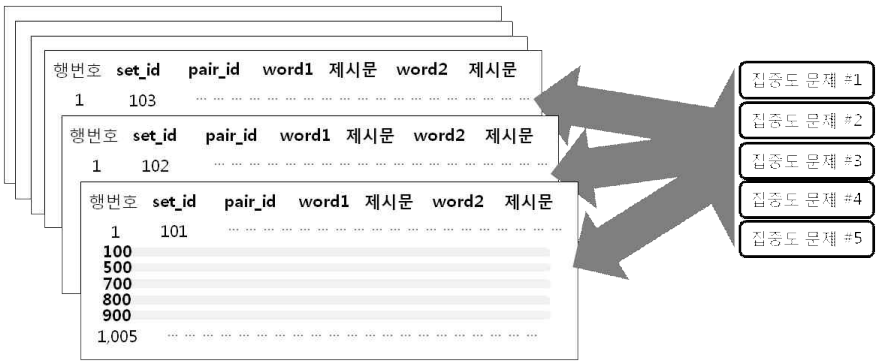
상기의 정제 및 처리 결과를 모두 포함하는 분석용 전체 데이터를 1~5주 차의 설문 진행 기간 동안 주 차별로 가공하여 저장하였고, 전체 데이터로부터 설문 기관에 전달될 설문용 데이터를 가공하는 작업을 진행하였다.



〈그림 6〉 분석용 전체 데이터 분할 및 설문 세트 ID 부여

먼저, 약 4만 어휘쌍의 주 차별 전체 데이터 중에서 설문 공정에서 필요한 세트 ID(set_ID), 어휘쌍 번호(pair_ID), 표제어(word1), 표제어 제시문, 관련어 대응 표제어(word2), 관련어 대응 표제어 제시문 등 6개 정보만을 남긴 상태에서, 각각 1,000개 어휘쌍을 갖는 개별 세트를 분할한 후 세트 ID를 부여하였다. 각 세트를 저장한 개별적인 파일명에는 'set_ID'가 포함되어 있어 설문지 배포 현황과 설문 결과 회수 현황 파악이 용이하도록 하였다.

생성된 각 세트의 설문지에는 집중도 문항을 적용하였다. 상기 설명된 집중도 문항은 각 설문 세트의 100, 500, 700, 800, 900번 문항 위치에 삽입되어 세트별 문항의 수는 1,005개가 됨을 확인한 후 설문 담당 기관(본 컨소시엄의 ㈜IRC)에 전달되었다.



〈그림 7〉 집중도 문제 적용 과정의 예

4. 설문 조사

가. 조사 설계

○ 설문 조사는 전국 만19세 이상 성인 남녀를 대상으로 4가지 유형(비슷한말, 반대말, 상위어, 하위어)의 어휘 관계 평가용 어휘쌍을 제시하고, 제시된 두 단어의 관계를 평가하는 방식으로 설계하였다.

○ 본 조사 진행에 앞서 어휘 제시 방법, 질의 및 평가 방법에 대해 검토하고자 993개 어휘쌍을 대상으로 시험 공정(파일럿 테스트)를 진행하였다. 시험 공정(파일럿 테스트)는 기 구축된 온라인 설문 응답 패널을 대상으로 2019년 12월 17~18일 2일간 진행되었으며, 총 200명의 유효 응답자를 확보하였다.

조사 참여 요청	참여자	완료자	미완료자	제외 (부실 응답 등)	유효 응답자
760	259	224	35	24	200

〈표 28〉 예비조사 진행 결과

		빈도(명)	비율(%)
[전 체]		200	(100.0)
[성별]	남자	92	(46.0)
	여자	108	(54.0)
[연령]	29세 이하	46	(23.0)
	30대	55	(27.5)
	40대	57	(28.5)
	50세 이상	42	(21.0)
[학력 (재학/중퇴포함)]	고등학교 이하	25	(12.5)
	2~3년제 대학	29	(14.5)
	4년제 대학	128	(64.0)
	대학원	18	(9.0)

〈표 29〉 예비조사 유효응답자 특성

○ 예비조사 결과에 나타난 응답 소요 시간 및 이에 따른 응답자 피로도 등을 고려하여, 어휘 관계 평가용 어휘쌍 200,000개를 1,000개 어휘쌍으로 나누어 200개 세트를 구성하고, 각 세트별로 200명의 유효 응답자를 확보하는 방식으로 조사 방식을 최종 결정하였다. 최종 결정된 조사 방식에 따라 비슷한말 60개 세트(60,000개 어휘쌍), 반대말 10개 세트(10,000개 어휘쌍), 상위어 70개 세트(70,000개 어휘쌍), 하위어 60개 세트(60,000개 어휘쌍) 등 총 200개의 응답 세트를 구성하였다.

구분	어휘쌍 수	응답 세트 수	set_ID
비슷한말	60,000	60	101~160
반대말	10,000	10	201~210
상위어	70,000	70	301~370
하위어	60,000	60	401~460
합계	200,000	200	-

〈표 30〉 유형별 어휘 관계 평가용 어휘쌍 및 응답 세트 수

나. 웹 설문 구축

○ 1개 세트당 1,000개 문항으로 구성된 200개 응답 세트를 웹 설문으로 구축하였으며, 유형별 웹 설문의 질의 형태는 다음 그림과 같이 구성하였으며, 모바일 환경과 개인용 컴퓨터 환경에서 모두 조사 진행이 가능하도록 설계하였다.

〈그림 8〉 웹 설문 질의 형태

다. 표본 설계

○ 표본 추출은 기 구축된 온라인 응답 패널을 대상으로 성별·연령별 유의 할당 및 편의 추출 방식으로 진행하였다. 각 응답 세트별 표본 배분은 다음 표와 같이 성별(남성, 여성), 연령별(40세 미만, 40세 이상)로 총 4개 층으로 구분하고 각층별로 유효 표본의 25%씩을 확보하는 것을 목표로 배분하였다. 응답 완료에 장시간 소요되는 본 조사의 특성과 난이도를 고려하여 성별·연령별 목표 표본수는 실사 과정에서 탄력적으로 적용하였고, 1개 세트에 응답을 완료한 응답자는 다른 세트에도 응답할 수 있도록 허용하였다.

구분	40세 미만	40세 이상	계
남성	50(25.0%)	50(25.0%)	100(50.0%)
여성	50(25.0%)	50(25.0%)	100(50.0%)
계	100(50.0%)	100(50.0%)	200(50.0%)

〈표 31〉 응답 세트당 성별·연령별 목표 표본수

라. 실사 진행

○ 본 조사 실사 진행은 2020년 1월 2일부터 2월 6일까지 36일간 진행하여, 40,730명의 유효 응답자를 확보하였다. 각 세트별 조사 참여자 및 유효 응답자 현황은 다음 표와 같다.

set_ID	조사 참여 요청	참여자	완료자	미완료자	제외 (부실 응답 등)	유효 응답자
101	749	293	203	90	0	203
102	754	287	201	86	0	201
103	751	323	202	121	2	200
104	748	308	203	105	1	202
105	696	270	204	66	4	200
106	729	280	207	73	7	200
107	735	269	200	69	0	200
108	719	273	204	69	4	200
109	727	276	204	72	4	200
110	725	279	203	76	2	201
111	1,001	279	207	72	5	202
112	1,001	291	208	83	4	204

set_ID	조사 참여 요청	참여자	완료자	미완료자	제외 (부실 응답 등)	유효 응답자
113	1,164	298	208	90	6	202
114	1,000	296	214	82	7	207
115	1,001	300	212	88	5	207
116	1,086	289	209	80	5	204
117	1,144	261	208	53	6	202
118	1,124	256	208	48	5	203
119	1,176	256	208	48	5	203
120	1,171	250	210	40	8	202
121	1,166	253	208	45	6	202
122	1,238	250	208	42	6	202
123	1,223	241	209	32	6	203
124	1,116	240	208	32	6	202
125	1,155	242	209	33	6	203
126	1,184	244	210	34	7	203
127	1,189	240	208	32	3	205
128	1,067	259	208	51	6	202
129	1,169	280	209	71	6	203
130	1,119	262	208	54	5	203
131	1,068	273	208	65	5	203
132	1,145	270	209	61	7	202
133	1,126	263	209	54	7	202
134	1,130	265	211	54	9	202
135	1,111	264	208	56	6	202
136	1,110	267	209	58	7	202
137	1,704	265	208	57	6	202
138	1,197	255	209	46	7	202
139	1,300	255	208	47	6	202
140	1,380	258	210	48	7	203
141	1,316	259	208	51	6	202
142	1,361	258	208	50	6	202
143	1,427	265	210	55	8	202

set_ID	조사 참여 요청	참여자	완료자	미완료자	제외 (부실 응답 등)	유효 응답자
144	1,361	261	211	50	9	202
145	1,734	258	209	49	7	202
146	1,681	256	209	47	7	202
147	1,774	253	209	44	7	202
148	1,330	237	208	29	1	207
149	1,483	249	210	39	2	208
150	1,500	265	209	56	3	206
151	1,271	246	209	37	0	209
152	1,370	254	209	45	2	207
153	1,240	243	209	34	2	207
154	1,270	256	208	48	1	207
155	1,245	270	208	62	0	208
156	1,186	261	209	52	0	209
157	1,353	267	210	57	1	209
158	1,242	272	208	64	0	208
159	1,310	288	210	78	0	210
160	1,320	265	208	57	3	205
201	1,093	266	208	58	2	206
202	1,131	264	208	56	6	202
203	1,177	240	208	32	5	203
204	1,203	235	208	27	5	203
205	1,199	251	208	43	6	202
206	1,160	241	208	33	5	203
207	1,230	249	209	40	7	202
208	1,728	255	212	43	10	202
209	1,794	256	211	45	7	204
210	1,774	269	211	58	8	203
301	741	284	205	79	5	200
302	730	309	203	106	3	200
303	745	281	203	78	3	200
304	748	280	205	75	5	200

set_ID	조사 참여 요청	참여자	완료자	미완료자	제외 (부실 응답 등)	유효 응답자
305	739	306	206	100	3	203
306	742	300	205	95	5	200
307	723	288	204	84	4	200
308	727	294	203	91	0	203
309	741	289	203	86	3	200
310	729	318	201	117	0	201
311	733	318	201	117	1	200
312	745	290	200	90	0	200
313	726	302	202	100	0	202
314	735	301	205	96	2	203
315	726	289	203	86	1	202
316	718	321	202	119	1	201
317	754	312	200	112	0	200
318	706	274	202	72	0	202
319	761	326	206	120	6	200
320	717	293	207	86	7	200
321	1,001	277	209	68	5	204
322	1,001	289	209	80	4	205
323	1,088	291	208	83	5	203
324	1,001	287	212	75	7	205
325	1,128	286	208	78	5	203
326	1,001	291	213	78	6	207
327	1,170	294	208	86	4	204
328	1,075	299	207	92	5	202
329	1,001	285	214	71	5	209
330	957	286	210	76	6	204
331	1,150	279	208	71	4	204
332	1,260	275	209	66	5	204
333	1,207	279	210	69	4	206
334	1,126	280	209	71	5	204
335	1,171	240	208	32	6	202

set_ID	조사 참여 요청	참여자	완료자	미완료자	제외 (부실 응답 등)	유효 응답자
336	1,220	272	208	64	6	202
337	1,198	266	208	58	6	202
338	1,078	249	208	41	6	202
339	1,072	250	209	41	7	202
340	1,147	254	209	45	7	202
341	1,146	256	209	47	7	202
342	1,258	271	208	63	6	202
343	1,180	265	208	57	6	202
344	1,601	259	210	49	8	202
345	1,693	268	211	57	6	205
346	1,688	267	211	56	9	202
347	1,688	273	212	61	4	208
348	1,774	266	212	54	7	205
349	1,833	281	212	69	8	204
350	1,848	262	213	49	7	206
351	1,794	280	213	67	8	205
352	1,748	266	212	54	8	204
353	1,781	277	218	59	11	207
354	1,841	260	209	51	7	202
355	1,800	261	213	48	3	210
356	1,869	263	211	52	4	207
357	1,280	243	212	31	5	207
358	1,259	247	209	38	7	202
359	1,334	276	203	73	1	202
360	1,156	274	208	66	1	207
361	1,297	273	209	64	1	208
362	1,337	294	208	86	0	208
363	1,314	276	208	68	1	207
364	1,217	280	208	72	2	206
365	1,261	264	207	57	0	207
366	1,382	284	208	76	1	207

set_ID	조사 참여 요청	참여자	완료자	미완료자	제외 (부실 응답 등)	유효 응답자
367	1,369	277	208	69	0	208
368	1,120	285	206	79	1	205
369	1,424	271	209	62	2	207
370	1,298	282	208	74	1	207
401	783	316	203	113	3	200
402	736	289	202	87	2	200
403	730	301	202	99	1	201
404	750	312	205	107	4	201
405	735	300	205	95	5	200
406	735	293	203	90	3	200
407	724	275	204	71	4	200
408	717	306	203	103	3	200
409	715	307	205	102	4	201
410	696	293	202	91	2	200
411	1,033	291	208	83	4	204
412	1,032	288	208	80	5	203
413	1,001	290	207	83	0	207
414	1,001	276	213	63	4	209
415	1,001	284	208	76	1	207
416	1,063	265	208	57	4	204
417	1,268	274	209	65	5	204
418	1,199	270	209	61	4	205
419	1,216	263	208	55	5	203
420	1,281	261	209	52	5	204
421	1,214	251	209	42	6	203
422	1,333	242	208	34	4	204
423	1,407	254	211	43	8	203
424	1,305	254	208	46	6	202
425	1,202	252	208	44	4	204
426	1,210	238	208	30	4	204
427	1,236	254	208	46	5	203

set_ID	조사 참여 요청	참여자	완료자	미완료자	제외 (부실 응답 등)	유효 응답자
428	1,317	246	209	37	6	203
429	1,146	263	208	55	6	202
430	1,180	266	209	57	5	204
431	1,174	257	208	49	6	202
432	1,169	261	207	54	5	202
433	1,190	266	208	58	6	202
434	1,140	264	208	56	6	202
435	1,230	266	209	57	7	202
436	1,177	260	208	52	6	202
437	1,223	267	208	59	6	202
438	1,183	262	208	54	6	202
439	1,100	247	207	40	5	202
440	1,116	265	209	56	3	206
441	1,307	258	208	50	6	202
442	1,221	247	209	38	7	202
443	1,400	240	208	32	4	204
444	1,449	246	208	38	1	207
445	1,350	255	209	46	2	207
446	994	261	205	56	2	203
447	1,209	247	208	39	2	206
448	1,333	252	208	44	0	208
449	1,323	258	210	48	5	205
450	1,329	243	208	35	1	207
451	1,325	250	210	40	0	210
452	1,392	261	209	52	1	208
453	1,325	270	209	61	0	209
454	1,262	265	208	57	1	207
455	1,296	263	209	54	4	205
456	1,290	262	208	54	1	207
457	1,357	261	209	52	1	208
458	1,494	290	209	81	1	208

set_ID	조사 참여 요청	참여자	완료자	미완료자	제외 (부실 응답 등)	유효 응답자
459	1,419	266	209	57	0	209
460	1,322	267	208	59	0	208
계	233,754	54,145	41,566	12,579	836	40,730

〈표 32〉 응답 세트별 조사 참여자 및 유효 응답자

○ 각 응답 세트별 응답자 특성은 다음과 같다.

세트 번호	전체	[성별]		[연령]				[학력(재학/중퇴포함)]			
		남자	여자	29세 이하	30대	40대	50세 이상	고등학교 이하	2~3년제 대학	4년제 대학	대학원
101	203 (100.0)	113 (55.7)	90 (44.3)	23 (11.3)	43 (21.2)	70 (34.5)	67 (33.0)	36 (17.7)	18 (8.9)	127 (62.6)	22 (10.8)
102	201 (100.0)	116 (57.7)	85 (42.3)	23 (11.4)	46 (22.9)	88 (43.8)	44 (21.9)	26 (12.9)	24 (11.9)	126 (62.7)	25 (12.4)
103	200 (100.0)	109 (54.5)	91 (45.5)	24 (12.0)	62 (31.0)	62 (31.0)	52 (26.0)	32 (16.0)	27 (13.5)	124 (62.0)	17 (8.5)
104	202 (100.0)	109 (54.0)	93 (46.0)	26 (12.9)	49 (24.3)	66 (32.7)	61 (30.2)	32 (15.8)	29 (14.4)	120 (59.4)	21 (10.4)
105	200 (100.0)	97 (48.5)	103 (51.5)	28 (14.0)	44 (22.0)	77 (38.5)	51 (25.5)	25 (12.5)	35 (17.5)	112 (56.0)	28 (14.0)
106	200 (100.0)	102 (51.0)	98 (49.0)	29 (14.5)	53 (26.5)	68 (34.0)	50 (25.0)	28 (14.0)	26 (13.0)	119 (59.5)	27 (13.5)
107	200 (100.0)	101 (50.5)	99 (49.5)	22 (11.0)	57 (28.5)	73 (36.5)	48 (24.0)	28 (14.0)	39 (19.5)	107 (53.5)	26 (13.0)
108	200 (100.0)	107 (53.5)	93 (46.5)	26 (13.0)	43 (21.5)	81 (40.5)	50 (25.0)	29 (14.5)	35 (17.5)	118 (59.0)	18 (9.0)
109	200 (100.0)	119 (59.5)	81 (40.5)	29 (14.5)	47 (23.5)	72 (36.0)	52 (26.0)	35 (17.5)	35 (17.5)	109 (54.5)	21 (10.5)
110	201 (100.0)	120 (59.7)	81 (40.3)	24 (11.9)	49 (24.4)	74 (36.8)	54 (26.9)	29 (14.4)	27 (13.4)	119 (59.2)	26 (12.9)
111	202 (100.0)	112 (55.4)	90 (44.6)	36 (17.8)	74 (36.6)	51 (25.2)	41 (20.3)	35 (17.3)	24 (11.9)	129 (63.9)	14 (6.9)
112	204 (100.0)	115 (56.4)	89 (43.6)	46 (22.5)	76 (37.3)	48 (23.5)	34 (16.7)	26 (12.7)	34 (16.7)	120 (58.8)	24 (11.8)
113	202 (100.0)	106 (52.5)	96 (47.5)	40 (19.8)	68 (33.7)	54 (26.7)	40 (19.8)	30 (14.9)	24 (11.9)	124 (61.4)	24 (11.9)
114	207 (100.0)	110 (53.1)	97 (46.9)	38 (18.4)	78 (37.7)	50 (24.2)	41 (19.8)	23 (11.1)	36 (17.4)	117 (56.5)	31 (15.0)
115	207 (100.0)	104 (50.2)	103 (49.8)	55 (26.6)	61 (29.5)	54 (26.1)	37 (17.9)	35 (16.9)	36 (17.4)	117 (56.5)	19 (9.2)
116	204 (100.0)	108 (52.9)	96 (47.1)	31 (15.2)	80 (39.2)	55 (27.0)	38 (18.6)	35 (17.2)	37 (18.1)	114 (55.9)	18 (8.8)

세트 번호	전체	[성별]		[연령]				[학력(재학/중퇴포함)]			
		남자	여자	29세 이하	30대	40대	50세 이상	고등학교 이하	2~3년제 대학	4년제 대학	대학원
117	202 (100.0)	98 (48.5)	104 (51.5)	52 (25.7)	62 (30.7)	50 (24.8)	38 (18.8)	40 (19.8)	25 (12.4)	122 (60.4)	15 (7.4)
118	203 (100.0)	104 (51.2)	99 (48.8)	54 (26.6)	74 (36.5)	40 (19.7)	35 (17.2)	29 (14.3)	20 (9.9)	133 (65.5)	21 (10.3)
119	203 (100.0)	98 (48.3)	105 (51.7)	41 (20.2)	70 (34.5)	52 (25.6)	40 (19.7)	30 (14.8)	38 (18.7)	125 (61.6)	10 (4.9)
120	202 (100.0)	108 (53.5)	94 (46.5)	44 (21.8)	70 (34.7)	53 (26.2)	35 (17.3)	32 (15.8)	24 (11.9)	126 (62.4)	20 (9.9)
121	202 (100.0)	84 (41.6)	118 (58.4)	42 (20.8)	74 (36.6)	47 (23.3)	39 (19.3)	31 (15.3)	27 (13.4)	121 (59.9)	23 (11.4)
122	202 (100.0)	102 (50.5)	100 (49.5)	39 (19.3)	76 (37.6)	47 (23.3)	40 (19.8)	38 (18.8)	26 (12.9)	113 (55.9)	25 (12.4)
123	203 (100.0)	113 (55.7)	90 (44.3)	38 (18.7)	77 (37.9)	57 (28.1)	31 (15.3)	27 (13.3)	40 (19.7)	113 (55.7)	23 (11.3)
124	202 (100.0)	108 (53.5)	94 (46.5)	51 (25.2)	65 (32.2)	52 (25.7)	34 (16.8)	34 (16.8)	31 (15.3)	117 (57.9)	20 (9.9)
125	203 (100.0)	112 (55.2)	91 (44.8)	39 (19.2)	74 (36.5)	52 (25.6)	38 (18.7)	34 (16.7)	35 (17.2)	118 (58.1)	16 (7.9)
126	203 (100.0)	105 (51.7)	98 (48.3)	55 (27.1)	57 (28.1)	50 (24.6)	41 (20.2)	32 (15.8)	24 (11.8)	129 (63.5)	18 (8.9)
127	205 (100.0)	103 (50.2)	102 (49.8)	31 (15.1)	91 (44.4)	55 (26.8)	28 (13.7)	32 (15.6)	39 (19.0)	112 (54.6)	22 (10.7)
128	202 (100.0)	93 (46.0)	109 (54.0)	44 (21.8)	73 (36.1)	48 (23.8)	37 (18.3)	38 (18.8)	29 (14.4)	121 (59.9)	14 (6.9)
129	203 (100.0)	101 (49.8)	102 (50.2)	39 (19.2)	88 (43.3)	40 (19.7)	36 (17.7)	34 (16.7)	35 (17.2)	115 (56.7)	19 (9.4)
130	203 (100.0)	102 (50.2)	101 (49.8)	48 (23.6)	66 (32.5)	50 (24.6)	39 (19.2)	31 (15.3)	34 (16.7)	112 (55.2)	26 (12.8)
131	203 (100.0)	107 (52.7)	96 (47.3)	44 (21.7)	73 (36.0)	45 (22.2)	41 (20.2)	35 (17.2)	37 (18.2)	114 (56.2)	17 (8.4)
132	202 (100.0)	98 (48.5)	104 (51.5)	50 (24.8)	58 (28.7)	61 (30.2)	33 (16.3)	32 (15.8)	27 (13.4)	123 (60.9)	20 (9.9)
133	202 (100.0)	91 (45.0)	111 (55.0)	43 (21.3)	75 (37.1)	43 (21.3)	41 (20.3)	32 (15.8)	29 (14.4)	115 (56.9)	26 (12.9)
134	202 (100.0)	98 (48.5)	104 (51.5)	39 (19.3)	75 (37.1)	54 (26.7)	34 (16.8)	33 (16.3)	31 (15.3)	122 (60.4)	16 (7.9)
135	202 (100.0)	109 (54.0)	93 (46.0)	45 (22.3)	82 (40.6)	38 (18.8)	37 (18.3)	35 (17.3)	22 (10.9)	127 (62.9)	18 (8.9)
136	202 (100.0)	106 (52.5)	96 (47.5)	46 (22.8)	69 (34.2)	52 (25.7)	35 (17.3)	36 (17.8)	27 (13.4)	125 (61.9)	14 (6.9)
137	202 (100.0)	110 (54.5)	92 (45.5)	42 (20.8)	69 (34.2)	44 (21.8)	47 (23.3)	36 (17.8)	28 (13.9)	123 (60.9)	15 (7.4)
138	202 (100.0)	125 (61.9)	77 (38.1)	45 (22.3)	75 (37.1)	48 (23.8)	34 (16.8)	29 (14.4)	33 (16.3)	116 (57.4)	24 (11.9)

세트 번호	전체	[성별]		[연령]				[학력(재학/중퇴포함)]			
		남자	여자	29세 이하	30대	40대	50세 이상	고등학교 이하	2~3년제 대학	4년제 대학	대학원
139	202 (100.0)	88 (43.6)	114 (56.4)	37 (18.3)	75 (37.1)	54 (26.7)	36 (17.8)	38 (18.8)	24 (11.9)	120 (59.4)	20 (9.9)
140	203 (100.0)	106 (52.2)	97 (47.8)	43 (21.2)	68 (33.5)	54 (26.6)	38 (18.7)	31 (15.3)	34 (16.7)	111 (54.7)	27 (13.3)
141	202 (100.0)	98 (48.5)	104 (51.5)	53 (26.2)	61 (30.2)	52 (25.7)	36 (17.8)	34 (16.8)	34 (16.8)	118 (58.4)	16 (7.9)
142	202 (100.0)	106 (52.5)	96 (47.5)	30 (14.9)	72 (35.6)	59 (29.2)	41 (20.3)	34 (16.8)	36 (17.8)	116 (57.4)	16 (7.9)
143	202 (100.0)	107 (53.0)	95 (47.0)	45 (22.3)	70 (34.7)	49 (24.3)	38 (18.8)	23 (11.4)	31 (15.3)	130 (64.4)	18 (8.9)
144	202 (100.0)	117 (57.9)	85 (42.1)	41 (20.3)	77 (38.1)	45 (22.3)	39 (19.3)	32 (15.8)	33 (16.3)	118 (58.4)	19 (9.4)
145	202 (100.0)	109 (54.0)	93 (46.0)	40 (19.8)	70 (34.7)	61 (30.2)	31 (15.3)	35 (17.3)	23 (11.4)	135 (66.8)	9 (4.5)
146	202 (100.0)	112 (55.4)	90 (44.6)	39 (19.3)	75 (37.1)	45 (22.3)	43 (21.3)	34 (16.8)	34 (16.8)	118 (58.4)	16 (7.9)
147	202 (100.0)	114 (56.4)	88 (43.6)	34 (16.8)	70 (34.7)	61 (30.2)	37 (18.3)	36 (17.8)	31 (15.3)	118 (58.4)	17 (8.4)
148	207 (100.0)	110 (53.1)	97 (46.9)	40 (19.3)	82 (39.6)	52 (25.1)	33 (15.9)	23 (11.1)	41 (19.8)	118 (57.0)	25 (12.1)
149	208 (100.0)	112 (53.8)	96 (46.2)	40 (19.2)	73 (35.1)	55 (26.4)	40 (19.2)	35 (16.8)	36 (17.3)	114 (54.8)	23 (11.1)
150	206 (100.0)	95 (46.1)	111 (53.9)	38 (18.4)	59 (28.6)	47 (22.8)	62 (30.1)	38 (18.4)	23 (11.2)	122 (59.2)	23 (11.2)
151	209 (100.0)	109 (52.2)	100 (47.8)	42 (20.1)	76 (36.4)	55 (26.3)	36 (17.2)	31 (14.8)	29 (13.9)	126 (60.3)	23 (11.0)
152	207 (100.0)	107 (51.7)	100 (48.3)	41 (19.8)	72 (34.8)	55 (26.6)	39 (18.8)	30 (14.5)	35 (16.9)	127 (61.4)	15 (7.2)
153	207 (100.0)	100 (48.3)	107 (51.7)	56 (27.1)	74 (35.7)	37 (17.9)	40 (19.3)	32 (15.5)	20 (9.7)	136 (65.7)	19 (9.2)
154	207 (100.0)	108 (52.2)	99 (47.8)	45 (21.7)	64 (30.9)	55 (26.6)	43 (20.8)	35 (16.9)	29 (14.0)	130 (62.8)	13 (6.3)
155	208 (100.0)	108 (51.9)	100 (48.1)	43 (20.7)	75 (36.1)	52 (25.0)	38 (18.3)	22 (10.6)	23 (11.1)	140 (67.3)	23 (11.1)
156	209 (100.0)	113 (54.1)	96 (45.9)	43 (20.6)	75 (35.9)	54 (25.8)	37 (17.7)	29 (13.9)	31 (14.8)	117 (56.0)	32 (15.3)
157	209 (100.0)	111 (53.1)	98 (46.9)	44 (21.1)	75 (35.9)	55 (26.3)	35 (16.7)	35 (16.7)	28 (13.4)	134 (64.1)	12 (5.7)
158	208 (100.0)	101 (48.6)	107 (51.4)	41 (19.7)	84 (40.4)	46 (22.1)	37 (17.8)	35 (16.8)	23 (11.1)	132 (63.5)	18 (8.7)
159	210 (100.0)	103 (49.0)	107 (51.0)	44 (21.0)	72 (34.3)	57 (27.1)	37 (17.6)	33 (15.7)	29 (13.8)	128 (61.0)	20 (9.5)
160	205 (100.0)	96 (46.8)	109 (53.2)	36 (17.6)	66 (32.2)	51 (24.9)	52 (25.4)	36 (17.6)	29 (14.1)	118 (57.6)	22 (10.7)

세트 번호	전체	[성별]		[연령]				[학력(재학/중퇴포함)]			
		남자	여자	29세 이하	30대	40대	50세 이상	고등학교 이하	2~3년제 대학	4년제 대학	대학원
201	206 (100.0)	104 (50.5)	102 (49.5)	43 (20.9)	72 (35.0)	57 (27.7)	34 (16.5)	30 (14.6)	32 (15.5)	112 (54.4)	32 (15.5)
202	202 (100.0)	109 (54.0)	93 (46.0)	36 (17.8)	72 (35.6)	50 (24.8)	44 (21.8)	27 (13.4)	25 (12.4)	129 (63.9)	21 (10.4)
203	203 (100.0)	113 (55.7)	90 (44.3)	40 (19.7)	78 (38.4)	50 (24.6)	35 (17.2)	25 (12.3)	29 (14.3)	128 (63.1)	21 (10.3)
204	203 (100.0)	118 (58.1)	85 (41.9)	39 (19.2)	84 (41.4)	46 (22.7)	34 (16.7)	29 (14.3)	32 (15.8)	123 (60.6)	19 (9.4)
205	202 (100.0)	111 (55.0)	91 (45.0)	40 (19.8)	71 (35.1)	57 (28.2)	34 (16.8)	35 (17.3)	22 (10.9)	132 (65.3)	13 (6.4)
206	203 (100.0)	111 (54.7)	92 (45.3)	40 (19.7)	69 (34.0)	50 (24.6)	44 (21.7)	32 (15.8)	27 (13.3)	128 (63.1)	16 (7.9)
207	202 (100.0)	109 (54.0)	93 (46.0)	35 (17.3)	71 (35.1)	57 (28.2)	39 (19.3)	39 (19.3)	22 (10.9)	124 (61.4)	17 (8.4)
208	202 (100.0)	109 (54.0)	93 (46.0)	32 (15.8)	74 (36.6)	53 (26.2)	43 (21.3)	31 (15.3)	30 (14.9)	116 (57.4)	25 (12.4)
209	204 (100.0)	113 (55.4)	91 (44.6)	45 (22.1)	74 (36.3)	41 (20.1)	44 (21.6)	27 (13.2)	32 (15.7)	130 (63.7)	15 (7.4)
210	203 (100.0)	105 (51.7)	98 (48.3)	30 (14.8)	76 (37.4)	55 (27.1)	42 (20.7)	32 (15.8)	36 (17.7)	116 (57.1)	19 (9.4)
301	200 (100.0)	101 (50.5)	99 (49.5)	29 (14.5)	51 (25.5)	71 (35.5)	49 (24.5)	27 (13.5)	24 (12.0)	119 (59.5)	30 (15.0)
302	200 (100.0)	111 (55.5)	89 (44.5)	27 (13.5)	50 (25.0)	75 (37.5)	48 (24.0)	27 (13.5)	34 (17.0)	114 (57.0)	25 (12.5)
303	200 (100.0)	112 (56.0)	88 (44.0)	26 (13.0)	50 (25.0)	82 (41.0)	42 (21.0)	40 (20.0)	25 (12.5)	122 (61.0)	13 (6.5)
304	200 (100.0)	110 (55.0)	90 (45.0)	28 (14.0)	55 (27.5)	73 (36.5)	44 (22.0)	38 (19.0)	27 (13.5)	116 (58.0)	19 (9.5)
305	203 (100.0)	108 (53.2)	95 (46.8)	27 (13.3)	54 (26.6)	70 (34.5)	52 (25.6)	32 (15.8)	30 (14.8)	112 (55.2)	29 (14.3)
306	200 (100.0)	108 (54.0)	92 (46.0)	20 (10.0)	57 (28.5)	58 (29.0)	65 (32.5)	28 (14.0)	28 (14.0)	118 (59.0)	26 (13.0)
307	200 (100.0)	102 (51.0)	98 (49.0)	16 (8.0)	55 (27.5)	69 (34.5)	60 (30.0)	27 (13.5)	42 (21.0)	106 (53.0)	25 (12.5)
308	203 (100.0)	114 (56.2)	89 (43.8)	28 (13.8)	53 (26.1)	70 (34.5)	52 (25.6)	36 (17.7)	26 (12.8)	114 (56.2)	27 (13.3)
309	200 (100.0)	112 (56.0)	88 (44.0)	24 (12.0)	43 (21.5)	79 (39.5)	54 (27.0)	34 (17.0)	25 (12.5)	109 (54.5)	32 (16.0)
310	201 (100.0)	97 (48.3)	104 (51.7)	31 (15.4)	47 (23.4)	62 (30.8)	61 (30.3)	31 (15.4)	31 (15.4)	120 (59.7)	19 (9.5)
311	200 (100.0)	111 (55.5)	89 (44.5)	24 (12.0)	55 (27.5)	56 (28.0)	65 (32.5)	27 (13.5)	34 (17.0)	120 (60.0)	19 (9.5)
312	200 (100.0)	106 (53.0)	94 (47.0)	28 (14.0)	56 (28.0)	70 (35.0)	46 (23.0)	29 (14.5)	24 (12.0)	122 (61.0)	25 (12.5)

세트 번호	전체	[성별]		[연령]				[학력(재학/중퇴포함)]			
		남자	여자	29세 이하	30대	40대	50세 이상	고등학교 이하	2~3년제 대학	4년제 대학	대학원
313	202 (100.0)	118 (58.4)	84 (41.6)	36 (17.8)	48 (23.8)	63 (31.2)	55 (27.2)	37 (18.3)	28 (13.9)	114 (56.4)	23 (11.4)
314	203 (100.0)	111 (54.7)	92 (45.3)	27 (13.3)	48 (23.6)	81 (39.9)	47 (23.2)	34 (16.7)	29 (14.3)	123 (60.6)	17 (8.4)
315	202 (100.0)	108 (53.5)	94 (46.5)	29 (14.4)	46 (22.8)	74 (36.6)	53 (26.2)	25 (12.4)	28 (13.9)	124 (61.4)	25 (12.4)
316	201 (100.0)	103 (51.2)	98 (48.8)	28 (13.9)	48 (23.9)	70 (34.8)	55 (27.4)	24 (11.9)	31 (15.4)	123 (61.2)	23 (11.4)
317	200 (100.0)	108 (54.0)	92 (46.0)	27 (13.5)	55 (27.5)	64 (32.0)	54 (27.0)	50 (25.0)	26 (13.0)	103 (51.5)	21 (10.5)
318	202 (100.0)	124 (61.4)	78 (38.6)	21 (10.4)	44 (21.8)	76 (37.6)	61 (30.2)	37 (18.3)	20 (9.9)	123 (60.9)	22 (10.9)
319	200 (100.0)	100 (50.0)	100 (50.0)	23 (11.5)	45 (22.5)	83 (41.5)	49 (24.5)	22 (11.0)	27 (13.5)	124 (62.0)	27 (13.5)
320	200 (100.0)	109 (54.5)	91 (45.5)	29 (14.5)	46 (23.0)	66 (33.0)	59 (29.5)	27 (13.5)	25 (12.5)	115 (57.5)	33 (16.5)
321	204 (100.0)	115 (56.4)	89 (43.6)	40 (19.6)	70 (34.3)	49 (24.0)	45 (22.1)	27 (13.2)	29 (14.2)	122 (59.8)	26 (12.7)
322	205 (100.0)	122 (59.5)	83 (40.5)	40 (19.5)	84 (41.0)	44 (21.5)	37 (18.0)	28 (13.7)	33 (16.1)	121 (59.0)	23 (11.2)
323	203 (100.0)	116 (57.1)	87 (42.9)	44 (21.7)	70 (34.5)	62 (30.5)	27 (13.3)	37 (18.2)	23 (11.3)	125 (61.6)	18 (8.9)
324	205 (100.0)	120 (58.5)	85 (41.5)	41 (20.0)	75 (36.6)	49 (23.9)	40 (19.5)	35 (17.1)	33 (16.1)	123 (60.0)	14 (6.8)
325	203 (100.0)	114 (56.2)	89 (43.8)	29 (14.3)	80 (39.4)	60 (29.6)	34 (16.7)	40 (19.7)	26 (12.8)	123 (60.6)	14 (6.9)
326	207 (100.0)	111 (53.6)	96 (46.4)	46 (22.2)	69 (33.3)	51 (24.6)	41 (19.8)	31 (15.0)	33 (15.9)	119 (57.5)	24 (11.6)
327	204 (100.0)	109 (53.4)	95 (46.6)	44 (21.6)	78 (38.2)	32 (15.7)	50 (24.5)	29 (14.2)	29 (14.2)	132 (64.7)	14 (6.9)
328	202 (100.0)	98 (48.5)	104 (51.5)	31 (15.3)	70 (34.7)	55 (27.2)	46 (22.8)	28 (13.9)	37 (18.3)	122 (60.4)	15 (7.4)
329	209 (100.0)	116 (55.5)	93 (44.5)	47 (22.5)	67 (32.1)	61 (29.2)	34 (16.3)	35 (16.7)	28 (13.4)	118 (56.5)	28 (13.4)
330	204 (100.0)	87 (42.6)	117 (57.4)	37 (18.1)	68 (33.3)	58 (28.4)	41 (20.1)	35 (17.2)	33 (16.2)	112 (54.9)	24 (11.8)
331	204 (100.0)	103 (50.5)	101 (49.5)	41 (20.1)	72 (35.3)	52 (25.5)	39 (19.1)	36 (17.6)	31 (15.2)	124 (60.8)	13 (6.4)
332	204 (100.0)	104 (51.0)	100 (49.0)	36 (17.6)	80 (39.2)	48 (23.5)	40 (19.6)	33 (16.2)	36 (17.6)	113 (55.4)	22 (10.8)
333	206 (100.0)	102 (49.5)	104 (50.5)	53 (25.7)	66 (32.0)	52 (25.2)	35 (17.0)	29 (14.1)	34 (16.5)	116 (56.3)	27 (13.1)
334	204 (100.0)	108 (52.9)	96 (47.1)	47 (23.0)	70 (34.3)	47 (23.0)	40 (19.6)	34 (16.7)	37 (18.1)	118 (57.8)	15 (7.4)

세트 번호	전체	[성별]		[연령]				[학력(재학/중퇴포함)]			
		남자	여자	29세 이하	30대	40대	50세 이상	고등학교 이하	2~3년제 대학	4년제 대학	대학원
335	202 (100.0)	107 (53.0)	95 (47.0)	44 (21.8)	73 (36.1)	47 (23.3)	38 (18.8)	30 (14.9)	29 (14.4)	121 (59.9)	22 (10.9)
336	202 (100.0)	114 (56.4)	88 (43.6)	44 (21.8)	80 (39.6)	35 (17.3)	43 (21.3)	25 (12.4)	29 (14.4)	133 (65.8)	15 (7.4)
337	202 (100.0)	102 (50.5)	100 (49.5)	25 (12.4)	69 (34.2)	59 (29.2)	49 (24.3)	28 (13.9)	36 (17.8)	120 (59.4)	18 (8.9)
338	202 (100.0)	110 (54.5)	92 (45.5)	48 (23.8)	74 (36.6)	52 (25.7)	28 (13.9)	29 (14.4)	28 (13.9)	121 (59.9)	24 (11.9)
339	202 (100.0)	94 (46.5)	108 (53.5)	34 (16.8)	65 (32.2)	59 (29.2)	44 (21.8)	30 (14.9)	35 (17.3)	114 (56.4)	23 (11.4)
340	202 (100.0)	106 (52.5)	96 (47.5)	41 (20.3)	70 (34.7)	46 (22.8)	45 (22.3)	29 (14.4)	28 (13.9)	130 (64.4)	15 (7.4)
341	202 (100.0)	114 (56.4)	88 (43.6)	46 (22.8)	74 (36.6)	45 (22.3)	37 (18.3)	26 (12.9)	32 (15.8)	122 (60.4)	22 (10.9)
342	202 (100.0)	108 (53.5)	94 (46.5)	40 (19.8)	71 (35.1)	53 (26.2)	38 (18.8)	32 (15.8)	25 (12.4)	122 (60.4)	23 (11.4)
343	202 (100.0)	106 (52.5)	96 (47.5)	41 (20.3)	66 (32.7)	52 (25.7)	43 (21.3)	23 (11.4)	44 (21.8)	112 (55.4)	23 (11.4)
344	202 (100.0)	120 (59.4)	82 (40.6)	47 (23.3)	74 (36.6)	53 (26.2)	28 (13.9)	30 (14.9)	27 (13.4)	121 (59.9)	24 (11.9)
345	205 (100.0)	93 (45.4)	112 (54.6)	33 (16.1)	65 (31.7)	59 (28.8)	48 (23.4)	30 (14.6)	35 (17.1)	120 (58.5)	20 (9.8)
346	202 (100.0)	101 (50.0)	101 (50.0)	43 (21.3)	74 (36.6)	46 (22.8)	39 (19.3)	40 (19.8)	26 (12.9)	124 (61.4)	12 (5.9)
347	208 (100.0)	119 (57.2)	89 (42.8)	40 (19.2)	84 (40.4)	45 (21.6)	39 (18.8)	35 (16.8)	34 (16.3)	122 (58.7)	17 (8.2)
348	205 (100.0)	108 (52.7)	97 (47.3)	51 (24.9)	64 (31.2)	56 (27.3)	34 (16.6)	31 (15.1)	35 (17.1)	117 (57.1)	22 (10.7)
349	204 (100.0)	111 (54.4)	93 (45.6)	47 (23.0)	79 (38.7)	45 (22.1)	33 (16.2)	33 (16.2)	34 (16.7)	118 (57.8)	19 (9.3)
350	206 (100.0)	103 (50.0)	103 (50.0)	51 (24.8)	57 (27.7)	65 (31.6)	33 (16.0)	28 (13.6)	30 (14.6)	123 (59.7)	25 (12.1)
351	205 (100.0)	102 (49.8)	103 (50.2)	44 (21.5)	71 (34.6)	49 (23.9)	41 (20.0)	37 (18.0)	31 (15.1)	112 (54.6)	25 (12.2)
352	204 (100.0)	103 (50.5)	101 (49.5)	42 (20.6)	73 (35.8)	55 (27.0)	34 (16.7)	31 (15.2)	32 (15.7)	123 (60.3)	18 (8.8)
353	207 (100.0)	108 (52.2)	99 (47.8)	41 (19.8)	82 (39.6)	50 (24.2)	34 (16.4)	40 (19.3)	23 (11.1)	127 (61.4)	17 (8.2)
354	202 (100.0)	103 (51.0)	99 (49.0)	43 (21.3)	67 (33.2)	56 (27.7)	36 (17.8)	40 (19.8)	24 (11.9)	126 (62.4)	12 (5.9)
355	210 (100.0)	111 (52.9)	99 (47.1)	48 (22.9)	69 (32.9)	52 (24.8)	41 (19.5)	29 (13.8)	30 (14.3)	119 (56.7)	32 (15.2)
356	207 (100.0)	112 (54.1)	95 (45.9)	36 (17.4)	72 (34.8)	59 (28.5)	40 (19.3)	29 (14.0)	27 (13.0)	132 (63.8)	19 (9.2)

세트 번호	전체	[성별]		[연령]				[학력(재학/중퇴포함)]			
		남자	여자	29세 이하	30대	40대	50세 이상	고등학교 이하	2~3년제 대학	4년제 대학	대학원
357	207 (100.0)	104 (50.2)	103 (49.8)	45 (21.7)	69 (33.3)	51 (24.6)	42 (20.3)	41 (19.8)	28 (13.5)	124 (59.9)	14 (6.8)
358	202 (100.0)	112 (55.4)	90 (44.6)	49 (24.3)	79 (39.1)	36 (17.8)	38 (18.8)	29 (14.4)	20 (9.9)	135 (66.8)	18 (8.9)
359	202 (100.0)	103 (51.0)	99 (49.0)	45 (22.3)	57 (28.2)	66 (32.7)	34 (16.8)	32 (15.8)	31 (15.3)	118 (58.4)	21 (10.4)
360	207 (100.0)	107 (51.7)	100 (48.3)	48 (23.2)	74 (35.7)	48 (23.2)	37 (17.9)	35 (16.9)	31 (15.0)	121 (58.5)	20 (9.7)
361	208 (100.0)	110 (52.9)	98 (47.1)	51 (24.5)	64 (30.8)	55 (26.4)	38 (18.3)	31 (14.9)	37 (17.8)	114 (54.8)	26 (12.5)
362	208 (100.0)	78 (37.5)	130 (62.5)	28 (13.5)	61 (29.3)	63 (30.3)	56 (26.9)	31 (14.9)	28 (13.5)	136 (65.4)	13 (6.3)
363	207 (100.0)	97 (46.9)	110 (53.1)	39 (18.8)	73 (35.3)	55 (26.6)	40 (19.3)	36 (17.4)	27 (13.0)	131 (63.3)	13 (6.3)
364	206 (100.0)	91 (44.2)	115 (55.8)	41 (19.9)	68 (33.0)	50 (24.3)	47 (22.8)	35 (17.0)	29 (14.1)	119 (57.8)	23 (11.2)
365	207 (100.0)	116 (56.0)	91 (44.0)	49 (23.7)	74 (35.7)	57 (27.5)	27 (13.0)	32 (15.5)	29 (14.0)	124 (59.9)	22 (10.6)
366	207 (100.0)	101 (48.8)	106 (51.2)	27 (13.0)	78 (37.7)	59 (28.5)	43 (20.8)	25 (12.1)	43 (20.8)	121 (58.5)	18 (8.7)
367	208 (100.0)	111 (53.4)	97 (46.6)	48 (23.1)	78 (37.5)	38 (18.3)	44 (21.2)	28 (13.5)	30 (14.4)	137 (65.9)	13 (6.3)
368	205 (100.0)	121 (59.0)	84 (41.0)	38 (18.5)	73 (35.6)	52 (25.4)	42 (20.5)	28 (13.7)	29 (14.1)	122 (59.5)	26 (12.7)
369	207 (100.0)	112 (54.1)	95 (45.9)	32 (15.5)	78 (37.7)	58 (28.0)	39 (18.8)	35 (16.9)	27 (13.0)	130 (62.8)	15 (7.2)
370	207 (100.0)	118 (57.0)	89 (43.0)	40 (19.3)	78 (37.7)	51 (24.6)	38 (18.4)	34 (16.4)	34 (16.4)	125 (60.4)	14 (6.8)
401	200 (100.0)	123 (61.5)	77 (38.5)	22 (11.0)	46 (23.0)	77 (38.5)	55 (27.5)	34 (17.0)	31 (15.5)	105 (52.5)	30 (15.0)
402	200 (100.0)	113 (56.5)	87 (43.5)	24 (12.0)	47 (23.5)	80 (40.0)	49 (24.5)	34 (17.0)	22 (11.0)	125 (62.5)	19 (9.5)
403	201 (100.0)	107 (53.2)	94 (46.8)	25 (12.4)	49 (24.4)	77 (38.3)	50 (24.9)	33 (16.4)	27 (13.4)	122 (60.7)	19 (9.5)
404	201 (100.0)	116 (57.7)	85 (42.3)	18 (9.0)	55 (27.4)	77 (38.3)	51 (25.4)	30 (14.9)	26 (12.9)	123 (61.2)	22 (10.9)
405	200 (100.0)	115 (57.5)	85 (42.5)	23 (11.5)	54 (27.0)	76 (38.0)	47 (23.5)	37 (18.5)	33 (16.5)	122 (61.0)	8 (4.0)
406	200 (100.0)	114 (57.0)	86 (43.0)	22 (11.0)	50 (25.0)	78 (39.0)	50 (25.0)	31 (15.5)	31 (15.5)	125 (62.5)	13 (6.5)
407	200 (100.0)	109 (54.5)	91 (45.5)	29 (14.5)	45 (22.5)	64 (32.0)	62 (31.0)	39 (19.5)	35 (17.5)	107 (53.5)	19 (9.5)
408	200 (100.0)	112 (56.0)	88 (44.0)	29 (14.5)	41 (20.5)	64 (32.0)	66 (33.0)	36 (18.0)	37 (18.5)	100 (50.0)	27 (13.5)

세트 번호	전체	[성별]		[연령]				[학력(재학/중퇴포함)]			
		남자	여자	29세 이하	30대	40대	50세 이상	고등학교 이하	2~3년제 대학	4년제 대학	대학원
409	201 (100.0)	106 (52.7)	95 (47.3)	29 (14.4)	32 (15.9)	86 (42.8)	54 (26.9)	31 (15.4)	43 (21.4)	97 (48.3)	30 (14.9)
410	200 (100.0)	93 (46.5)	107 (53.5)	22 (11.0)	46 (23.0)	72 (36.0)	60 (30.0)	38 (19.0)	37 (18.5)	90 (45.0)	35 (17.5)
411	204 (100.0)	108 (52.9)	96 (47.1)	49 (24.0)	62 (30.4)	59 (28.9)	34 (16.7)	32 (15.7)	31 (15.2)	120 (58.8)	21 (10.3)
412	203 (100.0)	94 (46.3)	109 (53.7)	39 (19.2)	81 (39.9)	41 (20.2)	42 (20.7)	39 (19.2)	29 (14.3)	110 (54.2)	25 (12.3)
413	207 (100.0)	103 (49.8)	104 (50.2)	39 (18.8)	80 (38.6)	54 (26.1)	34 (16.4)	28 (13.5)	31 (15.0)	129 (62.3)	19 (9.2)
414	209 (100.0)	106 (50.7)	103 (49.3)	45 (21.5)	89 (42.6)	40 (19.1)	35 (16.7)	35 (16.7)	25 (12.0)	129 (61.7)	20 (9.6)
415	207 (100.0)	114 (55.1)	93 (44.9)	42 (20.3)	72 (34.8)	58 (28.0)	35 (16.9)	37 (17.9)	25 (12.1)	128 (61.8)	17 (8.2)
416	204 (100.0)	110 (53.9)	94 (46.1)	43 (21.1)	72 (35.3)	54 (26.5)	35 (17.2)	29 (14.2)	32 (15.7)	113 (55.4)	30 (14.7)
417	204 (100.0)	101 (49.5)	103 (50.5)	35 (17.2)	79 (38.7)	47 (23.0)	43 (21.1)	24 (11.8)	24 (11.8)	136 (66.7)	20 (9.8)
418	205 (100.0)	105 (51.2)	100 (48.8)	48 (23.4)	66 (32.2)	55 (26.8)	36 (17.6)	39 (19.0)	25 (12.2)	125 (61.0)	16 (7.8)
419	203 (100.0)	109 (53.7)	94 (46.3)	57 (28.1)	68 (33.5)	40 (19.7)	38 (18.7)	30 (14.8)	22 (10.8)	129 (63.5)	22 (10.8)
420	204 (100.0)	105 (51.5)	99 (48.5)	37 (18.1)	71 (34.8)	55 (27.0)	41 (20.1)	30 (14.7)	39 (19.1)	120 (58.8)	15 (7.4)
421	203 (100.0)	109 (53.7)	94 (46.3)	44 (21.7)	69 (34.0)	53 (26.1)	37 (18.2)	30 (14.8)	30 (14.8)	120 (59.1)	23 (11.3)
422	204 (100.0)	92 (45.1)	112 (54.9)	40 (19.6)	74 (36.3)	52 (25.5)	38 (18.6)	27 (13.2)	29 (14.2)	124 (60.8)	24 (11.8)
423	203 (100.0)	113 (55.7)	90 (44.3)	38 (18.7)	83 (40.9)	45 (22.2)	37 (18.2)	32 (15.8)	28 (13.8)	116 (57.1)	27 (13.3)
424	202 (100.0)	115 (56.9)	87 (43.1)	43 (21.3)	69 (34.2)	54 (26.7)	36 (17.8)	30 (14.9)	35 (17.3)	115 (56.9)	22 (10.9)
425	204 (100.0)	104 (51.0)	100 (49.0)	53 (26.0)	64 (31.4)	52 (25.5)	35 (17.2)	34 (16.7)	33 (16.2)	118 (57.8)	19 (9.3)
426	204 (100.0)	119 (58.3)	85 (41.7)	45 (22.1)	68 (33.3)	54 (26.5)	37 (18.1)	35 (17.2)	30 (14.7)	118 (57.8)	21 (10.3)
427	203 (100.0)	105 (51.7)	98 (48.3)	54 (26.6)	53 (26.1)	53 (26.1)	43 (21.2)	32 (15.8)	28 (13.8)	124 (61.1)	19 (9.4)
428	203 (100.0)	108 (53.2)	95 (46.8)	32 (15.8)	87 (42.9)	56 (27.6)	28 (13.8)	30 (14.8)	33 (16.3)	120 (59.1)	20 (9.9)
429	202 (100.0)	107 (53.0)	95 (47.0)	54 (26.7)	63 (31.2)	51 (25.2)	34 (16.8)	33 (16.3)	37 (18.3)	112 (55.4)	20 (9.9)
430	204 (100.0)	106 (52.0)	98 (48.0)	28 (13.7)	76 (37.3)	59 (28.9)	41 (20.1)	38 (18.6)	34 (16.7)	115 (56.4)	17 (8.3)

세트 번호	전체	[성별]		[연령]				[학력(재학/중퇴포함)]			
		남자	여자	29세 이하	30대	40대	50세 이상	고등학교 이하	2~3년제 대학	4년제 대학	대학원
431	202 (100.0)	106 (52.5)	96 (47.5)	45 (22.3)	71 (35.1)	46 (22.8)	40 (19.8)	29 (14.4)	32 (15.8)	120 (59.4)	21 (10.4)
432	202 (100.0)	109 (54.0)	93 (46.0)	40 (19.8)	82 (40.6)	44 (21.8)	36 (17.8)	31 (15.3)	31 (15.3)	120 (59.4)	20 (9.9)
433	202 (100.0)	114 (56.4)	88 (43.6)	42 (20.8)	65 (32.2)	61 (30.2)	34 (16.8)	38 (18.8)	24 (11.9)	126 (62.4)	14 (6.9)
434	202 (100.0)	113 (55.9)	89 (44.1)	38 (18.8)	75 (37.1)	48 (23.8)	41 (20.3)	32 (15.8)	32 (15.8)	122 (60.4)	16 (7.9)
435	202 (100.0)	106 (52.5)	96 (47.5)	33 (16.3)	72 (35.6)	58 (28.7)	39 (19.3)	36 (17.8)	24 (11.9)	125 (61.9)	17 (8.4)
436	202 (100.0)	108 (53.5)	94 (46.5)	44 (21.8)	74 (36.6)	48 (23.8)	36 (17.8)	29 (14.4)	27 (13.4)	120 (59.4)	26 (12.9)
437	202 (100.0)	113 (55.9)	89 (44.1)	45 (22.3)	77 (38.1)	33 (16.3)	47 (23.3)	31 (15.3)	30 (14.9)	129 (63.9)	12 (5.9)
438	202 (100.0)	103 (51.0)	99 (49.0)	27 (13.4)	72 (35.6)	55 (27.2)	48 (23.8)	24 (11.9)	43 (21.3)	119 (58.9)	16 (7.9)
439	202 (100.0)	109 (54.0)	93 (46.0)	46 (22.8)	74 (36.6)	52 (25.7)	30 (14.9)	30 (14.9)	30 (14.9)	120 (59.4)	22 (10.9)
440	206 (100.0)	92 (44.7)	114 (55.3)	35 (17.0)	65 (31.6)	60 (29.1)	46 (22.3)	35 (17.0)	29 (14.1)	121 (58.7)	21 (10.2)
441	202 (100.0)	99 (49.0)	103 (51.0)	36 (17.8)	70 (34.7)	54 (26.7)	42 (20.8)	31 (15.3)	33 (16.3)	127 (62.9)	11 (5.4)
442	202 (100.0)	105 (52.0)	97 (48.0)	41 (20.3)	75 (37.1)	47 (23.3)	39 (19.3)	28 (13.9)	23 (11.4)	129 (63.9)	22 (10.9)
443	204 (100.0)	92 (45.1)	112 (54.9)	38 (18.6)	69 (33.8)	48 (23.5)	49 (24.0)	28 (13.7)	22 (10.8)	130 (63.7)	24 (11.8)
444	207 (100.0)	99 (47.8)	108 (52.2)	46 (22.2)	68 (32.9)	48 (23.2)	45 (21.7)	36 (17.4)	30 (14.5)	120 (58.0)	21 (10.1)
445	207 (100.0)	112 (54.1)	95 (45.9)	36 (17.4)	81 (39.1)	57 (27.5)	33 (15.9)	27 (13.0)	38 (18.4)	119 (57.5)	23 (11.1)
446	203 (100.0)	112 (55.2)	91 (44.8)	48 (23.6)	63 (31.0)	51 (25.1)	41 (20.2)	33 (16.3)	31 (15.3)	121 (59.6)	18 (8.9)
447	206 (100.0)	108 (52.4)	98 (47.6)	40 (19.4)	76 (36.9)	58 (28.2)	32 (15.5)	34 (16.5)	34 (16.5)	122 (59.2)	16 (7.8)
448	208 (100.0)	111 (53.4)	97 (46.6)	52 (25.0)	63 (30.3)	53 (25.5)	40 (19.2)	35 (16.8)	27 (13.0)	126 (60.6)	20 (9.6)
449	205 (100.0)	109 (53.2)	96 (46.8)	28 (13.7)	85 (41.5)	59 (28.8)	33 (16.1)	28 (13.7)	36 (17.6)	121 (59.0)	20 (9.8)
450	207 (100.0)	107 (51.7)	100 (48.3)	40 (19.3)	76 (36.7)	47 (22.7)	44 (21.3)	34 (16.4)	28 (13.5)	132 (63.8)	13 (6.3)
451	210 (100.0)	117 (55.7)	93 (44.3)	47 (22.4)	71 (33.8)	55 (26.2)	37 (17.6)	31 (14.8)	31 (14.8)	124 (59.0)	24 (11.4)
452	208 (100.0)	111 (53.4)	97 (46.6)	41 (19.7)	71 (34.1)	57 (27.4)	39 (18.8)	33 (15.9)	20 (9.6)	142 (68.3)	13 (6.3)

세트 번호	전체	[성별]		[연령]				[학력(재학/중퇴포함)]			
		남자	여자	29세 이하	30대	40대	50세 이상	고등학교 이하	2~3년제 대학	4년제 대학	대학원
453	209 (100.0)	115 (55.0)	94 (45.0)	40 (19.1)	76 (36.4)	54 (25.8)	39 (18.7)	32 (15.3)	30 (14.4)	120 (57.4)	27 (12.9)
454	207 (100.0)	111 (53.6)	96 (46.4)	44 (21.3)	81 (39.1)	45 (21.7)	37 (17.9)	23 (11.1)	30 (14.5)	132 (63.8)	22 (10.6)
455	205 (100.0)	112 (54.6)	93 (45.4)	27 (13.2)	75 (36.6)	59 (28.8)	44 (21.5)	32 (15.6)	32 (15.6)	122 (59.5)	19 (9.3)
456	207 (100.0)	103 (49.8)	104 (50.2)	53 (25.6)	66 (31.9)	50 (24.2)	38 (18.4)	37 (17.9)	30 (14.5)	122 (58.9)	18 (8.7)
457	208 (100.0)	102 (49.0)	106 (51.0)	40 (19.2)	74 (35.6)	56 (26.9)	38 (18.3)	24 (11.5)	36 (17.3)	125 (60.1)	23 (11.1)
458	208 (100.0)	118 (56.7)	90 (43.3)	45 (21.6)	70 (33.7)	56 (26.9)	37 (17.8)	42 (20.2)	28 (13.5)	113 (54.3)	25 (12.0)
459	209 (100.0)	115 (55.0)	94 (45.0)	48 (23.0)	71 (34.0)	53 (25.4)	37 (17.7)	30 (14.4)	25 (12.0)	125 (59.8)	29 (13.9)
460	208 (100.0)	105 (50.5)	103 (49.5)	38 (18.3)	77 (37.0)	46 (22.1)	47 (22.6)	37 (17.8)	29 (13.9)	126 (60.6)	16 (7.7)
합계	40,730 (100.0)	21,466 (52.7)	19,264 (47.3)	7,710 (18.9)	13,547 (33.3)	11,138 (27.3)	8,335 (20.5)	6,395 (15.7)	6,010 (14.8)	24,230 (59.5)	4,095 (10.1)

〈표 33〉 세트별 응답자 특성

5. 설문 결과 분석

가. 설문 데이터의 품질 분석

1) 설문 데이터에 대한 품질 확보 방안

설문 데이터에 대한 품질 확보를 위해, 어휘쌍에 대한 사용자 평가 데이터 구축 관련 문헌(Simlex-999(Hill et al. 2015), wordsim-353(Finkelstein et al. 2001), Zesche & Gurevych (2009))을 참고하여 품질 절차와 기준을 수립하고 이에 따라 데이터를 분석하고 품질을 확보하였다.

○ 관리적 품질 확보 절차

다음과 같은 품질 확보 절차를 수립하고 이행하였다.

첫째, 품질을 확보하고 검증하기 위한 기준을 수립하기 위하여 시험 공정(파일럿 테스트)를 수행하였다. 이때 비슷한말/반대말/상위어/하위어 모두를 포함하는 1,000개의 문항에 대해 200명의 응답을 확보한다.

둘째, 20만 쌍 응답의 유효성을 높이기 위하여 응답할 수 있는 최대 수치인 1,000개의 문항으로 구성하였다.

셋째, 200개의 세트(트랜치)로 구성하되, 전체 응답의 일관성과 균질성을 확보하기 위하여 각 세트에 일관성 확인 문제들(consistency questions set)을 추가하였다.

넷째, 일관성 확인 문제를 포함하여 불성실한 응답을 제거하였다(예를 들어, 중간보고 시점 기준 116개 세트 총 624명의 응답을 확인하였고, 확인된 세트의 일부는 200명의 충분조건을 달성하지 못하여 재설문을 실시하였다. 최종적으로 선택된 응답은 <표 34> 참조).

○ 정량적 품질 확보 기준

다음과 같은 품질 확보 기준을 수립하고 검증하였다. 다만, 대상이 되는 어휘쌍의 난이도를 고려할 때, 아래의 품질 기준을 하나도 빠짐없이 모두 만족해야 한다는 것은 아니다. 모든 기준을 전반적으로 고려하여, 이상치가 발생할 수 있는 데이터의 존재 여부를 확인하고, 이를 조정하는 것을 목적으로 하였다.

첫째, 각 세트(트랜치)의 일관성 확인 문제의 응답값 평균이 전체 평균의 응답값 및 평균과 비교하여 표준편차 범위 내에 있어야 한다. 만약, 현저한 차이로 미달성된 경우, 전체 일관성 확인 문제들의 평균과 비교하여 해당 문제에 대한 응답자의 평균 응답값과 비교하여 1 이상 차이가 있는 경우 응답자의 결과를 조정한다(일괄적으로 1점을 상향 조정하되, 5점인 경우는 5점을 그대로 사용).

둘째, 응답자의 일관적인 응답을 측정하기 위하여 IAA(Inter Annotator Agreement) 측정 방안을 수립하고 품질 기준을 수립하였다. IAA는 설문 대상 자료의 난이도를 고려하여 IAA hard 유형과 IAA normal 유형으로 구분하여 측정하되, 각 세트의 일관성 확인 문제에 대한 IAA normal 수치가 0.5 이상이며, 각 세트에 포함된 응답의 IAA hard 수치가 0.3 이상이어야 한다.

미달성시, 응답된 결과의 전체 상관 비교(pair-wise correlation)를 수행하고 그 응답의 평균값이 전체 평균값의 표준편차 범위보다 낮게 나오는 응답자의 응답을 제외한다(예: 다른 사람 평가와의 상관계수가 낮은 경우).

셋째, 응답된 결과는 다른 응답의 결과와 높은 상관계수를 가져야 한다. 본 설문에서는 시험 공정(파일럿 테스트)에서 수집된 1,000쌍의 데이터와 본 공정에서 수집된 1,000쌍의 데이터는 여러 세트와 여러 문제 번호와 여러 응답자로 분산되어 수집되었음)를 비교하여 0.7 이상인 경우를 품질 높은 설문의 기준으로 수립하였다.

〈참고 중간 결과 - 2월 5일 기준〉〉

세트 구분: 수집된 응답수: IAA (hard): IAA (medium): 유효응답

'유의101' '203' '0.406992336202535' '0.775727758220652' '202'
'유의102' '201' '0.340490862629225' '0.806120490641982' '201'
'유의103' '200' '0.412496108403743' '0.734185131297813' '197'
'유의104' '203' '0.385378623284923' '0.84769047355328' '202'
'상위305' '200' '0.386887881527891' '0.798500996793914' '198'
'상위308' '203' '0.415920417750039' '0.700999797674619' '202'
'상위310' '201' '0.351411269554263' '0.755974614853988' '200'
'상위311' '201' '0.363811413468187' '0.703167361984228' '200'
'상위316' '202' '0.355951285817096' '0.622601870242398' '200'
'상위319' '200' '0.400038932195886' '0.693909887021226' '200'

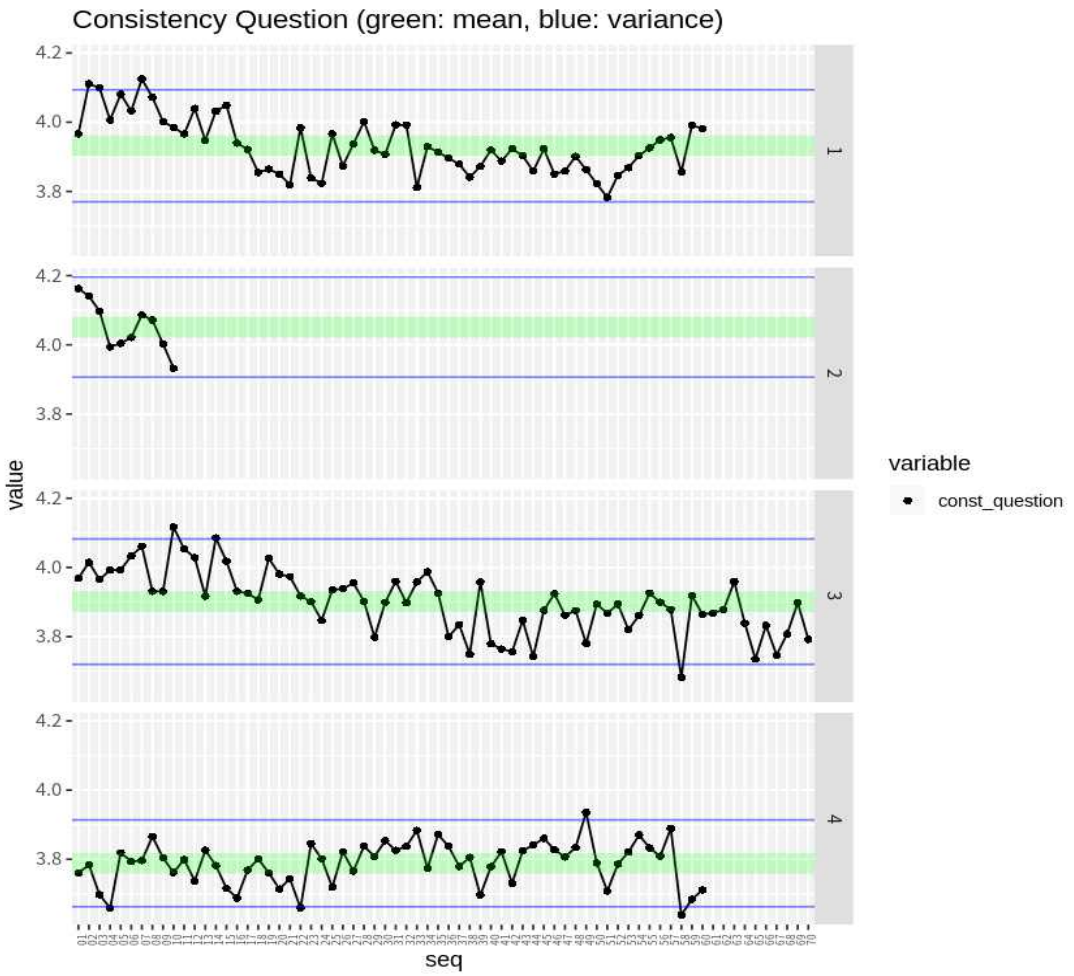
'하위403' '202' '0.329350102425785' '0.561708916435688' '202'
'하위408' '201' '0.344052789356044' '0.603877943809403' '201'

2) 설문 데이터에 대한 품질 검증 및 분석

○ 전체 세트의 일관성 및 품질

본 검증은 여러 세트로 분할되어 응답을 하는 그룹들 간의 유의미한 편차가 있는지를 확인하고, 편차가 발견되면 이를 조정하여 데이터의 균질성을 높이기 위한 것이다.

다음 <그림 9>에서와 같이, 일관성 확인 문제에 대한 평균은 각각 '비슷한말(1), 반대말(2), 상위어(3), 하위어(4)'로 구분하여 평균하였다. 주의할 점은 본 설문에 포함된 일관성 문제는 사전의 '내용 검토(contents validity)'를 통하여, 다른 응답에 비하여 거리가 가까울 것으로 생각되는 문항들을 선택한 것이다. 따라서 일관성 문제의 평균이 전체 평균에 대한 표준편차 범위보다 더 높은 경우(예를 들어 비슷한말(1)의 2, 3, 7번 세트 및 상위어의 10번 세트) 해당 세트의 응답자들은 본 일관성 확인 설문의 취지에 비교적 성실히 응답한 것으로 판단할 수 있다.



〈그림 9〉 세트별 일관성 문제 검토

○ 응답 결과의 일관성 및 품질

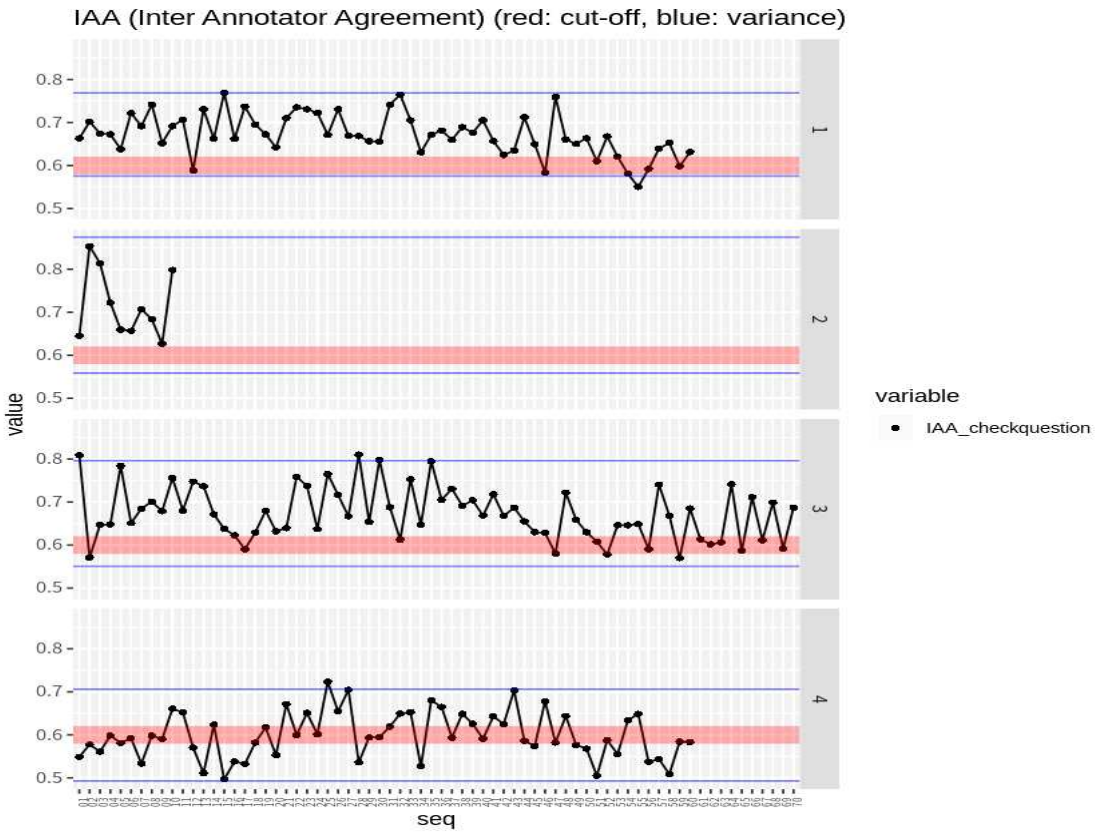
본 검증의 목적은 Simlex-999 등 어휘쌍에 대한 설문 조사에서 설문자간 일관성 검증(Inter Annotator Agreement)을 수행하는 것이다. 세부 방법은 기존의 다양한 지표를 검토하여 전체 평가에 대한 쌍대 비교(pair-wise correlation)를 수행하는 것이다. 관련 연구에 의하면 스피어만 상관계수, 피어슨 상관계수, 피셔 트랜스포메이션(Fisher'Z) 등으로 측정할 수 있다.

본 연구는 Zesche & Gurevych (2009)의 연구를 준용하여, 상대적으로 어휘 간 거리 측정 특성에 합하는 피셔 트랜스포메이션 수치를 적용하였다. 다만 수용 기준점(cut-off)은 응답의 난이도에 따라 다르게 적용될 수 있기 때문에(Zesche & Gurevych, 2009), 본 과제에서는 설문에 따라 IAA 유형과 IAA-hard 유형으로 구분하였다.

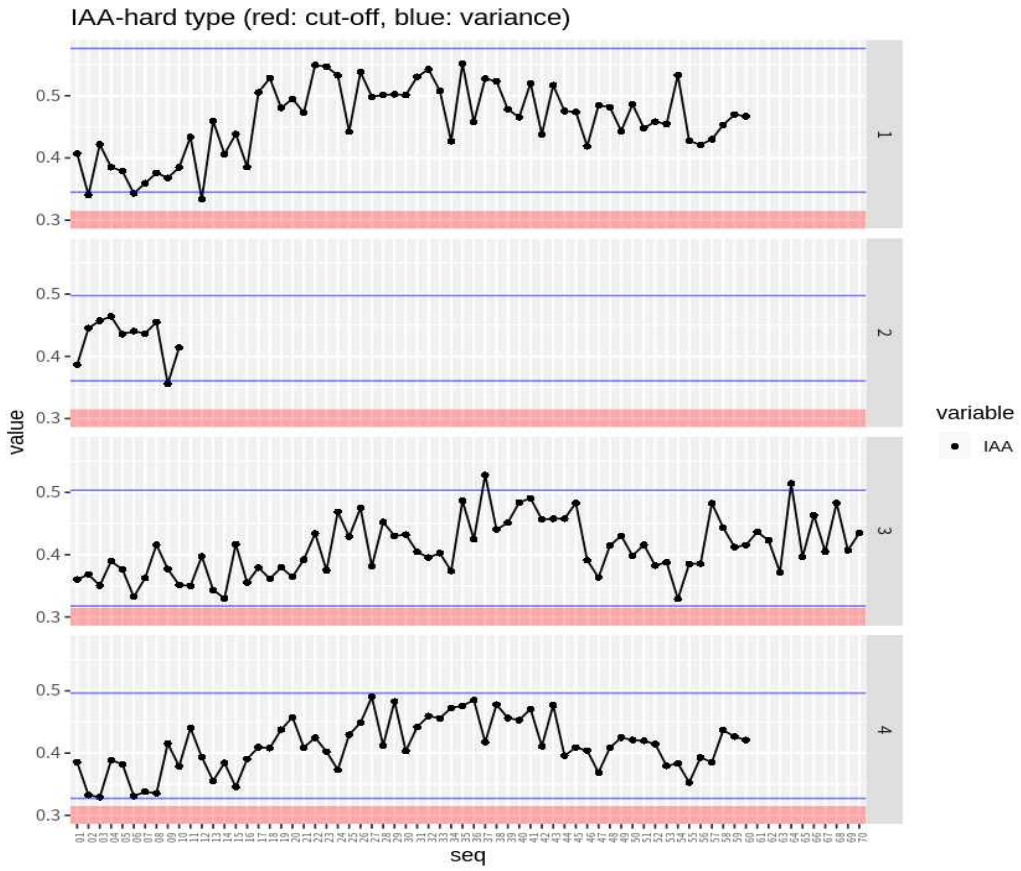
전자는 일관성 문제에 대한 평가자들의 응답 결과의 일관성을 측정한 것이다. 이러한 평가에 있어 표제어와 관련어가 갖는 구체성의 정도가 상이한 경우(예를 들어, 상하위어 관계), 응답자는 혼동을 더 많이 할 수 있기 때문에 응답간의 편차가 일반적인 유의어보다 더 크다(Hill et al., 2015). 따라서 '비슷한 말/반대말'보다는 '상위어/하위어'의 IAA값이 더 낮을 것으로 예측되었으며, 다음 <그림 10>에서 확인할 수 있듯이 상위어/하위어 IAA가 조금 더 낮을 것을 확인할 수 있다.

그러나 이 경우에 하위어를 제외하고는 대부분이 일반적인 수용 기준점(cut-off)인 0.6 이상을 만족하였으며, 하위어에서도 본 설문 조사를 통해 수립한 0.5 이상을 충족하였다.

후자는 전체 문제에 대한 평가자들의 응답 결과의 일관성을 측정한 것이다. 일관성 검증에서는 IAA_hard 유형의 기준 값인 0.3 이상을 모두 충족하였다.



<그림 10> 세트별 IAA 검토(1)



〈그림 11〉 세트별 IAA 검토(2)

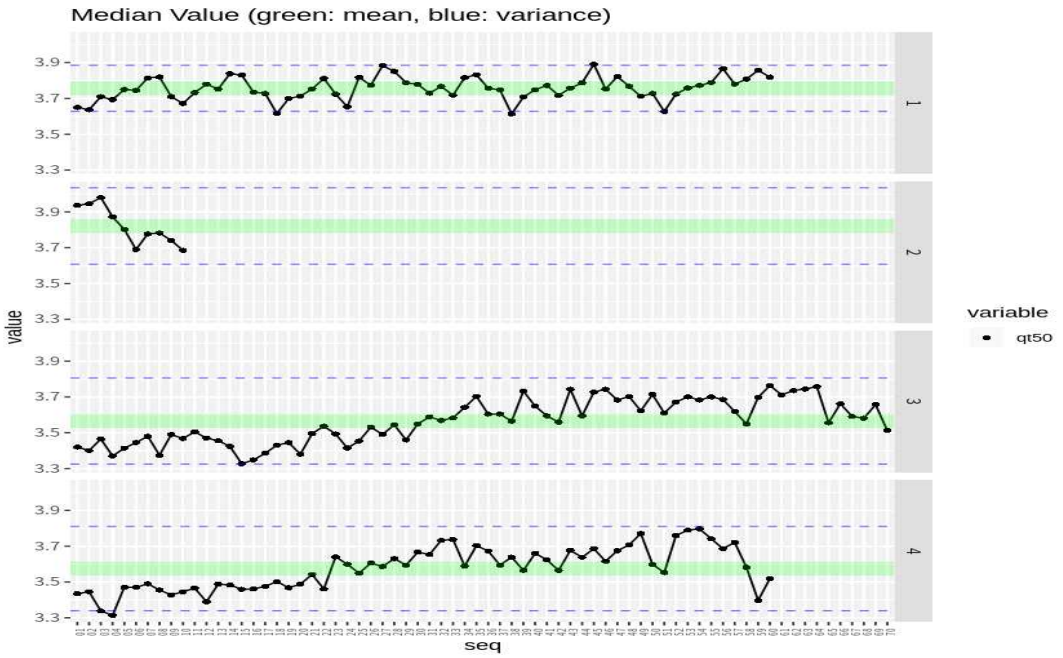
○ 세트별 응답의 특성 차이 분석

본 분석의 목적은 서로 다른 어휘쌍에 대한 평가를 하였음에도 불구하고, 세트간의 응답에 대한 유의미한 편차가 존재하는지를 검증하기 위한 것이다. 이를 위해 각 응답의 중위수 값(median), 그리고 평균을 표준편차로 나눈 계수(beta coefficient)를 비교하였다.

첫 번째 중위수는 평균의 왜도가 있는 경우에 이를 보완할 수 있는 수치이며, 평균을 표준편차로 나눈 계수는 평균의 분포를 더 확대하여 정밀히 살펴볼 수 있는 지표이다.

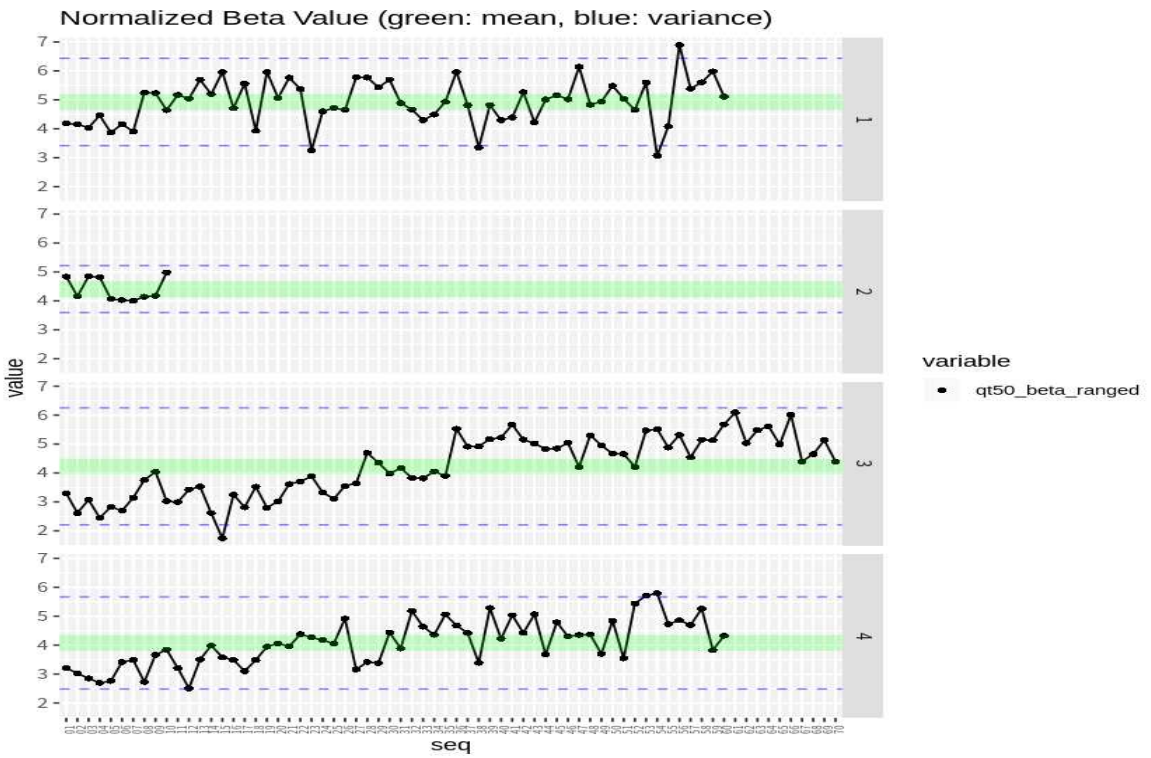
중위수 관점에서 살펴보면 (<그림 12>의 첫 번째 그래프의 녹색 부분), 반대말의 중위수가 다른 어휘쌍보다 높은 것을 알 수 있는데, 이는 반대말이 유사어보다 더 거리가 멀게 인지된다는 일반적인 견해와 다른 부분이다.

이는 Hill et al.(2015) 등과 달리, 반대말에 대해 “얼마나 더 반대가 되는 개념입니까?”로 명확하게 질문하여 응답자가 좀 더 분명하게 인지할 수 있었고 상대적으로 반대말에 속한 어휘 난이도가 낮았기 때문인 것으로 추정된다.



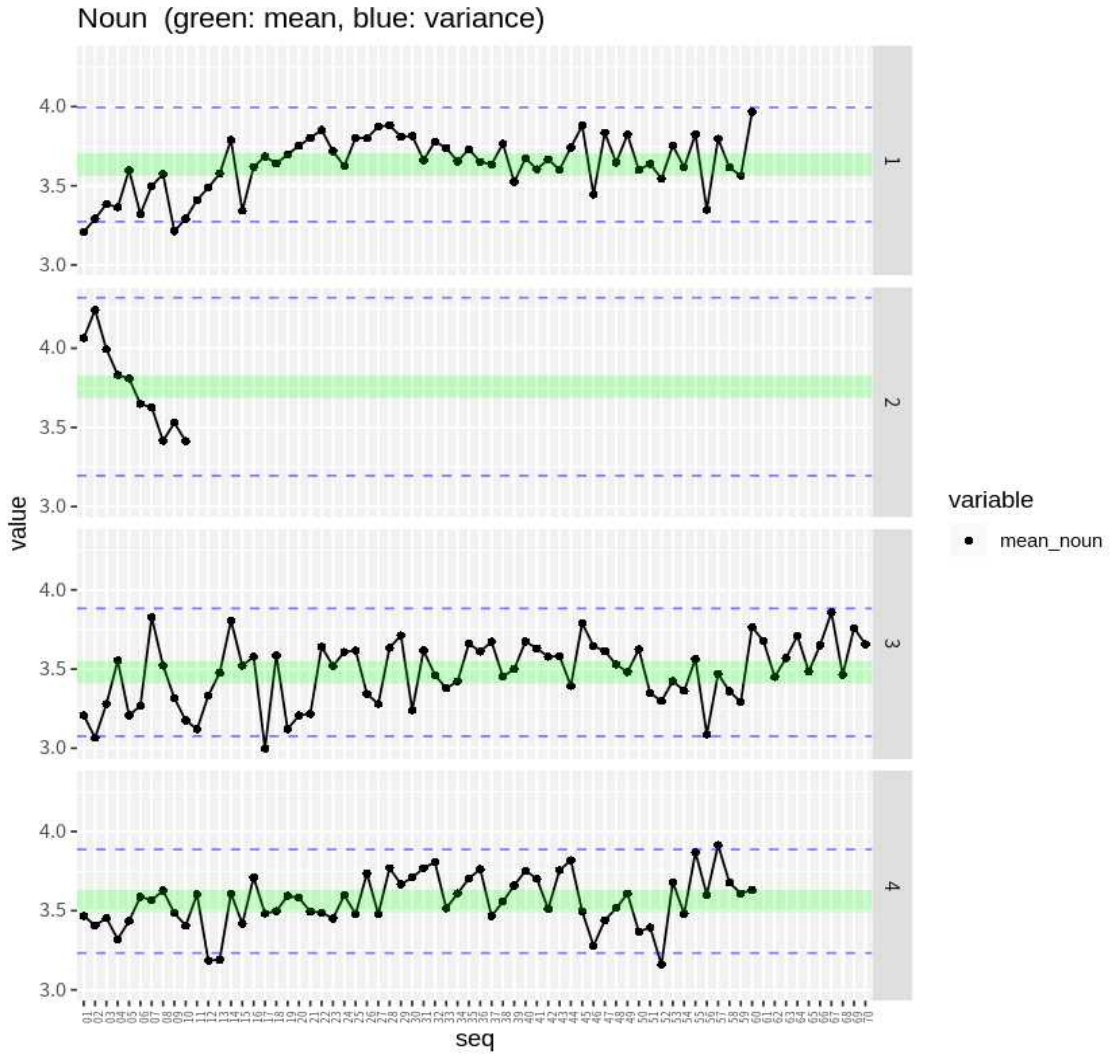
〈그림 12〉 세트별 중위수 검토

두 번째인 베타계수를 통해 반대말이 비슷한말보다 낮은 경향을 다시 확인할 수 있다. 즉, 전반적인 중위도는 높지만, 비슷한 말보다는 더 높은 표준편차를 보여주고 있어 베타계수를 낮추는 요인이 되기 때문이다.



〈그림 13〉 세트별 베타계수 검토

세 번째 그래프인 품사별 비교에서는 우리말 사전의 표제어 대부분이 명사이므로, 전 세트에 걸쳐서 추이를 확인할 수 있는 명사로 한정하여 분석하였다. 그 결과 반의어와 다른 어휘 관계 사이에 극명한 차이를 보인다. 평균의 차이는 없으나, 각 세트별로 편차가 높다. 이를 통해 비교적 변동 폭이 좁은 명사의 경우에도 반의어에 대한 응답은 변동성이 높다는 것을 알 수 있다.



〈그림 14〉 세트별 명사 - 평균 검토

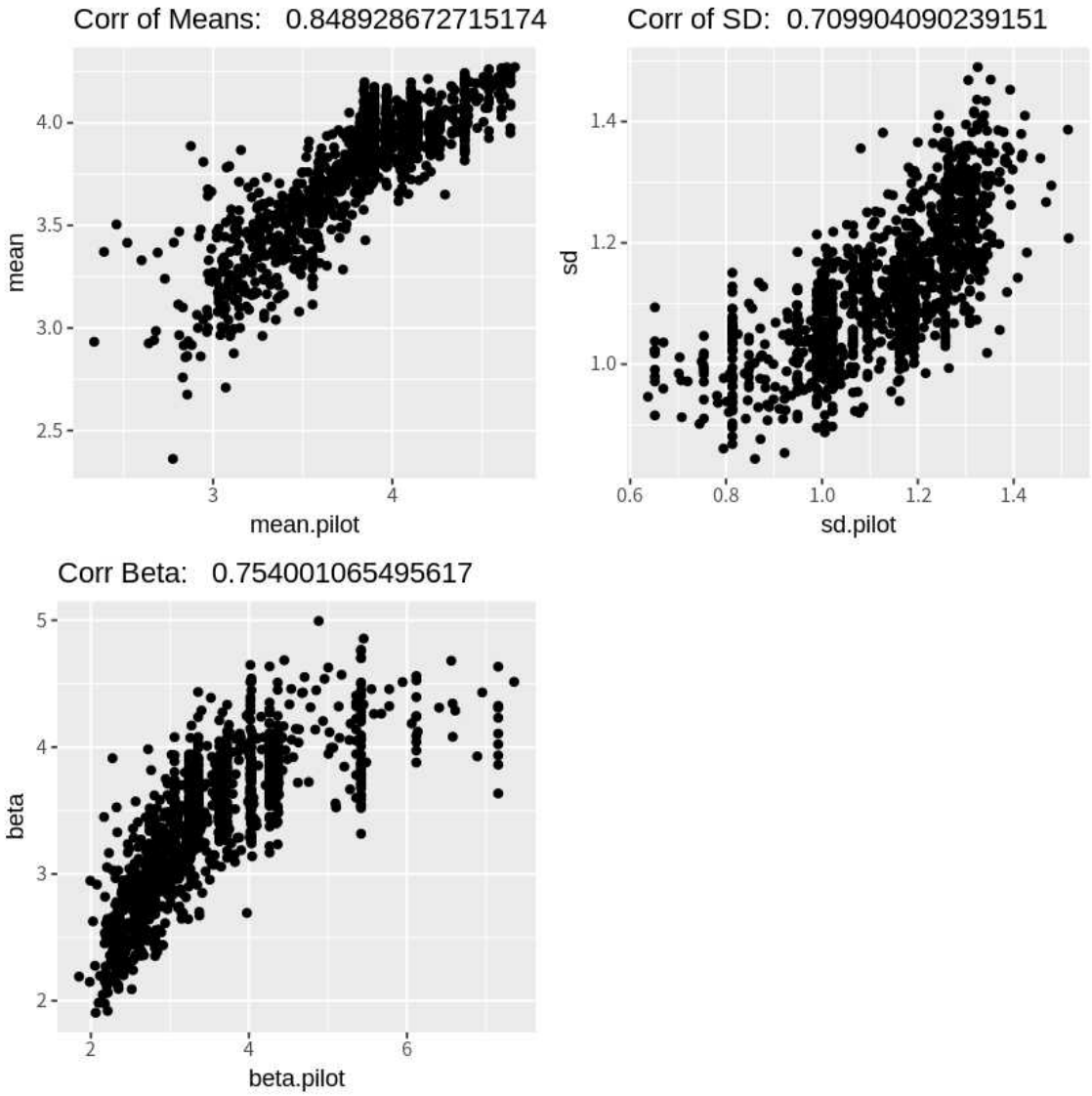
○ 최종 설문 품질 분석

본 검증은 최종적으로 응답된 결과에 대한 품질을 측정하기 위한 것이다. 기존의 연구자들(Hill et al. 2015, Zesche & Gurevych 2009)은 기존의 응답 결과를 wordSim-353을 기준으로 한 상관 분석을 통하여 설문 결과의 품질을 평가하고 보고하였다.

본 설문 조사 이전에 한국어 어휘쌍에 대한 평가가 전무한 관계로, 시험 공정(파일럿 테스트)으로 수집된 총 200 응답자의 1,000쌍을 본 공정에서 수집된 총 200 응답자의 결과와 비교하여 품질 분석을 하였다. 상관 분석 대상은 일반적으로 수행하는 평균 외에도 각 응답의 표준편차 및 베타계수(평균/표

준편차)를 모두 고려하였다.

아래 그래프에서 확인할 수 있듯, 평균에 대한 상관계수(0.84)는 물론, 표준편차와 베타계수에 대한 상관계수 모두 고품질 설문 결과 수용 기준점(cut-off) 기준인 0.7을 만족하는 것을 알 수 있다.



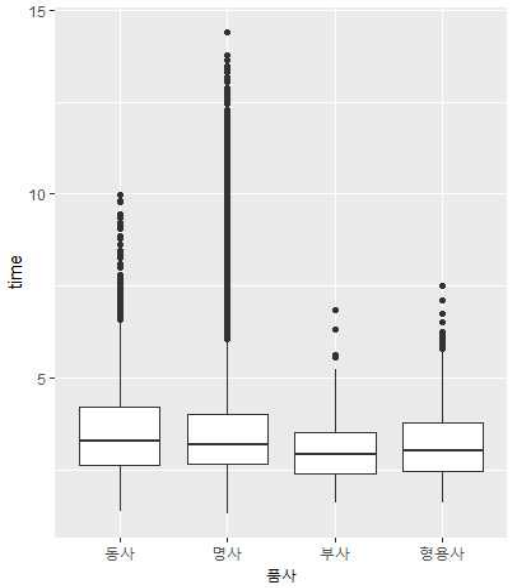
〈그림 15〉 최종 설문 품질 결과

나. 응답 시간 분석

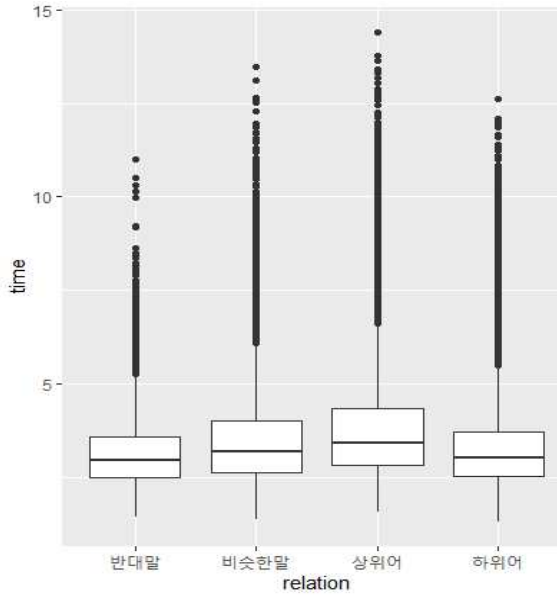
제공되는 각 설문에 대해 응답자가 해당 설문에 대한 완료 버튼을 누른 순간을 기준으로 하여 전후 시간차를 통하여 응답 시간을 추정하였다.

1) 개요

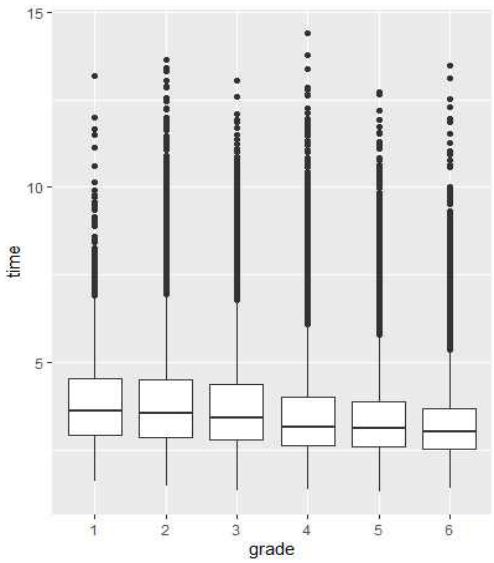
품사별 응답 시간은 부사가 가장 짧았고 동사가 가장 길었다(부사<형용사<명사<동사). 어휘 관계별로는 '반대말 < 하위어 < 비슷한말 < 상위어'순으로 응답 시간이 길어졌다. 한편 난이도별로는 등급이 높아질수록 그에 반비례하게 응답 시간이 짧아졌다.



〈그림 16〉 품사별 응답 시간



〈그림 17〉 어휘 관계별 응답 시간



〈그림 18〉 난이도별 응답 시간

2) 범주별 평균 응답 시간

품사	time
부사	2.999199
형용사	3.201965
명사	3.475681
동사	3.532048

〈표 34〉 품사별

relation	time
반대말	3.16167
하위어	3.263714
비슷한말	3.439604
상위어	3.72911

〈표 35〉 어휘 관계별

grade	time
1	3.836388
2	3.836372
3	3.72887
4	3.472058
5	3.374724
6	3.210394

〈표 36〉 난이도별

3) 교차 비교

품사를 기준으로 보았을 때, 어휘 등급에 따른 응답 시간은 일관성을 찾기 어렵다. 평균을 보았을 때, 6등급의 어휘에 대한 응답 시간이 짧을 것으로 기대되었으나 품사별로 차이를 보이기도 하고(형용사는 반대의 분포를 보임) 동일 품사 내에서도 어휘 등급에 따른 순차적인 변화를 보이지 않는 경우도 나타났다(부사, 동사).

<품사별>

grade	time
1	3.8775
2	3.838077
3	3.721352
4	3.471281
5	3.3619
6	3.204903

〈표 37〉 명사난이도별

grade	time
1	2.480257
2	3.184584
3	3.296118
4	3.263421
5	3.375602
6	3.113706

〈표 38〉 형용사난이도별

grade	time
1	3.063158
2	3.020106
3	3.04519
4	3.093521
5	3.072338
6	2.896735

〈표 39〉 부사난이도별

grade	time
1	3.678301
2	4.188651
3	3.991209
4	3.937489
5	3.633012
6	3.2951

〈표 40〉 동사난이도별

<어휘 관계-난이도별>

grade	time
1	2.959137
2	3.265282

grade	time
1	4.12357
2	4.077095

grade	time
1	4.042925
2	4.136513

grade	time
1	3.531665
2	3.52212

3	3.340175
4	3.012814
5	3.169265
6	3.102256

〈표 41〉 반대말 난이도별

3	4.075875
4	3.679317
5	3.674686
6	3.193149

〈표 42〉 비슷한말 난이도별

3	3.872108
4	3.670863
5	3.464127
6	3.391368

〈표 43〉 상위어 난이도별

3	3.477536
4	3.08296
5	3.072931
6	3.177162

〈표 44〉 하위어 난이도별

<품사-어휘 관계별>

relation	time
반대말	3.155707
비슷한말	3.431805
상위어	3.72893
하위어	3.263694

〈표 45〉 명사어휘 관계별

relation	time
반대말	2.531092
비슷한말	3.367195
상위어	3.604038

〈표 46〉 형용사어휘 관계별

relation	time
반대말	2.498533
비슷한말	3.020263

〈표 47〉 부사어휘 관계별

relation	time
반대말	3.442327
비슷한말	3.53922
상위어	5.355744
하위어	3.845553

〈표 48〉 동사어휘 관계별

이상에서 품사나 어휘 관계에 비해서 어휘 등급의 응답 시간은 예상과 달리 등급이 올라갈수록 응답 시간이 짧아지는 경향을 보인다.

품사는 부사와 형용사가 응답 시간이 짧은 데 비해, 명사나 동사는 상대적으로 응답 시간이 길다. 부사와 형용사는 전형적인 반대말이 다수 포함되어 있기 때문인 것으로 보이고, 이와 달리 명사와 동사는 상하위관계나 비슷한말과 같이 기본적으로 관계 판단이 반대말에 비해 어렵기 때문인 것으로 보인다.

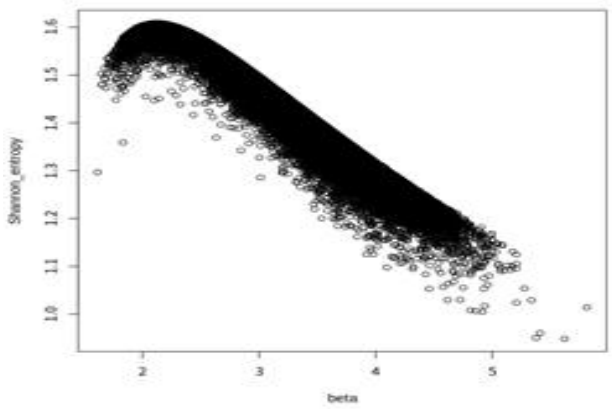
어휘 관계에 있어서 다른 관계에 비해 반대말은 품사나 어휘 등급과 상관없이 응답 시간이 일관성 있게 가장 짧았다. 이는 기본적으로 반의관계가 이분법적인 의미관계를 맺고 있어 어휘 관계에 대한 판단이 쉽기 때문에 응답 시간이 짧은 반면, 나머지 어휘 관계는 다양한 어휘항목과 관계를 맺고 있거나(비슷한말), 관계에 대한 개념적 지식체계를 요구하기(상위어, 하위어) 때문인 것으로 보인다.

다. 정보량과 무질서도 분석

주어진 어휘쌍에 있어 응답자가 답변하는 유형은 전체 답변에 넓게 분포되어 있거나(200명의 응답에 대해 약 40명이 1번을, 40명이 2번을, 40명이 3번을, 40명이 4번을, 40명이 5번을 선택하는 경우), 특정한 답변이 몰려 있을 수 있다(200명의 응답에 대해 200명 모두가 3번을 선택하는 경우). 전자와 후자의 경우 단순히 평균을 구하면 3으로 동일하지만, 표준편차가 더 큰 전자의 응답은 본 연구에서 제안한 베타계수에 의하면 더 많은 제재(페널티)를 받아 더 낮은 점수를 가지게 된다. 후자의 경우는 극단적으로 무한대의 값을 가지게 된다.

한편, 전자의 사례는 응답에 존재하는 정보의 양이 많다는 것을 의미함과 동시에 응답이 내포하고 있는 무질서도가 더 높다는 것을 의미한다. 정보 이론에 따르면 정보가 전달할 수 있는 양을 측정하기 위해 사용되는 정보 엔트로피를 통해서 무질서도를 측정할 수 있다. 이러한 측정에 가장 많이 사용되는 공식이 샤논의 엔트로피이다. 이는 확률과 로그 함수의 결합을 통해서 확률이 1과 0인 경우에 대한 제재(페널티)를 주는 방법이다. 로그 함수의 밑이 2인 경우가 정보의 단위인 비트로 알려져 있다.

본 연구에서는 각 응답에 포함되는 답변의 비율을 활용하여 샤논 엔트로피 계수를 도출하였다. 아래 그림과 같이 각 20만 쌍 설문에 대한 엔트로피 지수는 약 1~1.6까지로 분포하였으며, 베타계수와의 관계는 위로 볼록한 2차 함수형이다.



〈그림 19〉 20만 쌍에 대한 엔트로피 지수 분포

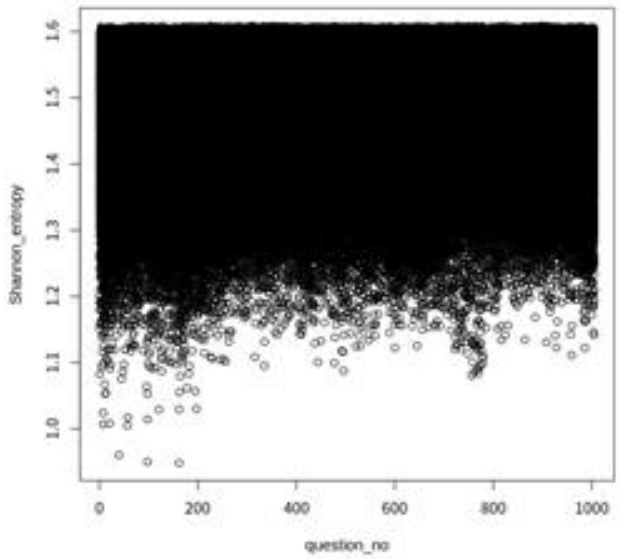
엔트로피가 높은 어휘쌍은 단어 자체는 쉽더라도 베타계수가 낮기 때문에 어느 한 쪽으로 확실하게 답변하기가 어려운 의미를 가지는 단어들의 쌍이었다.

full_ID	난이도등급	베타계수	사분위수 등급	엔트로피
3060826 [306]_242969: 젖빛-흰빛	3	2.138344	약	1.609188
3330897 [333]_249790: 최고-시간	2	2.129136	약	1.609157
3180727 [318]_233900: 우편-업무	2	2.125968	약	1.60912
3200309 [320]_196564: 귀신-넋	1	2.118063	약	1.60906
3100353 [310]_201150: 뇌물-물건	2	2.107438	약	1.609019

〈표 49〉 상위 엔트로피 어휘쌍 예시

주어진 문제는 1,000쌍에 대한 문제 및 5쌍의 일관성 확인 문제인 1,005쌍으로 구성된다. 피로도나 응답 학습효과 등으로 초반에 비해 후반부에 정보 손실이나 무질서도의 증감이 발생하였는지 여부를 확인하기 위하여 엔트로피와 문제 번호의 관계를 확인하였다.

아래 그림과 같이 엔트로피와 문제 번호 사이의 상관관계는 발견되지 않았으며, 응답자들은 문제의 위치와 상관없이 일관된 응답 패턴을 보였음을 확인할 수 있었다.



〈그림 20〉 엔트로피와 문제 번호 사이의 상관관계

6. 어휘별 등급화

가. 개요

5절에서는 데이터 품질을 검증하기 위해 세트별 응답 및 일관성 문제의 평균과 중위수, 표준편차, IAA 등을 살펴보았고, 해당 어휘쌍의 특징을 살펴볼 수 있는 어휘쌍별 응답 시간 및 엔트로피 지수를 검토하였다. 특히 어휘쌍별 응답 시간에 대한 분석에서는 품사 및 어휘 관계별 분석은 물론 어휘 고유의 난이도에 따른 난이도 등급별로 분류하였다.

이 중 난이도 등급은 어휘가 갖는 고유의 난이도에 근거하여 어휘쌍에 난이도를 부여한 것이다. 즉, 두 단어가 모두 난이도 '하'인 경우는 1, 둘 중의 하나가 '하'이고 나머지 하나가 '중'인 경우는 2, '중-중'인 경우는 3, '하-상'인 경우는 4, '중-상'인 경우는 5 '상-상'인 경우는 6이다.

어휘에 대한 등급화는 산업의 요구에 따라 다양한 방안으로 산정이 가능하다. 예를 들어 각 어휘가 어휘 관계망에서 차지하는 중심성(centrality)에 각 어휘 관계별(edge) 가중치(weight)를 적용하는 방법을 고려할 수 있다. 이 경우, 중심성은 등급화가 필요한 업무(비즈니스) 영역에 부합하는 각종 중심성(예를 들어, betweenness, closeness, eigen vector 등)을 고려해야 하며, 가중치는 앞 절에서 제시한 다양한 수치를 고려하여 제시할 수 있다(예를 들어, 난이도 등급, 응답 시간, 응답 평균, 정보 엔트로피 등).

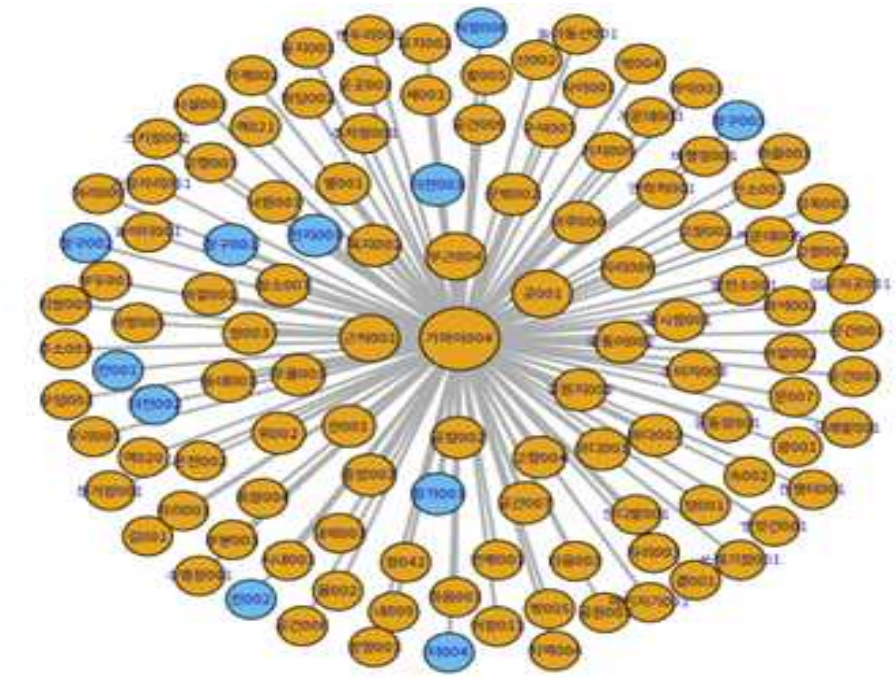
이 경우 요구되는 업무(비즈니스) 영역이 '대응어 제시를 위한 유의어 사전'을 도출하는 것이라면, 『우리말샘』의 어휘 간의 관계에만 의존하는 등급화는 적용에 한계가 있을 수 있다. 예를 들어 '가까이'와 관련된 어휘는, 『우리말샘』에 의하면 반대말인 '멀리'와 상위어인 '곳' 2가지의 경우가 있으며, 이

경우 상기의 중심성과 가중치를 평면적으로 적용한 등급화는 '가까이'와 대체하여 사용할 수 있는 어휘(예를 들어 '근처')를 즉각적으로 제시하기가 어렵다.

pair_id	head.word_id	head.voca	head.의미어	head.의미번호	tail.word_id	tail.voca	tail.의미어	tail.의미번호	relation
122362	250	가까이	가까이 004	4	59826	멀리	멀리 004	4	반대말
187855	250	가까이	가까이 004	4	15834	곳	곳 001	1	상위어

〈표 50〉 '가까이004' 관련어

다시 말해, 산업이 필요로 하는 다양한 업무(비즈니스) 요구를 모두 고려한 단 하나의 해결책으로서의 완벽한 등급화 원칙은 존재하기 어렵다. 본 사업을 통해 제시된 어휘의 기초 정보를 활용하여 해당 산업과 기업에서 다양한 알고리즘을 강구하고 서비스에 적용할 수 있을 것으로 기대된다. 예를 들어, 아래 그림은 기초적인 『우리말샘』 어휘 관계에 근거하여 ㈜나라지식정보 인문-AI연구소에서 제시한 '가까이'와 관련된 유의어 및 어휘별 등급을 도식화한 것이다 표제어인 '가까이004' 근처에 더 의미적으로 유사한 단어의 정보가 표출되어 있다.



〈그림 21〉 '가까이'와 관련된 유의어 및 어휘별 등급 도식화

*어휘의 Lexical Cohesion과 구조적 네트워크 중심성 알고리즘(나라지식정보연구소) 적용 및 계산 예시(주황색: 1등급 어휘, 파란색: 2등급 어휘)

본 절에서는『우리말샘』사전 정보의 시각화 방안을 고려하여, 어휘쌍에 대한 등급화 원칙 및 등급화 결과를 제공한다. 이 같은 등급화는 어휘쌍에 대한 사용자 인지 평가를 기초로 하여 제시한다.

등급화를 위해 본 컨소시엄이 제시하는 원칙은 ①정보의 중복성을 최소화하고, ②사용자 평가의 특성을 나타내는 정보를 최대한 제공하는 것이다.

나. 사분위수(Quantile) 기준

1) 등급화 대상

설문 조사를 통해 수집된 전체 어휘쌍 총 200,000에 대한 등급화를 실시하였다. 일관성 확인을 위한 문제에 있어서는 수집된 세트 중 표준편차가 가장 작은 응답을 선별하였으며, 최종적으로 수집된 데이터는 <표 51>과 같다.

분류	품사	소계	비율	계
비슷한말	동사	8,524	4%	60,000
	명사	49,525	25%	
	부사	927	0%	
	형용사	1,024	1%	
반대말	동사	851	0%	10,000
	명사	8,856	4%	
	부사	39	0%	
	형용사	254	0%	
상위어	동사	8	0%	70,000
	명사	69,989	35%	
	형용사	3	0%	
하위어	동사	2	0%	60,000
	명사	59,998	30%	
계				200,000

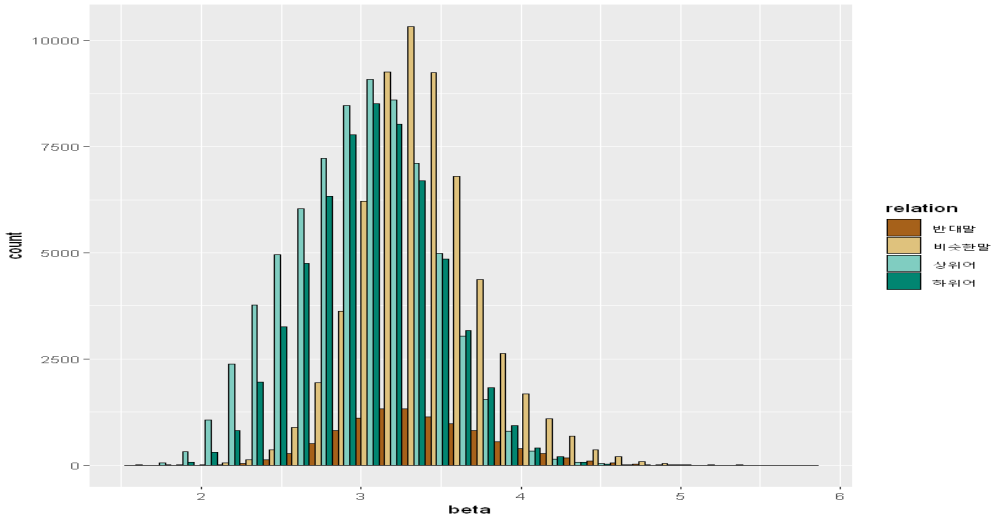
〈표 51〉 관계 유형/품사별 어휘쌍 분포

위에서 확인할 수 있듯이, 상/하위어는 주로 명사로 구성되어 있으며, 비슷한 말/반대말의 경우에도 명사의 개수가 단연 압도적이다. 이와 같은 상태에서 등급화는 어휘쌍의 종류 및 품사의 종류에 따라서 구분되어야 한다.

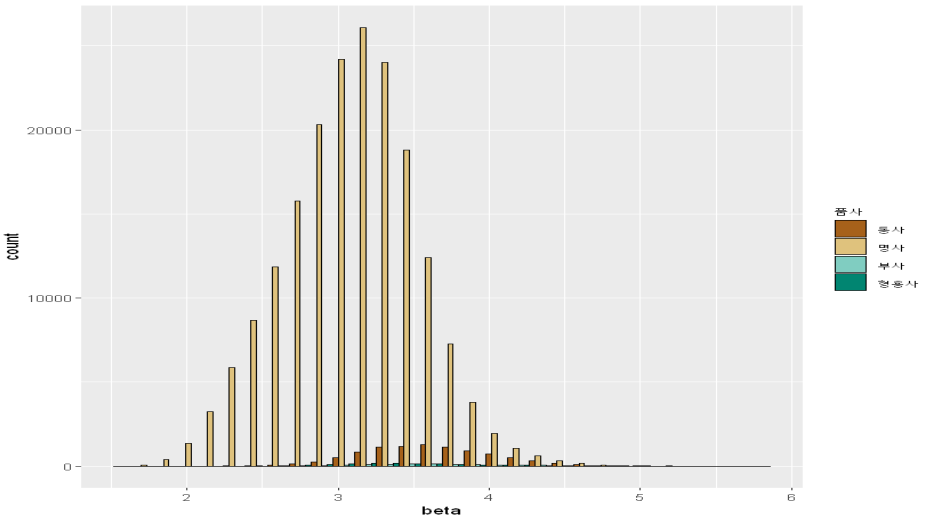
2) 등급화 기준

등급을 부여하기 위한 기준으로 평균 및 표준편차를 고려하였으며, 동일 평균이라고 하더라도, 변동성 요소에 좀 더 많은 제재(페널티)를 주기 위하여 베타계수(평균을 표준편차로 나눈 계수)를 각 어휘 종류별, 품사별 분포를 확인하였다. 베타의 범위는 표준편차에 따라 무한대로 확장 가능하나, 설문에서 파악된 최대 베타계수는 약 5.8이다.

베타계수의 분포는 어휘쌍의 종류 및 품사의 종류에 따라 정규분포형의 종모양의 곡선을 보여주고 있다. 또한 속한 데이터의 개수에 따라 t-분포가 가지는 패턴을 보이는데, 반대말과 같이 데이터의 개수가 작은 분포인 경우는 조금 더 납작한 종모양의 분포가 관찰되었다.



〈그림 22〉 관계 유형별 베타계수 분포



〈그림 23〉 품사별 베타계수 분포

다음으로는 각 어휘쌍 종류별, 품사별로 구분된 베타계수에 대한 사분위수를 기준으로 각각 66%, 33% 지점을 기준으로 하여, 각각의 위치에 속하는 어휘쌍을 각각 '강(높은 관련성)/중/약(낮은 관련성)'으로 구분하였다. 각 관계 유형과 품사별 사용자 응답의 차이는 관련 통계적 분석을 통해 면밀한 분석이 가능할 것이다. <표 52>에서 대략적인 관계 유형과 품사별 경향을 살펴 볼 수 있다.

예를 들어, 비슷한말과 반대말에 비해 상위어와 하위어에 대한 베타계수가 더 낮은 경향이 뚜렷하게 확인되었다. 형용사의 경우, 비슷한말은 다른 품사에 비해 더 낮은 계수(더 낮은 관련성)를, 반대말의 경우 다른 품사에 비해 더 높은 계수(더 높은 관련성)를 보인다. 향후 본 기초 데이터를 기반으로 한 서비스 개발이나 학술 연구를 통해 더욱 풍부하고 흥미로운 시사점을 제공할 것으로 기대한다.

분류	품사	개수	중위수	높은 관련성 (Quantile66 평균)	낮은 관련성 (Quantile33 평균)
비슷한말	동사	8,524	3.60	3.77	3.41
	명사	49,525	3.33	3.46	3.19
	부사	927	3.63	3.84	3.43
	형용사	1,024	3.33	3.50	3.17
반대말	동사	851	3.61	3.83	3.41
	명사	8,856	3.30	3.50	3.13
	부사	39	3.60	3.72	3.36
	형용사	254	3.94	4.20	3.67
상위어	동사	8	2.38	2.53	2.23
	명사	69,989	2.98	3.16	2.77
	형용사	3	2.24	2.64	2.16
하위어	동사	2	2.53	2.54	2.53
	명사	59,998	3.05	3.21	2.87
평균			3.19	3.38	3.03

〈표 52〉 관계 유형/품사별 개수 및 등급별 기준점

3) 등급화 결과

위에서 수립된 기준에 따라, 분류된 최종 등급화 결과는 다음과 같다. 각 어휘 관계별 품사별 분포를 기준으로 부여하였다. 본 등급은 향후 『우리말샘』 사전의 시각화에 표제어와 관련어 간의 관계에 대해 '강/중/약'으로 직관적인 표기 및 이해가 가능할 것으로 기대한다.

구분	사분위수 등급	개수	비율
비슷한말	강	20,399	34%
비슷한말	중	19,801	33%
비슷한말	약	19,800	33%
반대말	강	3,400	34%
반대말	중	3,299	33%
반대말	약	3,301	33%
상위어	강	23,800	34%
상위어	중	23,098	33%
상위어	약	23,102	33%
하위어	강	20,400	34%
하위어	중	19,799	33%
하위어	약	19,801	33%

〈표 53〉 관계 유형별 등급별 어휘쌍 개수

어휘쌍 분류별로 가장 높은 베타계수를 가진 어휘쌍은 다음과 같다. 상대적으로 배치한 어휘에 전문어가 많이 포함된 상위어가 있으며 하위어 그룹에는 표제어와 관련어 중에서 최소 한 개가 어려운 단어일 가능성이 높다. 흥미로운 점은 반대가 될 것으로 예측되는 어휘쌍의 경우 난이도가 쉬운 어휘가 아닌 비교적 난이도가 높은 어휘가 포함되어 있다는 것이다(예를 들어 '낙선하다-당선하다'가 '죽이다-살리다'보다 더 반대되는 말로 이해함). 반대말에 대한 대중의 인식이 다른 일반적인 어휘 관계보다 의미와 폭이 더 좁은 어휘일수록 잘 드러난다는 것은 시사하는 바가 크다고 하겠다.

분류	어휘쌍	품사	*난이도 등급	베타계수	사분위수 등급
비슷한말	어머니-엄마	명사	1	5.811976	강
반대말	낙선하다-당선하다	동사	3	5.336234	강
상위어	요골-뼈	명사	4	5.620987	강
하위어	그림-상상도	명사	4	5.168991	강

※ 참고: 난이도 등급 (하하:1, 중하:2, 중중:3, 상하:4, 상중:5, 상상:6)

〈표 54〉 관계 유형별 최고 베타계수 어휘쌍

모든 어휘쌍에 대해서 베타계수가 가장 높은 순으로 나타난 결과는 다음과 같다. 아래의 난이도 등급은 해당 어휘가 가지는 고유한 난이도를 기준으로 부여한 것이다. 이중 난이도 등급은 어휘별 난이도에

따른 것이며, '어머니-엄마'의 경우 두 어휘가 '하(쉬움)' 등급의 어휘이므로 가장 낮은 어휘도 등급인 '1' 등급이 부여되어 있다. '요골-뼈'의 경우, '요골'은 '상(어려움)', 뼈는 '하(쉬움)' 등급으로 난이도 등급 '4' 등급이 부여되어 있다.

'셰퍼드'는 '중' 등급으로 '세퍼드'와 '개'는 2등급이다(하하:1, 중하:2, 중중:3, 상하:4, 상중:5, 상상:6). 베타계수가 가장 높은 10개의 단어에는 1등급부터 6등급까지의 어휘쌍이 모두 등장한다. 난이도 등급과 베타계수와의 관계는 "다. 관련 정도를 이용한 어휘별 등급화"에서 더욱 상세히 논의한다.

full_ID	품사	난이도 등급	베타계수	사분위수 등급
1070098 [107]_127826: 어머니-엄마	명사	1	5.811976	강
3150162 [315]_95017: 요골-뼈	명사	4	5.620987	강
3150040 [315]_21232: 셰퍼드-개	명사	2	5.409537	강
3150098 [315]_51520: 새송이버섯-버섯	명사	4	5.376529	강
2020121 [202]_121670: 낙선하다-당선하다	동사	3	5.336234	강
1060013 [106]_9524: 사이버대-사이버대학교	명사	6	5.276552	강
1070252 [107]_139118: 곰탕-곰국	명사	3	5.208544	강
1010008 [101]_9475: 기말시험-기말고사	명사	5	5.208068	강
1070521 [107]_163271: 추천서-추천장	명사	3	5.205185	강
1030335 [103]_146476: 물안경-수경	명사	5	5.205082	강

〈표 55〉 최고 베타계수 어휘쌍

다. 관련 정도를 이용한 어휘별 등급화(어휘쌍에 대한 관련 정도 결과 검토)

관련 정도를 이용한 어휘별 등급화는 설문 조사가 완료되고 유효한 데이터로 확정된 200개 세트(200,000 어휘쌍)의 설문 조사 결과를 토대로 관련어별로 관련 정도를 반영하여 어휘를 등급화한 것이다. <표 56>은 어휘별 등급화에 사용한 데이터에 대한 기본 통계 정보이다.

구분	어휘쌍 개수	평균(5점 척도)	표준편차
비슷한말	60,000	3.73	1.11
반대말	10,000	3.80	1.14
상위어	70,000	3.55	1.21

하위어	60,000	3.57	1.18
총합계	200,000	3.62	1.17

〈표 56〉 등급화에 사용된 데이터 수

1) 관련 정도 기준

관련 정도의 기준은 5점 척도에 의한 설문 조사 결과의 응답 평균을 가지고 <표 57>과 같이 정한다. 설문에 응답하는 분들이 직관에 의해 5점 척도 중 하나를 선택했다는 전제하에 이 기준(절대기준)을 정한 것이다.

구분	기준
하(관련성 낮음)	응답평균 〈 2.3
중(관련성 보통)	2.3 〈= 응답평균 〈 3.7
상(관련성 높음)	응답평균 〉= 3.7

〈표 57〉 관련 정도 기준

2) 어휘별 등급화 결과

관련 정도의 기준에 따라 어휘를 등급화한 후 그 결과를 보면 <표 58>과 같다. 비슷한말과 반대말은 '관련성_상'에 응답한 비율이 60% 이상이지만 상위어와 하위어는 '관련성_중'에 응답한 비율이 60% 이상으로 나왔다. 이것은 일반인들이 비슷한말과 반대말이라는 개념에는 익숙하지만 상위어와 하위어라는 개념에는 익숙하지 않은 결과라고 볼 수 있다. 상위어와 하위어의 개념을 잘 알고 있는데도 그런 결과가 나왔다면 그것은 하위어들을 대표하는 상위어가 하위어 모두를 포함하는 개념이 아니기 때문일 가능성이 높다. 이런 상위어들은 다른 단어로 교체하여 평가하는 것이 결과의 신뢰도를 더욱 높일 수 있을 것이다.

구분	관련성 등급(개수)				관련성 등급(비율)			
	상	중	하	총합계	상	중	하	총합계
비슷한말	36,915	23,084	1	60,000	61.53%	38.47%	0.00%	100%
반대말	6,793	3,207		10,000	67.93%	32.07%	0.00%	100%
상위어	23,059	46,937	4	70,000	32.94%	67.05%	0.01%	100%
하위어	20,135	39,865		60,000	33.56%	66.44%	0.00%	100%
총합계	86,902	113,093	5	200,000	43.45%	56.55%	0.00%	

〈표 58〉 관련정도 이용 어휘 등급화 결과 통계

관련 정도의 기준에 따라 어휘를 등급화한 결과를 관련어별로 관련어가 많은 어휘 위주로 정리한 것이 <표 59>(비슷한말), <표 60>(반대말), <표 61>(상위어), <표 62>(하위어)이다. 정리된 결과를 보면 관련어가 많은 하위어를 이 기준에 따라 구분하는 것이 효과적임을 알 수 있다. 비슷한말, 반대말, 상위어도 개별 어휘별로 일반인이 생각하는 관련 정도를 파악할 수 있어 의미가 있다고 볼 수 있다.

표제어	관련성_상	관련성_중	관련성_하
감례001		광악003,구술001,국생001,도깨빗국001,두강001,마주보기001,망우물001,맥곡지영001,미록002,벌성지광악001,부의004,아랑주002,어르광이001,어리광이001,홍우001,화천008	
디시안디아미드001	다이사이아노다이아마이드001,다이사이아노다이아미드001,다이시안다이아마이드001,다이시안다이아미드001,다이시안디아미드001,디시안다이아마이드001	다이시아노다이아마이드001,다이시아노다이아미드001,다이시아노디아미드001,디사이아노다이아마이드001,디시아노다이아마이드001,이사이아노구아니딘001,이시아노구아니딘001,이시안구아니딘001	
사021	잠매001,장면001,죽음001	귀토001,대귀004,반진001,사세007,영면001,영서003,장서003,졸004,화거001,획린001	
거년001	안해001	객세001,고세003,과년006,석년002,선년001,전세005,전재006,전해001,지난해001	
능인적묵001		모니001,본사008,불016,불성005,붓다009,석가004,세존001,정진002,천존001	

〈표 59〉 관련정도 이용 어휘 등급화 결과(비슷한말)

표제어	관련성_상	관련성_중	관련성_하
자옥세계001	극락계001,극락안양정토001,낙방003,무량청정토001,보계004,안락세계001,안양세계001,연화세계001,정역011	금색세계001,안락국001,안락정토001,안양003,안양계001,안양보국001,안양정토001,열반성001,천수국002	
극한003	극염001,대서002,대열003,된더위001,맹서001,서염	융서001,혹리002	

	001, 엄서001, 열서003, 염열001, 혹서001, 혹양002, 혹열001, 혹염001		
단수011	노수003, 대수008, 대춘지수001, 만수002, 만수006, 영수001, 용수009, 장생001, 장생구시001, 호수012	수령005, 하년002	
세말003	세초003, 연초001	설003, 세수006, 세시003, 수세016, 연두001, 연수002, 연시002, 정초001, 조세006, 조세009	
이승001	저승001, 지부002, 천양002, 황로003, 황천005	시왕청002, 유명006, 음부004, 중천004, 천대002, 황양002, 황토004	

〈표 60〉 관련정도 이용 어휘 등급화 결과(반대말)

표제어	관련성_상	관련성_중	관련성_하
알롱이001	무늬001	물건001, 점016, 짐승001	
얼루기001		무늬001, 물건001, 점016, 짐승001	
길005		목적004, 방향002, 전문 분야001, 지침003	
여닫이001	문007	방식001, 창문001	
최상위001	등급001	위치001, 지위004	

〈표 61〉 관련정도 이용 어휘 등급화 결과(상위어)

표제어	관련성_상	관련성_중	관련성_하
일001	가사005, 근로003, 꽃게잡이001, 날품팔이001, 노역001, 바느질001, 사무003, 삯일001, 새우잡이001, 잡역002, 잡일001, 포함하여 총 19개	갚음001, 개국004, 걷기001, 검거002, 겨울맞이001, 경기018, 계산001, 고리대금업001, 고수레002, 고수레003, 고시005, 고역001, 고용003, 곰돌이001, 공무005, 공사002, 공소의제기001, 공일001, 괭이질001, 교학002, 구구004, 구사002, 국사006, 군사005, 궂은일002, 근심거리001, 글다듬기001, 금물006, 기소003, 기적003, 기찰004, 꽃꽂이001, 꽃모종002, 꾸밈002, 나랏일001, 포함하여 총 604개	
일010	소일거리001, 아르바이트001	간산003, 같은자리001, 개계001, 검속002, 겸업001,	

		경숙002,계책002,과락001,급가속001,기회균등001,길쌈질001,끝마무리001,남북통일001,남자관계001,노가리002,노래자랑001,눈도장001,눈요기001,다대일001,단판걸이001,달구경001,달구질001,달맞이001,당질001,대역007,대용갈이001,대절006, 포함하여 총 504개	
상태001	뇌사003,만취004,백지상태002,백지상태003,사몽비몽001,영양실조001,잠결001,중태002,팽만001,포화009, 포함하여 총 19개	가식적001,가정적003,간접적001,감염002,강경007,개괄적001,개념적001,개방004,개방적001,개별적001,개연008,개인적001,개체적001,거미001,거침새001,건납001,결002,결딴001,결손001,겹001,경색002,경이적001,경제적002,경지009,경직004,경험적001,계절적001,계획적001,고갈003,고난001,고립001, 포함하여 총 479개	
부분001	돌출부001,뒷부분001,뒷부분002,몸체001,미부001,밑부분001,상반부001,상층부001,아랫부분001,앞부분001,윗부분001,일부분001,하반부001, 포함하여 총 16개	가랑이002,가슴팍001,거죽001,겨드랑이002,공감대001,국부001,깃007,꼬리001,날009,내측001,노른자001,노른자002,단003,단자008,달걀노른자001,대목002,대목003,동강001,동강이001,동부007,동부010, 포함하여 총 296개	
성질002	가변성001,고성004,공성001,구상성001,난연성001,내습성001,내열성001,늘림성001,다산성002,면역성001,무기질001,방부성001,부동성001,불규칙성001, 포함하여 총 54개	가능성001,가동성001,가연성001,가용성001,간결성001,개연성001,건강성001,건성004,건전성001,견고성001,경질003,계속성001,공공성001,공익성001,공통성001,관념성001,관련성001,관성002,교훈성001,구심성002,구질004,구체성001,극성004, 포함하여 총 229개	

〈표 62〉 관련정도 이용 어휘 등급화 결과(하위어)

3) 관련 정도를 이용한 어휘별 등급화 활용 방안

관련 정도의 기준에 따라 어휘를 등급화한 결과는『우리말샘』누리집에서 어휘지도를 보여줄 때 활용할 수 있다. 지금은 모든 관련어가 동일한 가중치로 표시된다. 이 정보를 활용하여 어휘지도를 수정하면 보다 명확한 어휘지도를 제공할 수 있다. 예를 들어, 관련성이 높은 어휘는 연결선을 진하게, 관련성이 보통인 것은 중간 정도로, 관련성이 낮은 어휘는 연하게 할 수 있다. 다른 방법으로 관련성이 높은 어휘와 보통인 어휘로 구분하여 보여줄 수도 있다.

또 다른 활용 방안에는 인공지능을 이용한 유의어/관련어/연관어 자동 검색 결과의 기준이나 척도로 사용할 수 있다. 예를 들어 <표 59>의 '거년'이란 어휘를 가지고 A라는 유의어 자동 추출 시스템이 대용량 말뭉치에서 찾은 유의어 검색 결과와 이 사업의 등급화 결과를 비교 분석할 수 있다. 관련성이 높은 '안해'가 관련성이 보통인 '객세, 고세, 과년, 석년, 지난해'보다 더 유사성이 높다고 나온다면 A라는 시스템의 유의어 자동추출 결과는 신뢰할 만하다고 볼 수 있다.

어휘 분석 단계에서 두 어휘가 서로 상위어 관계이거나 하위어 관계인 경우가 발견되었고, 둘 중 하나는 수정이 필요한데 관련 정도를 가지고 그 문제를 해결하려고 <표 63>과 같이 어휘쌍 3가지에 대해 설문 응답 결과를 비교하여 보았다. 3가지 어휘쌍 모두 평균, 표준편차, 평균/표준편차 각각에 대해 큰 차이를 보이지 않았다. 따라서 이들 쌍은 장기적인 안목으로 자료를 수정 보완할 필요가 있다.

구분	표제어	관련어	관련성 등급	평균(5점 척도)	표준편차	mean/SD
상위어	가식001	거짓001	중	3.38	1.19	2.85
	거짓001	가식001	중	3.45	1.21	2.86
상위어	가입001	참가001	중	3.41	1.23	2.78
	참가001	가입001	중	3.11	1.35	2.31
하위어	수월005	누월001	중	3.42	1.25	2.74
	누월001	수월005	중	3.45	1.26	2.74

〈표 63〉『우리말샘』의 데이터 정제가 요구되는 어휘쌍의 설문 결과

4) 생각해 볼 만한 분석 결과

관련 정도를 이용한 어휘별 등급화를 위한 기초 자료 검토 과정에서 나온 여러 가지 분석 결과 중 생각해 볼 만한 것을 정리해 보았다.

첫째, '평균'과 '평균/표준편차' 2개의 척도 중 어느 것을 사용하는 것이 나을지에 대한 것이다. 5점 척도에서 단순한 평균을 사용하는 것보다 'mean/SD'를 사용하는 것이 보다 의미 있는 결과를 보여준다고 알려져 있다. <표 64>에 2가지 척도를 가지고 비교한 관련어별 상위 5개 어휘쌍을 보면 상위에 위치한 어휘쌍 대부분 동일하다는 것을 볼 수 있다. 이 결과로 설문자들이 잘 아는 단어에 대해서는 2가지 척도가 비슷한 결과를 보여준다는 것을 알 수 있다.

구분	어휘쌍	평균	어휘쌍	mean/SD
비슷한말	어머니-엄마	4.39	어머니-엄마	5.81

	사이버대-사이버대학교	4.38	사이버대-사이버대학교	5.28
	기말시험-기말고사	4.42	기말시험-기말고사	5.21
	물안경-수경	4.29	물안경-수경	5.21
	추천서-추천장	4.23	추천서-추천장	5.21
반대말	낙선하다-당선하다	4.39	낙선하다-당선하다	5.34
	개회하다-폐회하다	4.32	개회하다-폐회하다	5.11
	작다-크다	4.35	작다-크다	5.07
	짧다-길다	4.32	짧다-길다	5.07
	투명하다-불투명하다	4.28	투명하다-불투명하다	5.01
상위어	요골-뼈	4.45	요골-뼈	5.62
	셰퍼드-개	4.49	셰퍼드-개	5.41
	새송이버섯-버섯	4.5	새송이버섯-버섯	5.38
	고기압-기압	4.38	고기압-기압	4.96
	수묵화-그림	4.42	수묵화-그림	4.94
하위어	그림-상상도	4.29	그림-상상도	5.17
	곤충류-매미	4.31	곤충류-매미	5.05
	혈관-뇌혈관	4.28	혈관-뇌혈관	4.96
	그림-풍경화	4.29	그림-풍경화	4.92
	세금-소득세	4.28	세금-소득세	4.92

〈표 64〉 관련어별 상위 5개 어휘쌍(척도 수치가 높은 것)

둘째, 설문자들이 하나의 어휘가 다양한 의미를 가진다는 것을 모르고 자신이 알고 있는 의미로 간주하고 답을 한 경우가 많다는 것이다. <표 65>는 '평균/표준편차'가 낮은 순으로 정리한 것 중 어휘 난이도가 낮은 것을 선별한 것이다. 여기서 어휘쌍은 실제 설문자에게 제시한 단어 형태이고, 표제어와 관련어는 『우리말샘』 의미별 표기를 간략히 한 것이다. 설명란에는 이 어휘쌍에 대해 설문자들의 응답 결과가 관련이 낮다고 나온 이유를 나름대로 설명해 본 것이다. 비슷한말인 경우 대부분 제시한 단어의 원래 의미를 자기가 아는 의미의 단어로 오해하였기 때문에 관련성이 낮다고 표시한 것이라 생각된다. 반대말의 경우는 정확한 의미를 몰랐거나 다른 단어로 착각한 듯하다. 상위어/하위어의 경우도 대부분 제시한 단어의 원래 의미를 자기가 아는 의미로 오해한 결과로 보인다. 하지만 왜 관련성이 낮다고 답했는지 의문이 가는 어휘쌍도 종종 보인다. 전문가의 자문 등을 통해서 검토할 필요가 있는 대목이다.

다양한 의미를 가진 단어의 경우 설문자에게 정확한 의미를 제시하는 방법에 대한 모색이 필요하고,

만약 그러한 방법을 찾아 제시하였음에도 불구하고 설문자가 자신이 알고 있는 의미로 평가한 결과를 확인할 수 있는 방법론을 수립할 필요가 있다.

구분	어휘쌍	mean/SD	표제어	관련어	설명
비슷한말	뚱보-뚱딴지	1.81	뚱보001	뚱딴지003	여기서 '뚱보'는 "심술 난 것처럼 뚱해서 붙임성이 적은 사람"을 의미하는데 "뚱뚱한 사람"이라 오해한 결과
	고기-물고기	1.85	고기003	물고기001	여기서 '고기'는 "물에서 사는 지느러미와 아가미가 있는 척추동물을 통틀어 이르는 말"을 의미하는데 가장 일반적인 의미인 "식용하는 온갖 동물의 살"로 오해한 결과
	하나님-아버지	1.89	하나님001	아버지008	종교 용어를 일반 단어로 오해한 것
	임금-왕자	1.91	임금001	왕자003	우리가 흔히 아는 '왕자'로 오해한 것
	병원-군병력	1.92	병원001	군병력001	우리가 흔히 아는 '병원'으로 오해한 것
반대말	금방-방금	2.13	금방002	방금004	여기서 '금방'은 "말하고 있는 때로부터 얼마 후에"를 의미하는데 "말하고 있는 시점보다 바로 조금 전에"라는 의미를 가진 '금방'으로 오해한 결과
	법인-자연인	2.21	법인002	자연인002	법률 용어를 일반 단어로 오해할 결과
	세계어-자연어	2.27	세계어001	자연어001	정확한 의미를 몰라서?
	덩어리-덩이	2.29	덩어리001	덩이001	정확한 의미를 몰라서?
	박수-무녀	2.3	박수001	무녀001	정확한 의미를 몰라서?
상위어	할복-땀	1.62	할복002	땀001	우리말샘 오류로 보인다.
	고층-다수	1.65	고층002	다수001	여기서 '고층'은 "건물의 층수가 많은 것"을 의미하지만 "여러 층으로 된 것의 높은 층"으로 오해한 결과
	돼지-사람	1.65	돼지003	사람001	사람을 놀릴 때 사용하는 '돼지'라는 표현을 일반적인 '사람'으로 오해한 결과
	곰-사람	1.65	곰004	사람001	오해

	개-사람	1.67	개005	사람003	오해
하위어	개구리-하마	1.68	개구리001	하마006	여기서 '하마'는 "청개구릿과의 하나"를 의미하는데 일반적인 하마로 오해한 결과
	곰-곰탕	1.76	곰001	곰탕002	여기서 '곰'은 "고기나 생선을 진한 국물이 나오도록 푹 삶은 국"을 의미하는데 일반적인 '곰'으로 오해한 결과
	남편-인형	1.78	남편001	인형008	여기서 '인형'은 "손위 누이의 남편을 이르거나 부르는 말"을 의미하는데 일반적인 '인형'으로 오해한 결과
	일-정조	1.78	일001	정조016	정조를 지키는 것이 일이 아니라고 본다?
	소-염소	1.79	소004	염소001	전문용어

〈표 65〉 관련어별 하위 5개 어휘쌍(척도 수치가 낮은 것)

셋째, 관련 정도에 따른 어휘 등급화 결과를 어휘 난이도를 이용하여 비교하여 보면 설문자들이 자기가 아는 어휘에 대해서는 자신 있게 답을 하지만 그렇지 않은 경우엔 '보통(중)'에 답했다는 것을 알 수 있다. <표 66>은 비슷한말에 대해서 표제어와 관련어의 어휘 난이도별 관련성 등급의 비율을 보여준다. 표제어나 관련어 중 하나라도 난이도 '상'인 어휘가 없으면 '관련성_상'의 비율이 70% 부근이다.

㈜낱말의 어휘 난이도 기준에 따르면, 난이도 하와 중은 일반인이 대부분 알고 있는 어휘 수준이다. 일반인을 대상으로 어휘의 관련성 정도를 파악하기 위해서는 가능한 한 일반인이 알 수 있는 어휘를 가지고 해야 기대하는 결과를 얻을 수 있다고 유추할 수 있다. 설문에 응하는 사람들의 입장에서는, 자신이 모르는 단어에 대해서는 『우리말샘』의 내용이 올바를 것이라 생각하면서도 그러한 확신이 없으므로 5점 척도 중 3이나 4를 선택했을 가능성이 매우 높다.

표제어난이도	관련어난이도	관련성_상	관련성_중	관련성_하
상	상	67.72%	32.28%	0.00%
상	중	46.94%	53.06%	0.00%
상	하	38.88%	61.12%	0.00%
중	상	46.58%	53.42%	0.00%
중	중	69.65%	30.35%	0.00%
중	하	69.59%	30.41%	0.00%

하	상	35.26%	64.74%	0.00%
하	중	69.43%	30.57%	0.00%
하	하	73.57%	26.11%	0.32%

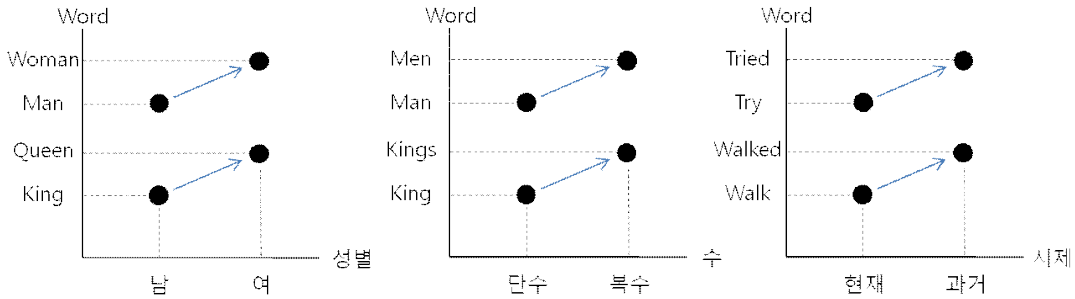
〈표 66〉 비슷한말의 어휘 난이도별 관련성 정도 비교

7. 추가 제안 설문 및 분석

가. 추가 제안 개요

본 컨소시엄에서는 과제 목표 달성 완료에 더하여 개별 어휘가 가진 함의 분석을 위한 기초적인 분석을 추가적으로 제안하였다. 언어가 가진 '고유한 감성적 차원'과 '의미적 차이점(semantic differential)'에 대한 분석은 방대한 연구와 자원을 필요로 하는 작업이므로, 본 컨소시엄에서는 추가 분석의 착수를 위해 분석 대상과 범위를 설정하기 위한 논의를 진행하였다.

현재 인공지능을 통한 텍스트 학습 및 분석의 큰 줄기를 이루고 있는 것은 분산 표현, 즉, word2Vec과 같이 word Embedding을 통한 벡터화(Vector Presentation)이다.



〈그림 24〉 word Embedding의 구현 예

그간 이루어진 인공지능 언어연구의 성과는 눈부신 성장을 거듭하였고, 이제 누구나 간단한 프로그래밍 지식만을 가지고도 word Embedding을 구현할 수 있는 수준으로까지 일반화되었다고 해도 과언이 아니다. 다만, 방대한 문건의 학습을 통해 이루어지는 방식은 아직 인간의 언어표현 상황에 반영된 감성적인 부분에 대한 연구에 많은 숙제를 남기고 있다. 이에 대해 화자의 감성분석 등을 위한 많은 솔루션들이 연구되고 있고 그 성과가 상용화되고 있다.

본 컨소시엄은 사업 과제의 진행을 통해 우리말의 품사 중 명사와 동사, 특히 명사의 수가 다른 품사에 비해 압도적으로 많은 것을 발견하였다. 본 공정의 대상인 어휘쌍의 경우 동사와 명사의 비중은 전체 어휘쌍의 98.8%에 이른다. 이중 명사만의 비중은 94%인데, 정도의 차이는 있겠지만 다른 언어에서도 이러한 상황은 유사할 것으로 추정되어 1~2%에 지나지 않는 형용사와 부사만으로 명사와 동사를

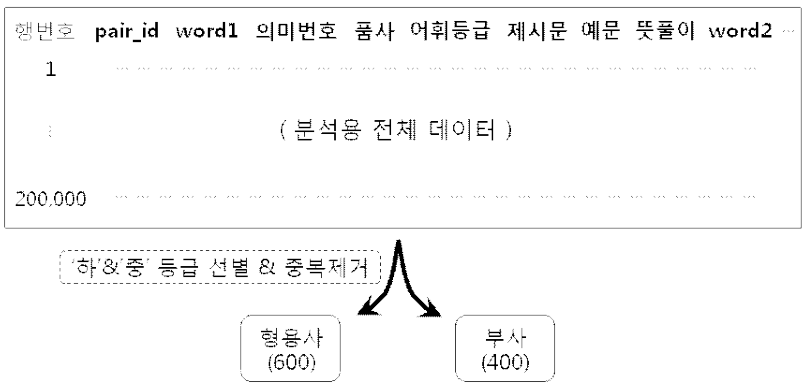
부연 설명할 수 있는 한국어의 특징을 파악하고자 하였다. 즉 형용사와 부사 어휘 내면에 존재하여 언어를 구현하는 시점에 암묵적으로 전달되는 잠재적 의미에 대한 고찰이 한국형 인공지능 언어분석 연구에 중요한 역할을 담당할 것으로 판단된다.

따라서 본 컨소시엄은 본 과제에서 설문 대상으로 삼았던 어휘쌍에 포함된 형용사와 부사 중 최대한 낮은 등급, 즉 쉬운 단어를 선별하여 잠재적인 의미의 차원을 탐구해 보는 것으로 범위를 결정하였고 다음과 같이 설문을 설계하였다.

나. 추가 제안용 설문 설계

1) 설문 대상 어휘 선정

1~5주 차 설문용으로 선발된 어휘쌍에 포함된 어휘들 중 '하' 또는 '중' 등급의 어휘를 추출하여 중복 제거 작업 후 형용사 600개와 부사 400개를 선정하였다. 본 공정에 사용된 부사의 비중은 형용사의 75%이므로 타당한 배분이라고 판단하였다.



〈그림 25〉 설문 대상 어휘 선정 과정

2) 설문용 의미 차원 선정

본 공정에서 진행된 설문과 다르게, 추가 제안을 위한 설문에서는 어휘쌍이 아닌 개별 어휘를 대상으로 한국어 사용자가 갖게 되는 정서적 느낌 또는 의미의 강도를 추출하는 것을 목표로 하였다. 따라서 내면적 의미 강도를 수치화하여 설문하는 방식이며, Harvard VI-4 분류체계의 기본 모델이었던 Osgood(1952)의 방식을 면밀히 검토하였다. 실제로 Osgood의 모델에서 'rounded－angular', 'rough－smooth', 'cold-hot'과 같은 구분은 초기 모델에서만 존재하고 Harvard VI-4 분류체계에는 포함되어 있지 않다. 그러나 본 컨소시엄은 이러한 차원이 한국어에서는 '격식-비격식'과 같은 차원으로 포착될 수 있을 것으로 판단하였다.

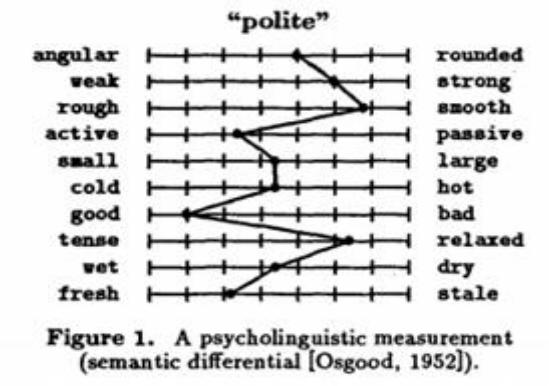


Figure 1. A psycholinguistic measurement (semantic differential [Osgood, 1952]).

〈그림 26〉 Osgood(1952) 모델

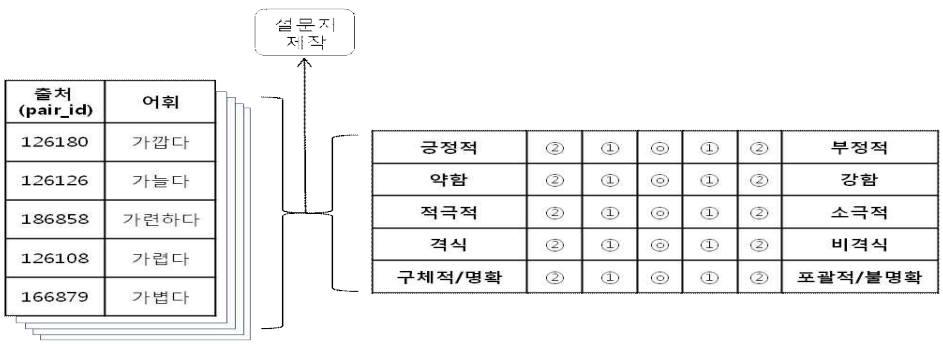
이를 위해 의미 차원 항목과 강도에 대한 응답 척도를 다음과 같이 결정하였다. 본 설문의 차원은 Osgood 모델에서 더 나아간 구체적/명확성 부분을 추가하였는데, 이는 어휘가 가지는 구체성의 정도에 따라서 어휘 관계의 양상이 다르게 나타날 수 있기 때문이다.

척도	의미 강도					척도
긍정적	②	①	⓪	①	②	부정적
약함	②	①	⓪	①	②	강함
적극적	②	①	⓪	①	②	소극적
격식	②	①	⓪	①	②	비격식
구체적/명확	②	①	⓪	①	②	포괄적/불명확

〈표 67〉 추가 제안 설문용으로 설정된 의미 차원

다. 설문지 제작 및 설문 패널 모집

추가 제안을 위한 설문은, 선정된 1,000개의 형용사와 부사 어휘 각각에 대해 5개 차원에 대한 질문을 실시하여 패널당 총 5,000개의 응답이 수집되도록 설계되었다.



〈그림 27〉 추가 제안용 설문 설계의 개념

추가 제안의 설문지는 본 공정과 마찬가지로 웹 또는 모바일 웹에서 답변 작성이 가능한 형태로 제작되었다. 이에 설문 패널 30명을 모집하였고 2020.02.06.~02.10의 기간 동안 설문을 진행하였다. 모집된 패널의 분포는 다음과 같다.

전체		고등학교		대학(4년제)		대학원(석사)		대학원(박사)		합계	
		N	%	N	%	N	%	N	%	N	%
		2	6.7	23	76.7	4	13.3	1	3.3	30	100
성별	남	0	0	10	83.3	1	8.3	1	8.3	12	100
	여	2	11.1	13	72.2	3	16.7	0	0	18	100
연령	~29	2	16.7	10	83.3	0	0	0	0	12	100
	30대	0	0	4	100	0	0	0	0	4	100
	40대	0	0	5	62.5	2	25.0	12.5	12.5	8	100
	50~	0	0	4	66.7	2	33.3	0	0	6	100

〈표 68〉 추가 제안 설문 패널 분포

라. 설문 분석

추가 제안 설문 분석의 목적은 ①어휘가 고유의 감성-극성의 차원(sentimental polarity)이 임의의 어휘쌍에 어떻게 부가적으로 적용될 수 있을지를 검토하고, ②각 차원별로 가장 극단적인 감성의 어휘 분석을 통하여, 본 분석 결과의 감성 분석 적용 가능성을 조망하고, ③특정 집단별로 어휘 수용이 차이가 있을 수 있음을 검토하고, ④마지막으로 본 극성 차원 차원에 따른 어휘별 군집화(클러스터링) 가능성을 모색하는 것이다.

본 설문은 어디까지나 의미 번호가 식별되지 않은 1,000개의 제한된 어휘에 대한 30명의 집단에 대한 응답을 대상으로 한 것이다. 본 설문 분석으로 유의미한 패턴과 유효한 시사점을 발견할 수 있을 경우, 설문을 확장하여 『우리말샘』 사전에 적용할 수 있을 것으로 기대한다.

1) 반대말 어휘쌍에 대한 적용

‘멀다-가깝다’, ‘가늘다-굵다’의 반대말 어휘쌍에 대한 분석을 실시하였다. 이들 단어는 모두 다의어를 가진 복수의 어휘 번호를 가지고 있다.

예를 들어 ‘멀다003’은 ‘가깝다001’과 반대말 관계이며, ‘멀다005’는 ‘가깝다002’와 반대말 관계이다. ‘멀다-가깝다’에 대한 어휘쌍은 총 4개로 각각의 어휘쌍에 각 어휘 번호에 고유한 예문을 제시하고 사용자 응답을 받았으며, 베타계수는 4.53 ~ 4.69로 모두 강한 관련성 관계이다. 엔트로피 정보 역시 1.15 ~ 1.17로 비슷한 수치이며, 비슷한 정보량을 가지고 있다. 수치상 눈에 띄는 차이는 응답 시간으로 ‘멀다005-가깝다002’의 응답 시간이 다른 어휘쌍에 비해 상대적으로 짧았다는 점이다. 참고로 이렇게 응답

자가 가장 빠르게 반응했던 '멀다005'는 "서로의 사이가 서먹서먹한 것"을 의미하여, '가깝다002'는 "서로의 사이가 다정하고 친한 것"을 의미한다.

베타계수가 가장 낮은(즉, 관련성이 적다고 인지한) 어휘쌍인 '가깝다005'는 "시간적으로 오래지 않다"는 의미이고 '멀다006'은 "시간적으로 사이가 길거나 오래"라는 의미이다. '다정함-서먹서먹함'과 같은 뉘앙스의 어휘는 더 친숙하게 이해하고, 시간에 대한 은유와 같은 어휘는 더 멀게 이해하는 것이다.

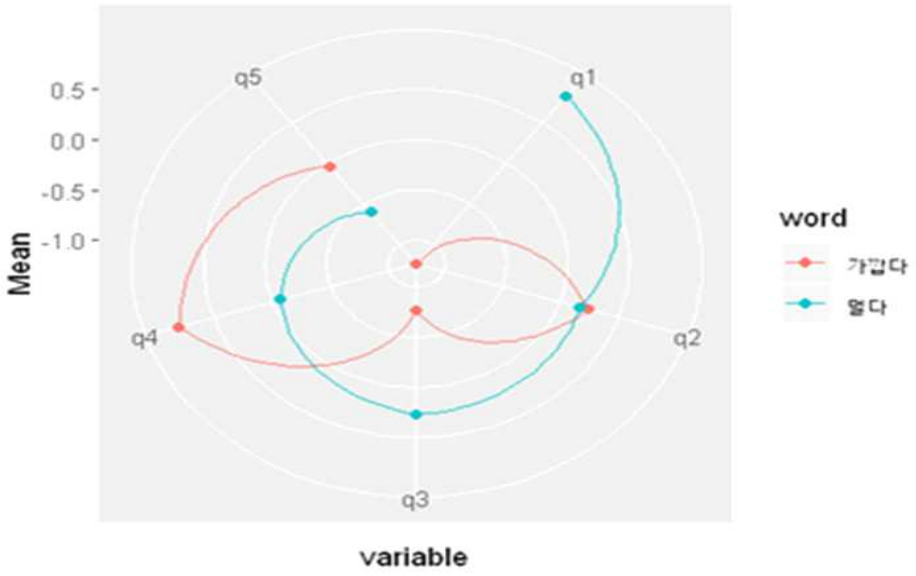
full_id	relation	head.의미어	tail.의미어	Shannon_entropy	beta	grade	Quantile_grade	mean.time
2010740 [201]_126179 : 멀다-가깝다	반대말	멀다006	가깝다005	1.174970	4.532351	1	강	2.689320
2020737 [202]_126177 : 멀다-가깝다	반대말	멀다003	가깝다001	1.154329	4.676162	1	강	2.569307
2020738 [202]_126180 : 멀다-가깝다	반대말	멀다007	가깝다006	1.149205	4.565998	1	강	2.435644
2030736 [203]_126178 : 멀다-가깝다	반대말	멀다005	가깝다002	1.154259	4.690445	1	강	1.679803

〈표 69〉 '멀다-가깝다' 어휘쌍에 대한 본 설문 응답

'멀다'와 '가깝다'라는 어휘가 의미 번호가 식별되지 않은 상태에서 직관적으로 사용자가 가진 감성의 차원은 아래 그림의 상위부분과 같다.

여기서 q1은 '긍정(-2)~부정(+2)', q2는 '약함(-2)~강함(+2)', q3는 '적극(-2)~소극(+2)', q4는 '격식(-2)~비격식(+2)', q5는 '명확(-2)~불명확(+2)'의 차원을 의미한다. 아래 그림에서 스파이더 차트의 안쪽은 각각 긍정, 약함, 적극, 격식, 명확성을 의미한다.

'가깝다'는 극단적인 긍정, '멀다'는 극단적인 부정, 약함과 강함은 두 어휘 간에 차이가 없으며, '가깝다'는 적극, '멀다'는 소극, '가깝다'는 조금 더 강한 비격식을, '가깝다'와 '멀다'는 모두 포괄적이나, '멀다'가 더욱 불명확하다.



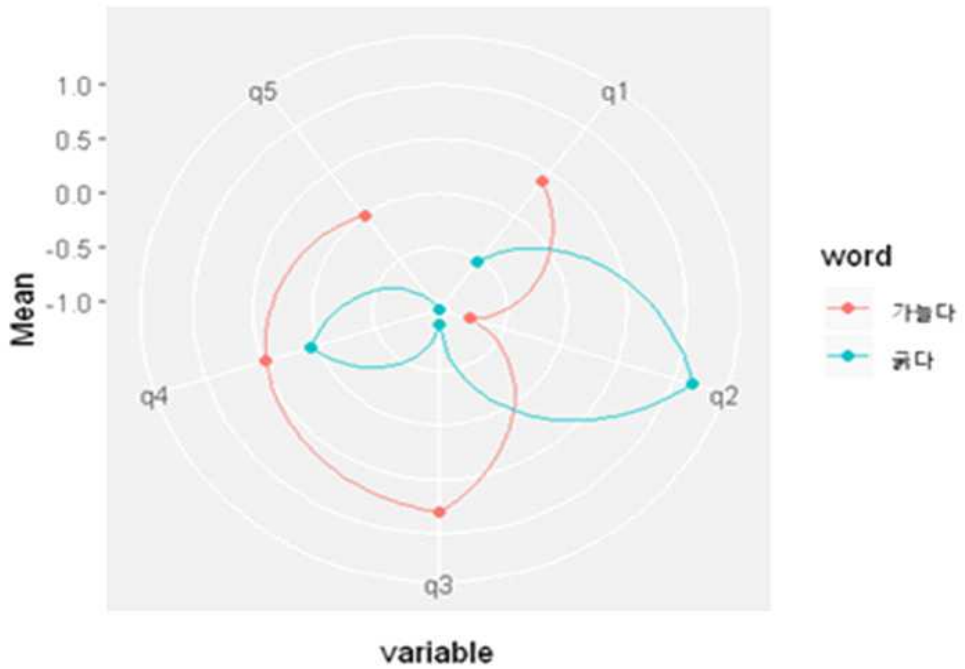
〈그림 28〉 '가깝다-멀다' 어휘쌍에 대한 추가 설문 응답

한편, '가늘다'와 '굵다'는 3가지 종류의 어휘쌍으로 평가되었는데, '가늘다001'이 "물체의 지름이 짧은 것", '가늘다002'는 "소리의 울림이 약한 것", '가늘다006'은 "사이가 좁고 촘촘한 것"을 의미한다. 상대적으로 형용사에 대한 관련성을 높게 평가한 만큼(4.2가 높은 관련성에 대한 기준 점), 처음의 어휘를 제외하고는 모두 관련성에 대한 평가 등급은 상대적으로 높은 베타계수에도 불구하고 '중'으로 평가되었다.

full_id	relation	head.의미어	tail.의미어	Shannon_entropy	beta	grade	Quantile_grade	mean.time
2010721 [201]_126101 : 가늘다-굵다	반대말	가늘다001	굵다001	1.186590	4.541952	1	강	2.815534
2020717 [202]_126102 : 가늘다-굵다	반대말	가늘다002	굵다006	1.208887	4.166229	1	중	2.460396
2030724 [203]_126126 : 굵다-가늘다	반대말	굵다008	가늘다006	1.228511	4.047619	1	중	2.640394

〈표 70〉 '가늘다-굵다' 어휘쌍에 대한 본 설문 응답

아래 그림은 '가늘다-굵다'에 대한 비교이다. 두 어휘의 비교에서 가장 극단적인 차이를 보이는 것은 q2와 q3으로 '가늘다'는 약하고 소극적이며, '굵다'는 강하고 적극적인 감성의 단어로 응답하였다.



〈그림 29〉 '가늘다-굵다' 어휘쌍에 대한 추가 설문 응답

본 분석은 해당 형용사에 대한 전반적인 느낌에 대한 응답자의 주관적인 느낌을 평가한 것이다. 이러한 어휘에 대한 사용자의 감성 평가는 어휘에 대한 잘못된 사용(예를 들어 '가늘다' 대신 '얇다', '굵다' 대신 '두껍다'를 사용하는 것)에 극성 차원의 해석이 가능할 수도 있을 것이다. 또한 대체어로 이와 비

슷한 속성 값을 가지는 단어로 대체할 수도 있을 것이다.

2) 가장 극단적인 감성의 어휘

5개의 극정 차원별로 가장 극단적인 어휘를 확인한 결과는 다음과 같다.
차원별로는 '긍정-부정'의 극단적 평가에 대한 범위가 가장 컸다('건강하다, 기쁘다'-1.80, '게을러터지다'-1.87). 눈에 띄는 점은 '게을러터지다'가 다른 어휘보다 더 극단적으로 부정적인 점수를 받았다는 사실이다. 또한 긍정에 대한 평가에서는 '공정하다', '공명정대하다'와 같은 어휘가 높은 긍정의 점수를 받았다는 점이다. 이 부분은 본 설문에 20대 응답자가 가장 많이 분포하였다는 사실로 미루어 20대에 '공정성'에 대한 젊은 층의 인지를 가늠해 볼 수 있는 대목이라 하겠다. 실제로 연령별, 성별로 어휘 수용에 다른 경향을 확인할 수 있었다.

〈 긍정										부정 〉
건강하다	기쁘다	공정하다	공명정대하다	귀하다	--	구질구질하다	사악하다	게걸스럽다	극악하다	게을러터지다
-1.80	-1.80	-1.76	-1.60	-1.60		1.70	1.73	1.73	1.73	1.87

〈 약함										강함 〉
약하다	연약하다	유약하다	가련하다	미약하다	--	기필코	과감하다	기운차다	극악하다	강건하다
-1.37	-1.30	-1.16	-1.13	-1.1		1.53	1.57	1.60	1.60	1.63

〈 적극										소극 〉
기운차다	공명정대하다	강직하다	기필코	과감하다	--	약하다	나른하다	데면데면히	의기소침하다	느릿느릿하다
-1.63	-1.46	-1.40	-1.40	-1.37		1.00	1.00	1.03	1.10	1.10

〈 격식										비격식〉
공명정대하다	공정하다	겸허하다	겸손하다	청렴하다	--	구질구질하다	걸신스럽다	구리다	게을러터지다	게걸스럽다
-1.57	-1.43	-1.30	-1.23	-1.23		1.36	1.40	1.50	1.57	1.60

〈 명확										불명확〉
까맣다	매월	배고프다	없다	매일매일	--	정성드릿이	가없이	여차하다	소연하다	누릇하다
-1.30	-1.27	-1.27	-1.27	-1.23		0.87	0.87	0.90	0.93	0.93

〈표 71〉 극단적 감정 어휘 결과

이와 같은 어휘에 대한 감성 평가의 정량화는 다양한 감성 분석에 사용될 수 있다. 예를 들어 '구질구질하다'와 '게을러터졌다'라는 평가에는 각각 1.70, 1.87의 부정 점수를 부여하여 분석할 수 있다. '공정하다'와 '귀하다'라는 평가에는 각각 1.76, 1.60의 긍정 점수를 부여할 수 있다.

3) 집단간 인지 어휘 수용의 차이

본 설문을 통해 연령별로 다소 상이한 어휘 수용 패턴을 파악할 수 있었다. 예를 들어 젊은 남성층은 다른 집단보다 '공정하다', '공명정대하다'에 대해 더 적극적이고 긍정적인 어휘로 인식하고 있으며, 여성층은 '온순하다'에 대해 더 소극적이고 약한 어휘로 인식하고 있었다. 성별 집단 간의 테스트 결과 '긍/부정'과 '명확/불명확'의 차원에는 차이가 없었으나, 다른 차원은 모두 유의미한 차이점이 발견되었다(대부분 여성의 단어에 대한 더 극단적인 인지에 의한 것이었다. 즉, 적극적이 단어는 더 적극적으로, 소극적인 단어는 더 소극적으로 인지하는 경향이 있었다). 그러나 이 같은 그룹별 분석은 30개라는 제한된 샘플로 섣불리 결론을 내리기 어려운 부분이다. 아래는 t 테스트로 성별 어휘에 대한 분석을 하고, 가장 높은 t 수치를 보여주는 상위 5개 단어를 제시한 것이다. 예를 들어 '허영다'라는 단어에 대해 남성은 강함의 0.33을, 여성은 약함의 −0.67의 감성으로 인지하였다. '미숙하다'라는 단어에 대해 모두 약한 감성의 단어로 인지하였으나, 여성이 더욱 극단적인 약한 감성의 단어로 인지하였다(significant, $p < 0.05$).

약함 - 강함

idx	target_word	t.stat	t.p	t.mx	t.my
548	허영다	3.511363	0.001542171	0.33333333	-0.6666667
209	미숙하다	3.510399	0.003472261	-0.16666667	-1.2777778
421	의아스럽다	3.442627	0.003533959	0.08333333	-0.7777778
600	간데온데없다	3.435693	0.002235811	0.83333333	-0.2222222
199	무소용하다	3.395202	0.002855403	0.00000000	-1.0000000

적극 - 소극

idx	target_word	t.stat	t.p	t.mx	t.my
279	빵빵하다	3.538753	0.001756021	-0.08333333	-1.1111111
392	우럼자다	3.477197	0.002788592	-0.75000000	-1.6111111
277	빠르다	3.286388	0.002944532	-0.41666667	-1.2777778
224	밝다	2.710057	0.013613690	-0.66666667	-1.3888889
606	가득가득히	2.611063	0.017383844	-0.58333333	-1.2777778

〈표 72〉 어휘 수용 차이가 높은 어휘(약함-강함, 적극-소극 차원)

4) 주성분 분석을 통한 어휘별 군집화(클러스터링)

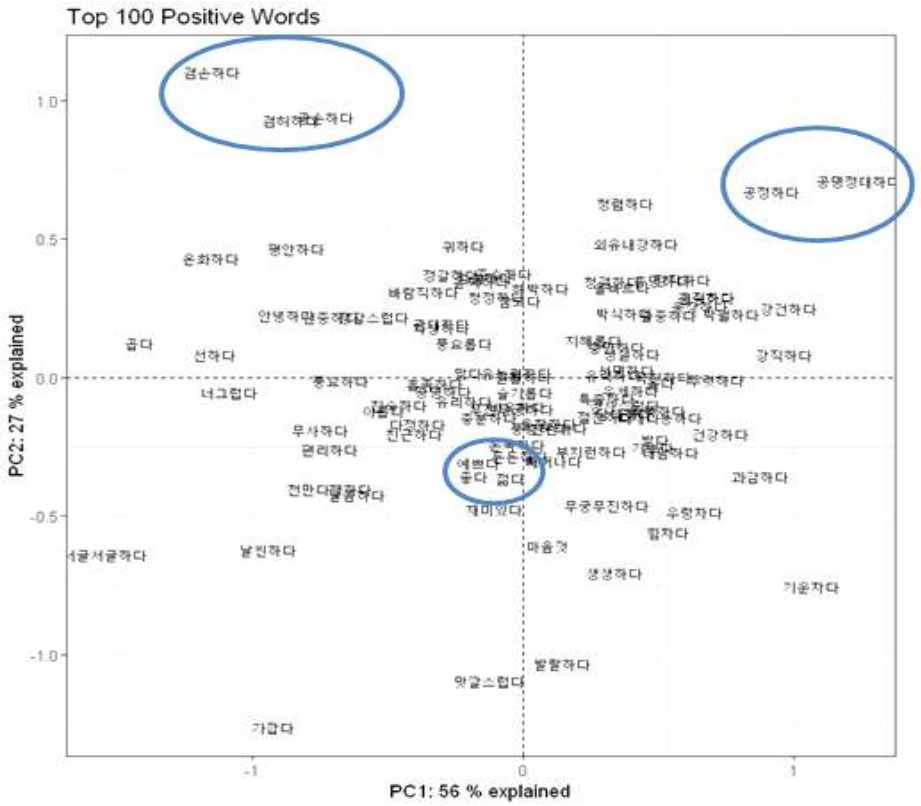
어휘별 군집화(클러스터링) 가능성을 조망하기 위하여, 5개의 차원별 속성값이 파악된 이들 전체 1,000개 어휘들을 대상으로 주성분 분석(principal components analysis)을 수행하였다. 주성분 분석은 각 차원의 어휘 벡터들(5차원의 벡터)을 임의의 어떤 벡터에 투영(projection)시켜 전체 1,000개의 어휘가 해당 벡터에 투영된 지점까지의 거리인 내적의 제곱의 합이 가장 큰 고유 벡터(eigen vector)를 찾는 것이다. 이 벡터를 제1의 주성분(PC1)이라고 부르며, 해당 벡터에 투영된 각 어휘의 적재값(loading variable, scalar)을 구할 수 있다. 마찬가지의 방법으로 제1 주성분이 고정된 상태에서 다음으로 내적의

합이 가장 큰 벡터를 제2의 주성분(PC2)이라고 부르며 임의의 N개의 주성분이 존재할 수 있다.

주성분의 개념은 직관적으로 설명하면 약 A% 비율의 긍정-부정 차원, B% 비율의 약함-강함 차원, C% 비율의 적극-소극차원 D% 비율의 격식-비격식 차원, E% 비율의 명확-불명확 차원을 가지는 임의의 성분으로 설명할 수 있다. 이 같은 주성분을 설명하는 고유 벡터를 찾는 방법은 최대 우도법 등이 있다. 또한 각 주성분이 가지는 분산을 통하여 주성분의 설명력을 설명할 수가 있는데, PC1의 설명력이 가장 높으며 다음으로 PC2, PC3, ...PC N순이다.

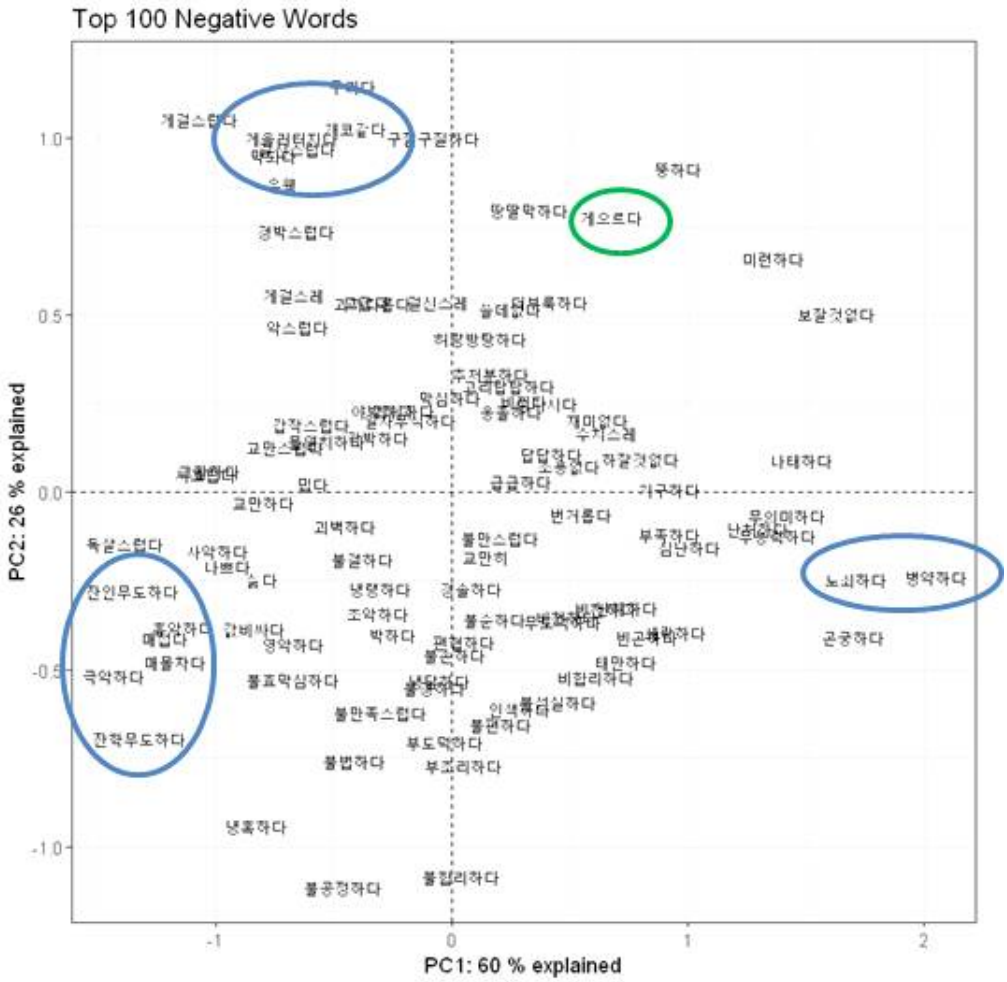
아래 그림은 긍정어 100개와 부정어 100개를 대상으로 각각 주성분 분석을 하고, 첫 번째와 두 번째 주성분 축에 해당하는 요인 적재값을 x, y 좌표로 하여 어휘를 위치시킨 것이다. 주성분 분석을 통한 차원의 축소와 시각화 방법은 정보의 손실이 발생할 수 있다는 단점이 있어 최신의 기계 학습에서는 auto-encoder나 또는 t-sne 알고리즘보다는 상대적으로 덜 선호되는 방법이다(딥러닝을 통한 월인천강지곡 시각화 방안, 김소정 등(2019)). 그러나 그럼에도 불구하고 주성분 분석은 어휘가 어떻게 군집화(클러스터링)이 될 수 있는지 풍부하고 직관적인 정보를 제공한다.

예를 들어 '공명정대하다'의 유사어는 『우리말샘』 사전에 따르면 '공명정당하다'뿐이다. 아래 그림에서는 이 단어는 '공정하다'와 비슷한 어휘로 분류될 수 있다. 마찬가지로, '겸손하다', '겸허하다' '공손하다' 등은 비슷한 감성을 가지는 대체어로 제시될 수 있을 것이다. 비슷한 예로 '예쁘다'라는 단어의 비슷한 말은 『우리말샘』 사전에는 '이쁘다'라는 어휘 하나이다. 이 단어는 '좋다', '젊다'와 비슷한 어휘로 분류될 가능성이 높다(군집화(클러스터링)을 위해서는 요인 분석 결과가 아닌, 원시 자료가 가지는 고유의 각 감성-극성 차원별 거리를 기반으로 한 알고리즘의 사용이 더욱 바람직하다).



〈그림 30〉 긍정어를 대상으로 한 주성분 분석 및 성분별 어휘 적재값(포지셔닝)

아래 그림은 부정어를 대상으로 한 주성분 분석이다. 여기서는 제1의 주요인인 PC1의 설명력이 긍정어보다는 다소 높았다. 어휘별로 유사한 단어들끼리 가까운 위치에 놓여 있는 패턴은 긍정어와 비슷하다. 예를 들어 '노쇠하다'는 '병약하다' 등으로 대체될 수 있을 것이다. 눈에 띄는 점은 '게을러터지다'와 '게으르다'가 다른 사분면에 위치해 있다는 점이다. 최소한 사용자 감성적 차원으로 두 단어의 의미는 다를 수 있음을 시사하는 부분이라 하겠다.



〈그림 31〉 부정어를 대상으로 한 주성분 분석 및 성분별 어휘 적재값(포지셔닝)

8. 결론

본 사업의 목적은 한국어 인공지능 서비스 개발과 국어사전 정보 고도화를 위해 한국어 분산 의미 모델을 개발하는 것으로서, 세밀한 의미 정보가 식별된 어휘별로 비슷한말, 반대말, 상위어, 하위어 등 어휘 간의 관련성에 대한 실제 사용 양상을 평가하고 이에 대한 통계 정보를 확보하는 것이다.

이를 위하여 『우리말샘』에 존재하는 약 33만개의 어휘쌍을 분석하여, 어휘쌍의 중복을 제외하고 사용자 양상 평가의 대상이 되는 20만 어휘쌍을 도출하였다. 제시된 어휘쌍에 대한 품질 높은 사용자 양

상 평가 결과를 수집하기 위해, 어휘 간 관련성에 대한 사용자 평가 및 수집 데이터를 공개한 캠브리지 대학의 Simlex-999 등 분산 의미 모델 수립 문헌을 참고하여, 응답자에 대한 문항 제시 및 질의 방법, 설문 결과에 대한 평가 방법을 설계하였다. 실제 설문 수행 전의 전문가 검토 및 1,000개 문항에 대한 200명 대상의 시험 공정(파일럿 테스트)를 수행하고 본 공정에 착수하였다.

본 공정에서는 20만 어휘쌍을 200개의 비슷한말, 반대말, 상위어, 하위어 세트로 구분하여 총 54,000명의 응답자가 평가를 하였으며, 응답자별로 1,005개의 어휘쌍에 대한 평가를 수행하도록 하였다. 최종적으로 유효한 응답 40,730건을 도출하였으며, 각 어휘쌍 별로 최소 200명의 응답을 확보하였다. 응답에 대한 평가 시간은 약 3~4초 소요되었던 것으로 분석되며, 응답 결과에 대한 본 공정과 시험 공정(파일럿 테스트)의 상관 계수 평가는 약 0.85점으로 일반적으로 국제 논문에서 통용되는 평가 점수를 상회하였다.

수집된 사용자 응답에 따라, 총 20만개의 평균과 표준편차를 활용한 베타계수, 응답의 분포의 정보량 제공을 위한 엔트로피 계수를 도출하고, 어휘 자체가 가지는 난이도에 따른 난이도 계수를 산출하였다. 이와는 별도로 각 응답에 소요되었던 시간을 산정하였다. 상호 배타적인 정보를 가지는 이들 3가지 계수 및 응답 시간은 수준 높고 다양한 인공 지능 서비스 개발 및 평가에 활용될 수 있을 것으로 기대한다. 그리고 일반인에게는 이해에 어려움이 따를 수 있으므로 직관적인 이해를 도울 수 있도록 베타계수를 이용하여 어휘 관계의 등급화를 제시하였다.

한편, 본 연구와 함께 진행된 추가 공정에서는 어휘쌍이 아닌 의미 번호가 비식별된 개별 어휘를 대상으로 한국어 사용자가 받는 내면적인 느낌 또는 의미의 강도를 추출하는 것을 목표로 하였다. 이를 위해 Harvard VI-4 분류체계의 기본 모델이었던 Osgood(1952)의 방식을 검토하고 여기에 구체적/명확성 부분을 추가하였다. 이는 어휘가 가지는 구체성의 정도에 따라서 어휘 관계의 양상이 다르게 나타날 수 있기 때문이다. 1,000개의 형용사와 부사 어휘 각각에 대해 5개 차원에 대한 질문을 실시하여 총 5,000개의 응답이 수집되도록 설계하였고, 총 30명의 유효한 응답을 확보하여 분석하였다. 추가 제안 설문 분석을 통해 ①어휘가 고유의 감성-극성의 차원(sentimental polarity)이 임의의 어휘쌍에 어떻게 부가적으로 적용될 수 있을지를 검토하고, ②각 차원별로 가장 극단적인 감성의 어휘 분석을 통하여 본 분석 결과의 감성 분석 적용 가능성을 조망하고, ③특정 집단별로 어휘 수용이 차이가 있을 수 있음을 검토하고, ④마지막으로 본 극성 차원 차원에 따른 어휘별 군집화(클러스터링) 가능성을 모색하였다.

본 과제를 통해 제시된 한국어 분산 의미 모델이 제시하는 시사점은 다음과 같다.

첫째, 방대한 분량의 사용자 평가 결과는 인공지능 기계 학습으로 포착하기 어려운 어휘별 주관적인 느낌에 대한 거리 측정이 사용자에 대한 설문 평가를 통하여 가능하다는 것을 밝혔다(예: 가장 관련성이 높은 비슷한 말은, '어머니－엄마', 가장 반대된다고 느끼는 반대말은 '당선－ 낙선'이었다). 특히 본 설문에서 제시된 설문의 질문은 어휘 관계별 거리 측정을 위한 가장 합리적인 문항 도출을 위해 내용 검토, 동료 검토, 시험 공정(파일럿 테스트) 등을 통하여 분석하고 제시한 것으로, 특히 반대말에 대한 설문에 있어서도 일관성 있는 설문 응답을 수집할 수 있었다.

둘째, 어휘가 가지는 구체성의 정도에 따라 사용자 응답이 달라질 수 있음을 확인하였다. 예를 들어

비슷한말-반대말에 대한 응답의 베타계수는 상위어, 하위어에 대한 계수보다 높다. 또한, 표제어-관련어 관계에서 표제어가 상위어인 경우와 하위어인 경우의 응답에 대한 편차가 발견되었다. 이는 적은 의미를 먼저 인지하고자 하는 심리적 요인에 기여할 수 있으며, 이 같은 사용자 인지의 '방향성'이나 '편견'의 요소의 존재는 인공지능 서비스 개발에 적용될 수 있는 여지가 크다. 뿐만 아니라 20만 쌍이라는 풍부한 기초 데이터는 이와 관련된 초기 연구의 초석이 될 수 있을 것으로 기대한다.

셋째, 본 설문 결과는 어휘를 의미별로 세분화하여 사용자가 평가한다는 측면에서 세계 최초의 시도이다. 본 결과를 통해 이러한 다의어 처리에 대한 다양한 전략 수립이 가능할 것으로 기대한다(예를 들어, '가깝다-멀다'에 대한 평가는 정서적이고 친숙한 의미인 경우와, 시간 등의 은유성을 포함한 경우의 강도와 응답 시간이 다르다. 마찬가지로 '가늘다-굵다'에 대한 평가는 1차적인 의미인 경우 '강'으로, 확장된 의미인 경우 관련성이 '중'으로 평가되었다).

넷째, 추가 설문을 통해 분석된 감석-극성 분석(sentimental polarity analysis)은 어휘 관계 분석에 대한 확장과 사전 분류체계의 제시(예를 들어, 공정하다는 '적극', '긍정', '격식'으로 분류)는 물론, 한국어에 대한 감성 분석에 대해 더 다채로운 분석이 가능함을 제시하였다(예를 들어, '극악무도하다', '보다', '게을러터지다'라는 단어에 더 부정적으로 반응할 수 있다). 또한, 다양한 대체어 제시 알고리즘에 적용될 수 있으며(예를 들어, '예쁘다'는 문맥에 따라 '좋다'로 대체할 수 있다), 집단 간 어휘 수용의 차이가 있음을 시사하였다(예를 들어, '공정하다'는 20대의 젊은 세대가 가장 긍정적인 단어로 인지하는 어휘이다).

본 연구결과와 제시된 기초 자료는 향후 『우리말샘』의 사전의 고도화와 시각화에 적용될 수 있을 것으로 기대한다. 이에 더 나아가, 우리말을 활용한 다양하고 유용한 인공지능 서비스의 개발 및 관련 연구에 활용될 수 있기를 기대한다.

[부록] 어휘쌍의 군집화(동의어 세트)

[부록] 어휘쌍의 군집화(동의어 세트)

1. 개요

본 사업에서 유의어 어휘쌍은 60,000쌍을 설문 조사하였고, 설문 조사에 사용된 어휘수는 91,618개이다. 설문 조사와는 별개로 이 유의어 어휘쌍은 다양한 용도로 활용될 수 있다. 특히 검색을 포함하여 언어처리를 위해서는 같은 의미의 서로 다른 어휘를 동치화시키는 작업이 필요한데, 유의어 데이터는 중요한 도구로 활용된다.

언어처리를 위해서는 단순한 A→B 관계의 유의어쌍보다 동일한 의미의 어휘들을 군집화(grouping)하는 것이 더 효과적이다. 유의어는 역방향도 유의어로 간주할 수 있고, 유의어의 유의어도 유의어로 간주할 수 있다는 특성이 있다(관련어의 경우에는 그렇지 않은 경우가 더 많다). 예컨대 '중추절001'의 유의어쌍은 '가우일001', '추석절001', '추석001', '한가윗날001', '추석날001', '가윗날001', '가우절001', '가배절001'이 있는데, '추석001'의 유의어쌍으로는 '가배일001', '한가위001'가 있기 때문에 이 모든 어휘들은 유의어로 간주될 수 있다.

일반적으로 동의어사전은 검색된 단어의 동의어만을 제공하고, 단어 간의 유사성은 제공하지 않는다. 동일한 의미를 가진 동의어 세트를 제작하면 훨씬 더 광범위하게 활용될 수 있다(Clustering Synonym sets in English wordNet, International Conference of Information and Communication Technology(ICoICT), 2019, IEEE, Jentrisi Priyatno; Moch Arif Bijaksana 참고).

따라서 본 컨소시엄에서는 설문 조사에 사용된 유의어 어휘쌍을 기반으로 동의어 세트를 제작하였다.

2. 방법론

동의어 세트는 동일한 개념의 어휘를 군집화한 것인데, 그 대표어를 지칭하기 곤란하므로 토픽이라 명명하고 일련번호를 부여하였다. 각 개념은 '토픽_ID'로 구분된다.

A→B, B→F, F→M의 유의어 관계가 존재하면 A, B, F, M을 동의어 세트로 군집화하였다. 『우리말샘』의 어휘는 '단어+의미 번호'로 구분되어 있기 때문에 정교한 동의어 세트를 제작할 수 있다는 장점이 있다. 동형이의어라 할지라도 의미 번호를 이용하여 개념을 구분할 수 있다. 따라서 동의어 세트의 어휘 기준은 '단어+의미 번호'를 사용하였다.

본 사업에서는 어휘쌍의 설문 조사를 실시하였기 때문에 그 결과를 활용하면 좀 더 풍부한 의미관계 정보를 얻을 수 있다. 다만 설문의 특성상 어휘쌍의 유사도에 대한 결과를 토픽과 단어 사이의 유사도로 환산하기 위하여 다음과 같은 방법을 사용하였다.

가. 토픽의 평균값

하나의 '토픽_ID'에 해당되는 어휘쌍의 어휘 설문 조사 평균값을 합산하여 평균값을 계산하고(평균값의 평균값 α), 어휘 설문 조사의 표준편차를 합산하여 평균값을 계산하였다(표준편차의 평균값β). 다시 $\alpha \div \beta$을 '토픽_ID'의 값으로 간주하였다.

나. 어휘의 평균값

어휘의 평균값은 그 어휘가 속한 토픽에서의 상대적 가치를 표현해야 하므로, 특정 '토픽_ID'에 속한 어휘쌍 내의 모든 어휘들의 평균값을 합산한 평균값(γ)과 표준편차의 평균값(δ)을 구한 후 $\gamma \div \delta$을 해당 토픽 내의 어휘값으로 간주하였다.

3. 결과

앞서 언급한대로 본 사업에서 유의어 어휘쌍은 60,000쌍을 설문 조사하였고, 설문 조사에 사용된 고유한 어휘수는 91,618개이다. 동의어 세트는 37,153개가 생성되었다. 즉 91,618개의 어휘가 37,153개의 토픽으로 군집화되었다.

topic_id	cluster
217	{'못서까래001', '못서001'}
218	{'묘두와001', '막새기와002', '방초003'}
219	{'무빙워크웨이001', '이동보도001', '자동보행기001', '자동길001', '무빙워크001', '자동복도001'}
220	{'무지개다리001', '아치교001'}
221	{'문광001', '문골001', '문얼굴001', '문틀002'}
222	{'문기둥001', '문설주001', '문주002'}
223	{'문둔테001', '둔테001'}
225	{'문중방001', '중방004', '중인방001'}
226	{'선단002', '문곁설중방001'}
227	{'문선006', '선틀001'}
228	{'물림004', '물림퇴001', '퇴004', '물림간001'}

〈그림 32〉 어휘의 군집화 예시

'토픽_ID'는 각각 평균화된 값을 가지고 있고, 군집화된 어휘는 각각 '토픽_ID'에 대한 값을 가지고 있다. 따라서 '토픽_ID'의 값과 어휘의 값의 차이가 개념에 대한 어휘의 거리라고 간주할 수 있다. 이런 계층화된 정보를 표현하기 위하여 JSON 형식으로 데이터를 생성하였다.

JSON(제이슨)은 JavaScript Object Notation의 약어로 사람이 읽고 쓰기에도 용이하고, 기계가 분석하고 생성하기에도 용이하여 데이터 교환 형식으로 많이 사용되고 있다.

생성된 동의어 세트를 표현하기 위하여 '토픽_ID'(topic_ID)와 토픽값(topic_value) 하위의 'children'이라는 Key Name에 어휘(word), 어휘값(word_value), 거리(distance)를 배열하였다.


```
{
    "topic_id" : "1",
    "topic_value" : "2.78847",
    "children" : "[{'word':'가톨릭002','word_value':2.70589', 'distance':'0.0825
    8'},{'word':'서학002','word_value':2.53038', 'distance':'0.25809'},{'word':'가
    톨릭-교001','word_value':3.02494', 'distance':'0.23647'},{'word':'공교002','word
    _value':3.08977', 'distance':'0.30130'},{'word':'천주-학001','word_value':3.101
    63', 'distance':'0.31316'},{'word':'진교002','word_value':2.88341', 'distanc
    e':'0.09494'},{'word':'천주-교001','word_value':2.67946', 'distance':'0.1090
    1'},{'word':'성교007','word_value':2.51470', 'distance':'0.27377'}]"
},
{
    "topic_id" : "3",
    "topic_value" : "3.25237",
    "children" : "[{'word':'가톨릭-교도001','word_value':3.05050', 'distance':'0.201
    87'},{'word':'천학-쟁이001','word_value':3.06220', 'distance':'0.19017'},{'wor
    d':'가톨릭^신자001','word_value':3.45950', 'distance':'0.20713'},{'word':'천주
    교-인001','word_value':3.09275', 'distance':'0.15962'},{'word':'가톨릭교-인00
    1','word_value':3.48269', 'distance':'0.23032'},{'word':'천주교-도001','word_va
    lue':2.95761', 'distance':'0.29476'},{'word':'구교-도001','word_value':3.5178
    8', 'distance':'0.26551'},{'word':'천주학-쟁이001','word_value':3.44534', 'dista
    nce':'0.19297'}]"
},
```


〈그림 33〉 JSON 형식의 배열 스크립트

각 토픽의 대표어로 어느 어휘를 선정할 것인지는 추후에 더 연구가 필요하다고 생각된다. 다만 동의어 군집화의 활용도는 더 커질 것이다. 생성된 동의어 세트 전체는 별도의 파일로 첨부하였다('topic.json').

4. 활용방안

동의어 세트는 검색 엔진에 바로 사용할 수 있다. 유명한 오픈소스 검색 엔진 Apache Lucene에서는 동의어 검색을 위해 'synonym.txt'라는 사전 파일을 사용하는데, 이 'synonym.txt'의 문법은 동의어로 군집화된 어휘의 배열을 사용하고 있다. 다른 검색 엔진에서도 동의어 확장 검색에 쓰이는 사전 파일 규

격은 대동소이하다.

언어처리에 있어서도 Java의 NLP나 Python의 NLTK도 동의어 세트를 이용하고 있다. 즉 어떤 어휘의 동의어를 나열하는데 A→B 형식의 동의어 사전을 조회하는 것이 아니라 동의어 세트에 포함되어 있는지 조회하고(one of them), 그 어휘 세트 전체를 반환하여 활용하고 있다.

감성 분석에 있어서도 서로 다른 표현의 어휘를 같은 의미로 묶어서 처리하는 데 동의어 세트는 유용하게 활용될 수 있다.

사업 책임자: 김소정 소장(나라지식정보)
사업 참여자: 윤택기, 윤택기, 김종수, 김동진, 전화자, 김혜연, 최운천, 김정호, 서민석, 안정균, 김장현, 권혁주, 전호섭, 김은경, 하지연, 김기형
담당 연구원: 이승재(국립국어원 언어정보과장)
최정도(국립국어원 학예연구사)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 152
전화 02-266-9775, 전송 02-2669-9727
인쇄일: 2020년 2월 20일
발행일: 2020년 2월 20일
인 쇄: ㈜나라지식정보

※ “이 책은 국립국어원의 용역비로 수행한 ‘한국어 정보 처리를 위한 어휘 관계 기초 자료 구축’ 사업의 결과물을 발간한 것입니다.”


국립국어원