

2016년도 연구과제 요약문(초록)

연구과제명	국 문	『한국민족문화대백과사전』 어휘 분석과 정보의 지능화 연구	
	영 문	A Study on the Lexical Analysis and Information Intellectualization of "Encyclopedia of Korean Culture"	
연구책임자	성 명	정 치 영 (한국학중앙연구원 교수)	
	생년월일		
	전화번호(휴대폰)		
	E-mail		
공동연구자	성 명	소 속	직 위
	김 병 선	한국학중앙연구원	교수
	이 건 식	단국대학교	부교수
	임 준 근	한국학중앙연구원	전문위원
	성 광 동	한국학중앙연구원	책임연구원
강 재 광	한국학중앙연구원	선임연구원	
연구 기간	2016년 5월 19일 ~ 2016년 11월 30일		연구형태 공동연구
연구목적	이 연구는 "『한국민족문화대백과사전』의 항목 풀이에 사용된 어휘를 조사하고 그 기본형을 추출하여, 이 어휘들이 '민백'의 표제어와 어떤 관련이 있는지를 분석하는 것을 목적으로 한다. 또한 명사 어휘를 중심으로 인접 어휘와의 관련성을 탐구하여 『민백』의 지식 정보의 처리에 대응하려는 것이다.		
연구내용	『민백』의 풀이에 사용된 2천만 개(token)의 어휘에 대하여 기본형(type)을 분석하였고, 항목명과 대조하는 작업을 했다. 『민백』 DB로부터 어휘를 추출하여 문맥과 함께 KWIC 형식으로 재구성하고, 어절의 기본형과 품사를 분석하였다. 특히 고유명사 어휘에 대해서는 그 세부 분류를 적용하였다. 명사 어휘와 동사의 명사 어근 어휘를 추출하고, 이 어휘로부터 항목화의 가능성이 있는 어휘를 탐색하였다. 이 보고서에서는 대량의 자료를 처리하기 위한 일련의 전산 처리 과정, 프로젝트를 위한 전산 환경 등도 밝혔으며, 프로젝트의 진행 내용은 물론 통계적 분석도 실시하였다.		
연구성과	연구의 결과 『민백』에는 약 100만 종(type)의 어휘가 사용되었으며, 명사류의 어휘가 압도적으로 많은 것을 밝혀냈다. 항목별로 어휘의 사용량과 주요 어휘를 분석해냈다. 이 연구결과는 향후 『민백』의 신규 표제항 개발에 적용할 수 있으며, 차년도에 실시할 어휘 정보 네트워크 연구의 자료로 사용될 수 있다. 연구결과물로는 다음과 같은 테이블 시트를 제출하였다. (1) 민백 KWIC (2) 민백 어휘사전 (3) 민백 항목의 구성 (4) 명사 어휘 목록 (5) 인명이 공통으로 출현하는 항목 분석 (6) 명사형 동사 어근 분석		
한글 키워드	한국민족문화대백과사전, 코퍼스, 통계처리, 기본형분석		
영문 키워드	encyclopedia, corpus, statistics, lemmatization		