2023-2 언어데이터과학 기말프로젝트 연구계획서 샘플

한국어 부사 '완전'의 통사적·의미적 확장에 관한 연구를 위한 '모두의 말뭉치' 활용

2020-20202 박수민

1 연구 개요

1.1 연구 배경

한국어 '완전'은 20세기 이전까지 명사로만 사용되었으나, 20세기 초의 문헌 자료에서부터 '모든 측면에서 두루 (완벽하게)'의 의미를 가진 부사적 쓰임이 발견되었고, 20세기 말에는 강조부사 '아주, 매우, 정말, 진짜' 등과 유사한 용법이 관찰된다(김다미 2021).

부사로 사용된 '완전'의 의미가 '아주'와 같이 추상화되는 현상은 비교적 최근에 일어난 언어 변화로, 현재에도 진행 중인 것으로 보인다. 그러므로 현대 한국어 사용자들 사이에서도 연령 등의 요인에 따라 '완전'의 분포가 달라질 수 있다.

1.2 연구 목적

이 프로젝트에서는 현대 한국어에서 '완전'의 용법이 발화자의 연령과 성에 따라 차이를 보이는지를 실제로 사용된 언어 데이터로 확인할 것이다.

1.3 데이터 선정

이 목적을 달성하기 위해서는 (i) 현실의 발화로 구성되고 (ii) 발화자의 연령·성 정보가 포함된 코퍼스를 사용해야 한다. 이런 두 조건을 만족하는 코퍼스로, 대한민국 국립국어원에서 공개한 [[모두의 말뭉치]] 중 [일상 대화 말뭉치]와 [온라인 발화 말뭉치]를 선택했다.

2 데이터 구축 및 활용 계획

2.1 데이터 수집

[[모두의 말뭉치]]는 공식 사이트(https://corpus.korean.go.kr)에서 신청한 후 승인을 받아 다운로드할 수 있다.

2.2 데이터 가공

[일상 대화 말뭉치]는 560MB, [온라인 대화 말뭉치]는 835MB로 파일이 매우 크다. GitHub 레포지토리에서 파일을 저장하고 사용하기 위해서는 각 파일의 크기가 100MB를 넘지 않아야 하므로, 파일이 압축된 상태에서 꼭 필요한 정보만을 로드해야 한다.

또한 [[모두의 말뭉치]]의 각 코퍼스는 모두 JSON 파일로 이루어져 있다. Python에서 '완전'의 용례와 발화자 정보를 손쉽게 검색하기 위해서는 JSON 형식을 데이터프레임 형식으로 변환할 필요가 있다.

2.3 데이터 활용

데이터프레임이 구축되면 pandas 라이브러리와 정규표현식을 사용하여 발화 형태로부터 '완전'의 출현 문맥을 추출할 수 있다.

참고 문헌

- 국립국어원 (2022). 국립국어원 온라인 대화 말뭉치(버전 1.1). URL: https://corpus.korean.go.kr
- 국립국어원 (2022). 국립국어원 일상 대화 말뭉치 2021(버전 1.0). URL: https://corpus.korean.go.kr
- 김다미 (2021). 〈부사적 쓰임을 보이는 '완전'에 대한 통시적 고찰〉. 《국어학》 97, 439-475. URL: https://kiss.kstudy.com/thesis/thesis-view.asp?key=3878140