한국어정보처리

2014130069 국어국문학과 박찬희

가사 코퍼스 분석 보고서
장범준 1-3집

1. 1차 분석

- 총 3386어절
- 규칙 적용 1245어절
- 미등록어 포함 110어절
- 가사는 벅스를 활용하여 검색하였다.
- 검색된 가사 내용 중 가장 빈번한 오류는 띄어쓰기 영역이었다.

ex) 별빛속으로, 떠나야만해, 생각하는건, 그날밤, 잠못, 노랠해, 할때, 그모습, 쉴때, 두눈에 등

- 오타로 볼 만한 단어는 '애태오르고' 하나 뿐이었다.

2. 2차 분석

- 총 3466어절
- 규칙 적용 1334어절
- 미등록어 포함 32어절
- 띄어쓰기를 적용한 결과 총 어절과 규칙 적용 어절이 모두 증가했음을 알 수 있다.
- 여전히 남아 있는 NF 태그 중 16개는 허밍음(워우워)이었다.
- 가사적 허용이라고 볼 수 있는 부분이 8개였다 (느지막에, 여어었어, 나는)

3. 키워드 분석

- R 프로그램을 이용하여 키워드를 분석해보았다.
- 총 3집에 걸친 앨범에서 가장 많이 언급된 명사의 순위는 다음과 같다.

1	사랑	91	11	상상	12
2	그대	72	12	노래	11
3	그녀	57	13	그때	10
4	당신	32	14	시간	10
5	엄마	25	15	천천	10
6	오늘	24	16	환상	10
7	노랠	21	17	별빛	9
8	마음	15	18	이별	9
9	모습	12	19	홍대	9
10	사람	12	20	빗속으로	8

- 사랑을 노래하는 곡이 많은 가수답게 '사랑'이 가장 많이 등장하였다.
- 마찬가지로 사랑을 고백하는 대상인 '그대', '그녀', '당신'이 뒤를 이어 2, 3, 4등을 차지했다.

- 5등의 엄마는 무려 한 곡 '엄마 용돈 좀 보내주세요'에서 나왔다.
- 6등의 '노랠'은 대개 '부르다', '하다'와 쌍으로 나왔다.

4. 용언 예측

- 프로그램 능력의 한계로 명사만 분석했지만, 가사 중에서 용언부 통계를 내릴 수 있다면 아마도 동사 중에선 '부르다', '보고 싶다', 형용사 중에선 '좋다', '예쁘다', 어미 중에선 '-요'가 가장 많을 것 같다.

5. 워드클라우드

- 3에서 키워드 분석했던 자료를 바탕으로 워드 클라우드를 만들어보았다.
- 총 빈도수가 5회 이상인 단어들이 워드클라우드에 수록되었다.