

문/이과 및 전공 구분에 따른 논문 작성 방식 비교

12조 박찬희 이다은 장윤나 전순표

1. 연구 소개

2. 코퍼스 구축

3. 코퍼스 분석

4. 결론

목차

1. 연구 소개

1. 연구 소개

연구의 필요성

- 문/이과 논문에 대한 일반적인 인식과 실재가 일치하는가?
- 문/이과 논문의 어휘, 구문에서 차이가 나타나는가?
- 논문 평가의 척도 제공
- 고급 한국어 학습자를 위한 학습자료
- 격식적 담화 코퍼스를 구축하여 향후 자동 수정, 자동 완성 등의 기술에 활용

1. 연구 소개

연구 방법

2. Word list, N-gram 분석

4. 분야별 특성 추출



1. 코퍼스 구축

3. 형태소 분석

5. Pajek을 이용한 시각화

2. 코퍼스 구축

2. 코퍼스 구축

주제 선정

KISS 분류			KCI 분류			KDC 분류		
어문학	인문과학	사회과학	자연과학	공학	의약학	농학	수해양	예체능
<input type="checkbox"/> 전체	<input type="checkbox"/> 언어학	<input type="checkbox"/> 국어학	<input type="checkbox"/> 국문학	<input type="checkbox"/> 중어중문학	<input type="checkbox"/> 일어일문학			
<input type="checkbox"/> 아랍어문학	<input type="checkbox"/> 영어학	<input type="checkbox"/> 영문학	<input type="checkbox"/> 불어학	<input type="checkbox"/> 독어학	<input type="checkbox"/> 독문학			
<input type="checkbox"/> 서반어문	<input type="checkbox"/> 이탈리아어문	<input type="checkbox"/> 노어노문	<input type="checkbox"/> 기타제어문					
<input type="text"/> 초기화								
<input type="button" value="검색"/>								




- KISS의 분류를 참고하여 분야 선정
- 논문 수집 - KISS와 DBpia 등 이용
- 계열별로 각각 300개의 논문을 수집
- 계열 내에서 특정 전공에 편중되지 않도록 각 학문에서 비슷한 비율로 수집
- 2000년대에 작성된 논문으로 한정

2. 코퍼스 구축

문과	인문과학	한국사, 동양사, 서양사, 인류학, 동양철학, 서양철학, 종교학, 민속학
	사회과학	법학, 정치/외교학, 행정학, 경제학, 경영학, 회계학, 지역개발, 관광학, 지역학, 사회학, 사회복지, 신문방송, 심리학, 문헌정보학, 교육학, 인문지리
	어문학	언어학, 국어국문학, 중어중문학, 일어일문학, 아랍어문학, 영어영문학, 불어불문학, 독어독문학, 서어서문학, 노어노문학
이과	자연과학	수학, 통계학, 물리학, 화학, 생물, 지질, 자연지리, 천문학, 대기화학
	공학	기계공학, 조선공학, 항공공학, 산업공학, 전기공학, 전자공학, 화학공학, 환경공학, 자원공학, 토목공학, 건축공학, 컴퓨터공학
	의약학	해부학, 병리학, 보건학, 내과학, 소아과학, 피부과학, 신경과학, 정신과학, 외과학, 산부인과학, 정형외과학, 신경외과학, 안과학, 마취과학, 방사선과학, 재활의학, 가정의학, 치의학, 간호학, 약화학, 한의학

2. 코퍼스 구축

전처리 과정

 어문학논문	2019-05-23 오후...	파일 폴더
 어문학 병합.pdf	2019-05-23 오후...	Adobe Acrobat 문...
 어문학.txt	2019-05-23 오후...	텍스트 문서

- 코퍼스 분석의 정확성을 높이기 위함
- 계열 별 논문 pdf 파일 병합 후 txt 파일로 변환
- 영문 초록, 서지 정보, 참고 문헌 등 불필요한 정보 삭제
- 영문, 숫자, 수식 필요에 따라 삭제
- R, Python을 이용하여 정제한 후 필요한 부분 수작업 처리

2. 코퍼스 구축

최종 코퍼스 구축 결과

구분	어휘 수(type)	어절 수(token)
문과	469,196 (0.70)	3,630,145 (0.69)
이과	193,732 (0.29)	1,662,741 (0.31)
전체	662,928	5,292,886

계열	어휘 수(type)	어절 수(token)
인문과학	227,054 (0.34)	1,260,043 (0.24)
사회과학	189,026 (0.29)	1,280,955 (0.24)
어문학	189,164 (0.29)	1,089,147 (0.21)
자연과학	75,520 (0.11)	551,134 (0.10)
공학	73,008 (0.11)	378,852 (0.07)
의약학	99,346 (0.15)	732,755 (0.14)

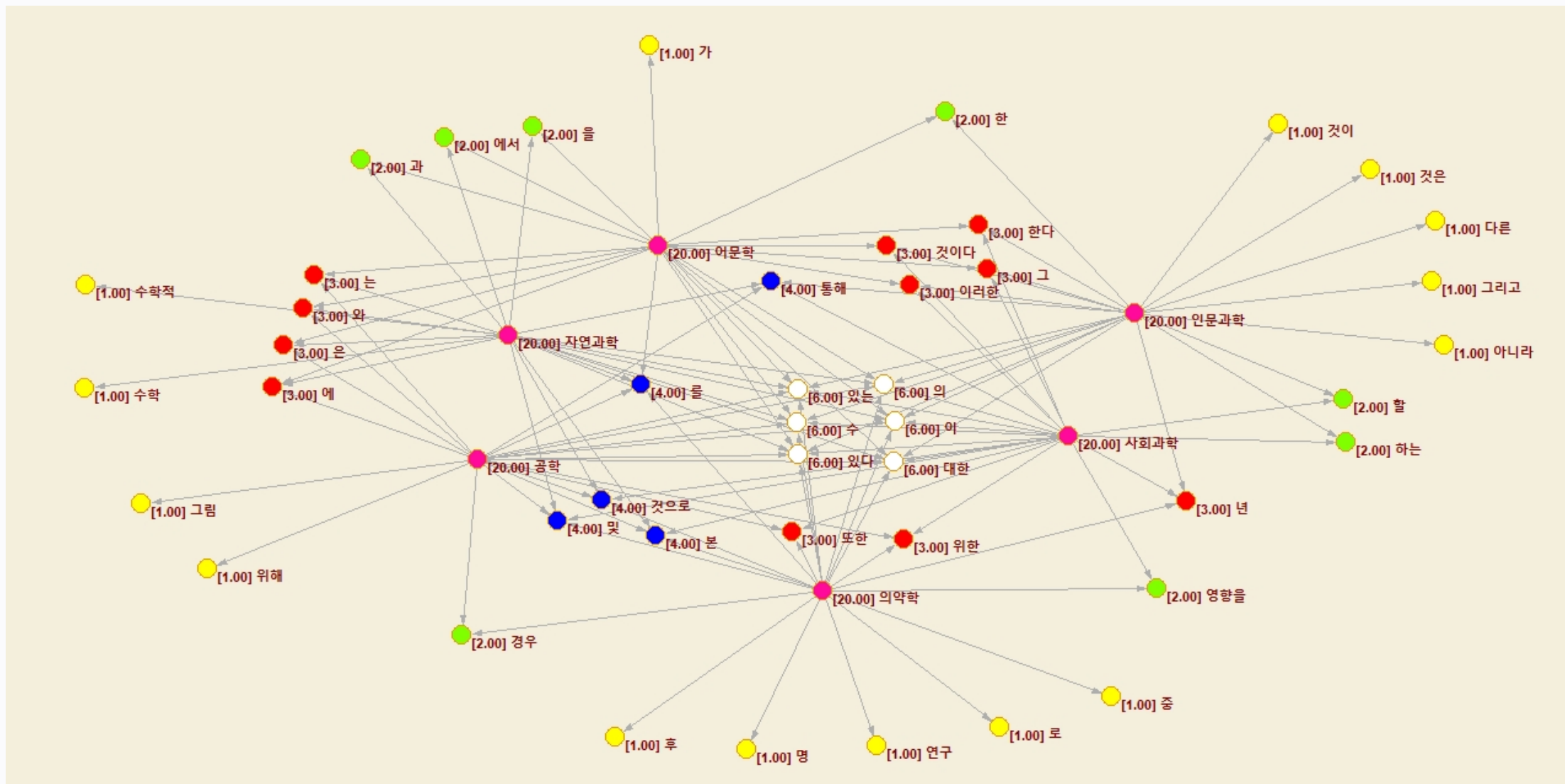
괄호 안의 숫자는 전체 대비 각 코퍼스의 비율을 의미 (소수점 셋째 자리에서 반올림)

3. 코퍼스 분석

3. 코퍼스 분석 - 상위 200개 word list

	인문과학		사회과학	2700	어문학		자연과학		공학		의약학	1600
1	11721	있다	13402	수	9818	수	6037	수	4207	있다	5840	및
2	11229	수	12638	있다	9677	있다	4613	있다	3896	수	5604	수
3	7513	대한	10139	대한	9361	의	4024	대한	2888	및	4896	대한
4	6726	이	6776	있는	6985	이	3685	의	2522	의	4752	있다
5	6680	있는	5963	및	5040	그	2446	는	1800	대한	4425	본
6	6064	그	4700	년	4767	있는	2434	및	1735	있는	3527	있는
7	5922	것이다	4682	것이다	4725	대한	2386	있는	1583	는	3431	의
8	4120	한다	4567	것으로	4712	는	2299	이	1444	를	3292	후
9	3631	이러한	4270	이러한	4293	것이다	1999	수학	1374	본	3259	것으로
10	3411	의	3910	이	4011	가	1926	를	1273	에	2136	로

3. 코퍼스 분석 - Pajek



3. 코퍼스 분석 - 형태소 분석 / 명사, 부사, 동사

문/ 이과 분석

- 문과와 이과로 코퍼스를 나누어 형태소 분석
- 일반 명사, 접속 부사, 일반 부사, 일반 동사에 대해 각각 상위 200여개 추출 후 비교

계열 별 분석

- 계열 안에서의 차이를 알아보기 위해 의학 / 공학 + 자연과학 / 사회과학 / 어문학 + 인문학으로 분류
- 4개의 계열 별 일반 명사, 접속 부사, 일반 부사, 일반 동사에 대해 각각 상위 100 - 150개 추출 후 비교
- ✓ 일반 명사의 상위 목록에서는 각 분야의 주제어들이 많이 등장, 확연한 차이 보임
- ✓ 접속 부사, 일반 부사, 일반 동사에서는 상위 목록에 비슷한 분포, 하위에서 차이 드러남

3. 코퍼스 분석 - 형태소 분석 / 명사

의학 명사 형태소 개수 : 1403922		자연_공학 명 형태소 개수 : 1713296		사회 명사 형태소 개수 : 2503263		어문_인문 명 형태소 개수 : 4385634	
kor tag num		kor tag num		kor tag num		kor tag num	
0 연구 NNG 13131		0 연구 NNG 6940		0 연구 NNG 13804		0 연구 NNG 13194	
1 환자 NNG 6548		1 수학 NNG 6556		1 사회 NNG 8513		1 도서관 NNG 7904	
2 결과 NNG 5244		2 사용 NNG 5139		2 관계 NNG 5760		2 의미 NNG 7671	
3 후 NNG 4341		3 분석 NNG 5016		3 결과 NNG 5328		3 사회 NNG 7493	
4 사용 NNG 3599		4 결과 NNG 4764		4 분석 NNG 5299		4 말 NNG 7154	
5 간호 NNG 3540		5 문제 NNG 4754		5 효과 NNG 4864		5 경우 NNG 6350	
6 경우 NNG 3516		6 학생 NNG 4555		6 영향 NNG 4740		6 사용 NNG 6173	
7 점 NNG 3376		7 경우 NNG 4064		7 국가 NNG 4718		7 문화 NNG 5643	
8 치료 NNG 3334		8 방법 NNG 4060		8 경우 NNG 4390		8 점 NNG 5609	
9 분석 NNG 3152		9 과정 NNG 3838		9 문제 NNG 4282		9 문제 NNG 5596	
10 효과 NNG 3109		10 이용 NNG 3641		10 학습 NNG 3815		10 때 NNG 5192	

3. 코퍼스 분석 - 형태소 분석 / 동사

의학 동사 형태소 개수 : 1403922		자연_공학 동 형태소 개수 : 1713296		사회 동사 형태소 개수 : 2503263		어문_인문 동 형태소 개수 : 4385634	
kor tag num		kor tag num		kor tag num		kor tag num	
0 있 W 8520		0 있 W 12310		0 있 W 17951		0 있 W 28006	
1 하 W 4862		1 하 W 5243		1 하 W 8092		1 하 W 16937	
2 되 W 2239		2 되 W 2422		2 되 W 3957		2 되 W 8352	
3 받 W 1786		3 나타내 W 1322		3 미치 W 2475		3 보 W 4700	
4 보 W 1500		4 보 W 1074		4 받 W 2100		4 받 W 3459	
5 미치 W 1387		5 가지 W 1052		5 보 W 1969		5 갖 W 2068	
6 보이 W 633		6 만들 W 855		6 가지 W 1186		6 보이 W 1994	
7 위하 W 617		7 받 W 824		7 살펴보 W 1106		7 가지 W 1927	
8 나타내 W 604		8 얻 W 821		8 보이 W 982		8 들 W 1907	
9 이 W 502		9 위하 W 807		9 갖 W 947		9 살펴보 W 1810	
10 가지 W 492		10 다루 W 756		10 나타나 W 873		10 만들 W 1719	

3. 코퍼스 분석 - 형태소 분석 / 명사, 부사, 동사



3. 코퍼스 분석 - N-gram

이과전 n=2	n=3	문과전 n=2	n=3	전체 n=2	n=3
4897 수 있다	633 알 수 있다	11570 수 있다	2171 할 수 있다	10467 수 있다	2767 할 수 있다
2861 수 있는	596 할 수 있다	6252 수 있는	1709 볼 수 있다	9113 수 있는	2253 볼 수 있다
1474 본 연구에서는	544 볼 수 있다	3677 할 수	1286 알 수 있다	4934 할 수	1919 알 수 있다
1257 할 수	272 확인할 수	2694 볼 수	996 수 있을 것이다	3562 볼 수	1259 수 있을 것이다
1162 본 연구는	271 확인할 수	2037 알 수	553 년 월 일	3152 알 수	663 것을 알 수
1115 알 수	270 것을 알 수	1791 년 월	442 할 수 있는	2506 본 연구에서는	637 년 월 일
1015 것으로 나타났다	263 수 있을 것	1630 수 있을	393 것을 알 수	2439 것으로 나타났	612 할 수 있는
945 수 있도록	238 수 있도록	1424 것으로 나타났다	342 있음을 알 수	2319 볼 수	611 확인할 수 있다
925 본 연구에서	208 수 있을 것	1338 있을 것이다	339 확인할 수 있다	2295 년 월	521 수 있다 또한
899 본 연구의	203 알 수 있었	1275 필요가 있다	317 영향을 미치는 것	2227 수 있을	499 있음을 알 수

-르 수 있다

다음과 같다

통계적으로 유의한 차이가

있는 것으로 나타났다

본 연구에서는

뿐만 아니라

3. 코퍼스 분석

문과

- ‘것이다’ 와 ‘것은’ 이과의 2.5배 이상

결론을 도출하는 방식, 결과 해석의 합리성을 입증

- 부정 표현의 빈번한 사용

‘아니라’, ‘않고’ 등, 기존 이론이나 관측에 대한 비판적 접근

- 논리에 기반한 전개

‘그러나’, ‘따라서’ 등 다수의 접속사

3. 코퍼스 분석

이과

- 관찰 및 통계에 기반한 차이에 대한 분석

분석, 확인, 결과, 평균 등

- 수치 정보 다수 포함

β , μ 등 통계 / 산술과 관련된 각종 문자, 수식, 그림 도표 등 인용 자료

- ‘것으로’ (+나타났다, 밝혀졌다, 드러났다 등의 동사) 결과에 대한 타당성 설명

- ‘및’ → 열거 / ‘각각’, ‘각’, ‘경우’ → 실험 결과에 대한 기술

4. 결론

한계

- 구조, 논지의 전개 등 어휘 이상의 단위 분석이 불가능
- 서론, 본론, 결론의 구분이 없는 코퍼스
- PDF -> TXT 변환 프로그램에서 맞춤법 및 띄어 쓰기가 완전하지 못한 부분 존재
- 영문과 숫자를 제거하다 보니 조사만 남는 경우가 많아 조사가 고빈도로 집계
- 코퍼스의 크기

4. 결론

의의 및 활용 방안

- 추측에 불과했던 문/이과 논문에서의 사용 어휘가 실제로 다르게 나타나는 것 확인
- N-gram 분석 결과를 통해 자주 사용되는 어구의 예시 제공 가능
- 논문 평가의 척도 제공
- 고급 한국어 학습자를 위한 학습자료로 사용
- 격식적 담화 코퍼스를 구축하여 향후 자동 수정, 자동 완성 등의 기술에 활용 가능

4. 결론

수정 및 보완 계획

- 계열 별로 제목만 모아서 사용 어휘 비교
- 서론, 본론, 결론을 나누어서 첫 문장의 서술 방식 비교
- 목차, 인용 방식 등 형식적인 부분 비교
- 계열 별 품사에 대한 자세한 비교

질의 응답 및 토론

감사합니다!