Korea Advanced Institute of Science and Technology

KAIST

카이스트 어휘망과 명사의미 연구

배선미, 시정곤
KORTERM/BOLA

Korea Advanced Institute of Science and Technology

차례

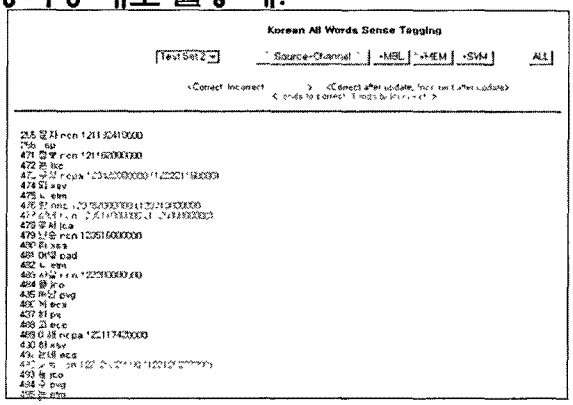
KI

2

Korea Advanced Institute of Science and Technology

카이스트 어휘망의 목적

- 카이스트 한국어 단어망은 자연언어처리를 함에 있어 의미의 애매성 해소를 위하여 개념체계를 기반으로 구축된 어휘 의미망이다.
- 중의성 해소 활용 예:



KAIST

3

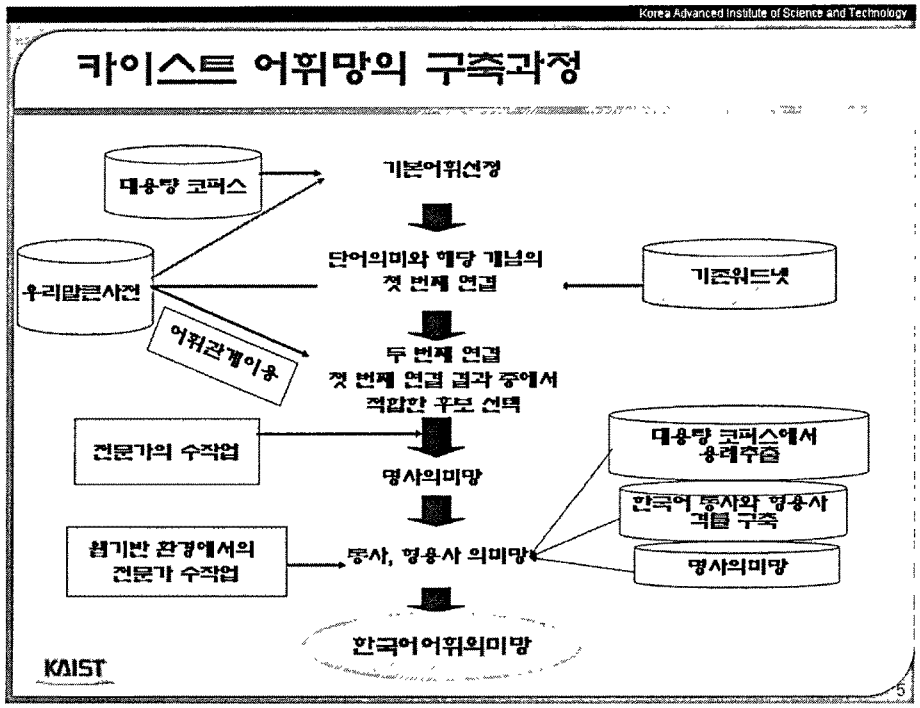
Korea Advanced Institute of Science and Technology

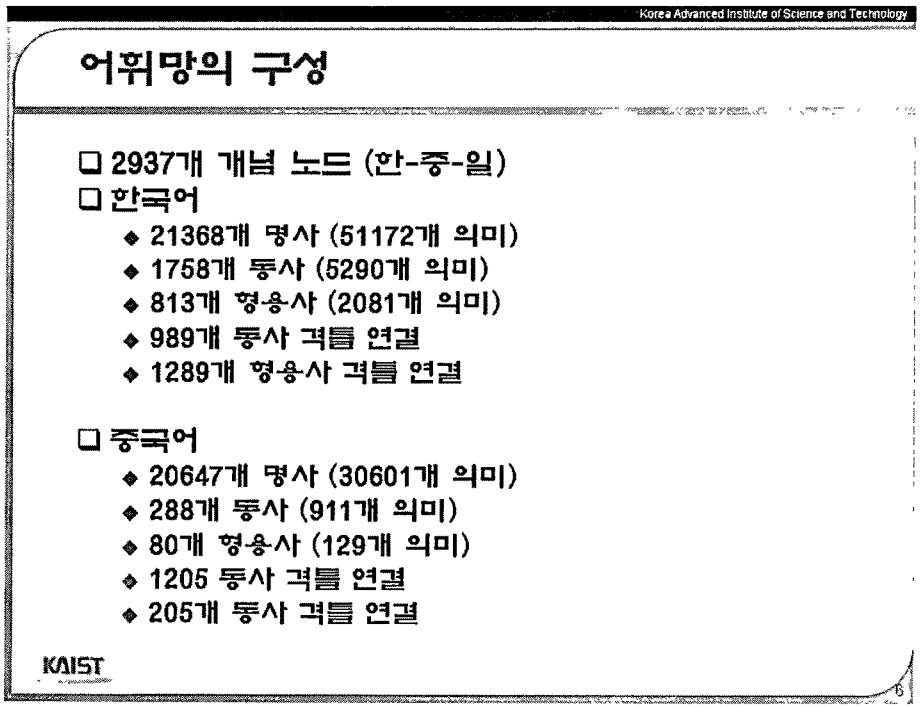
카이스트 어휘망의 특징

- 종합적 체계
  - 명사가 먼저 구축된 다음, 명사 어휘망을 활용하여 동사와 형용사 단어망을 만듦(하나의 개념체계 안에 명사/동사/형용사 연결)
  - 동사와 형용사 어휘망에서는 격틀과 명사 어휘망까지 연결하여 명사, 동사, 형용사, 격틀을 아우르는 입체적이고 종합적인 체계로 구축
- 코퍼스 기반
  - 한국어 어휘망은 어휘 선정에서부터 서술어 논항 목록까지 철저하게 코퍼스를 기반으로 구축되었다는 특징을 가짐
- 다국어로의 확장 가능성
  - 한-중-일 다국어 개념체계 공유
  - 한국어 단어망은 한국어를 중심으로 한/중/일/영 다국어 어휘망으로 확장(중국어는 첫 번째 버전 나옴)
  - 뜻풀이 부여
    - 구축된 모든 어휘에 우리말큰사전의 뜻풀이 부여

KAIST

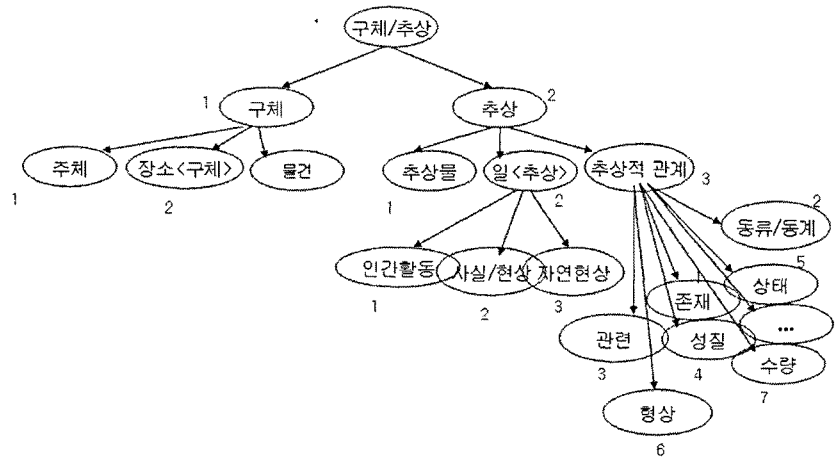
4


Korea Advanced Institute of Science and Technology
카이스트 어휘망의 구축과정
대용량 코퍼스
기본어휘선정
단어의미와 해당 개념의
첫 번째 연결
우리말큰사전
기존워드넷
어휘관계이용
두 번째 연결
첫 번째 연결 결과 중에서
적합한 후보 선택
전문가의 수작업
대용량 코퍼스에서
용례추출
명사의미망
한국어 동사와 형용사
격틀 구축
웹기반 환경에서의
전문가 수작업
동사, 형용사 의미망
명사의미망
한국어어휘의미망
KAIST
5


Korea Advanced Institute of Science and Technology
어휘망의 구성
❑ 2937개 개념 노드 (한-중-일)
❑ 한국어
◆ 21368개 명사 (51172개 의미)
◆ 1758개 동사 (5290개 의미)
◆ 813개 형용사 (2081개 의미)
◆ 989개 동사 격틀 연결
◆ 1289개 형용사 격틀 연결
❑ 중국어
◆ 20647개 명사 (30601개 의미)
◆ 288개 동사 (911개 의미)
◆ 80개 형용사 (129개 의미)
◆ 1205 동사 격틀 연결
◆ 205개 동사 격틀 연결
KAIST
6

Korea Advanced Institute of Science and Technology

개념체계 (2937개의 개념간의 계층관계)



KAIST

Korea Advanced Institute of Science and Technology

명사의미망

- ❑개념 노드와 단어 의미와의 연결
- ❑고빈도 단어 조사하여 목록 선정
- ❑자동으로 해당 개념번호 후보 제시
  - ◆기존 일본어 명사 어휘대계와 한국어 단일어 사전 이용
- ❑전문가의 수동 선택
- ❑한글학회 우리말큰사전 기준으로 의미 구분

KAIST

Korea Advanced Institute of Science and Technology

동사, 형용사 의미망

□작업 설명

- 동사와 형용사는 문장에서의 역할이 명사와 다르며, 의미 또한 동반하는 논항의 의미적, 통사적 특성에 의해 분화되므로 논항과 함께 고려하여야 한다.
- 한국어 동사 격틀 있으며 형용사 격틀도 구축할 계획이므로 동사 자체의 의미를 체계화한다면 격틀과 연동할 수 있을 것이다.
- "서술어가 상태와 동작을 기술하기 때문에 이를 실체화 하면 명사적 개념으로 표현 가능" (Ikehara S. et al., 1997)

KAIST

9

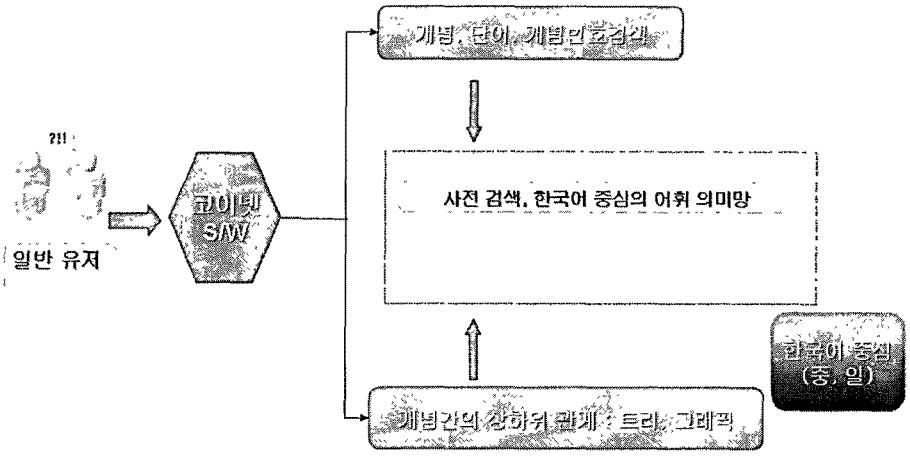
Korea Advanced Institute of Science and Technology

KAIST

코어넷 브라우저 데모

Korea Advanced Institute of Science and Technology

코어넷 개념, 단어, 개념번호 검색 브라우저



KAIST 11

Korea Advanced Institute of Science and Technology

브라우저를 통한 카이스트 어휘망의 검색

❑코어넷 브라우저의 기능

◆단어, 개념명, 코어넷 번호 검색

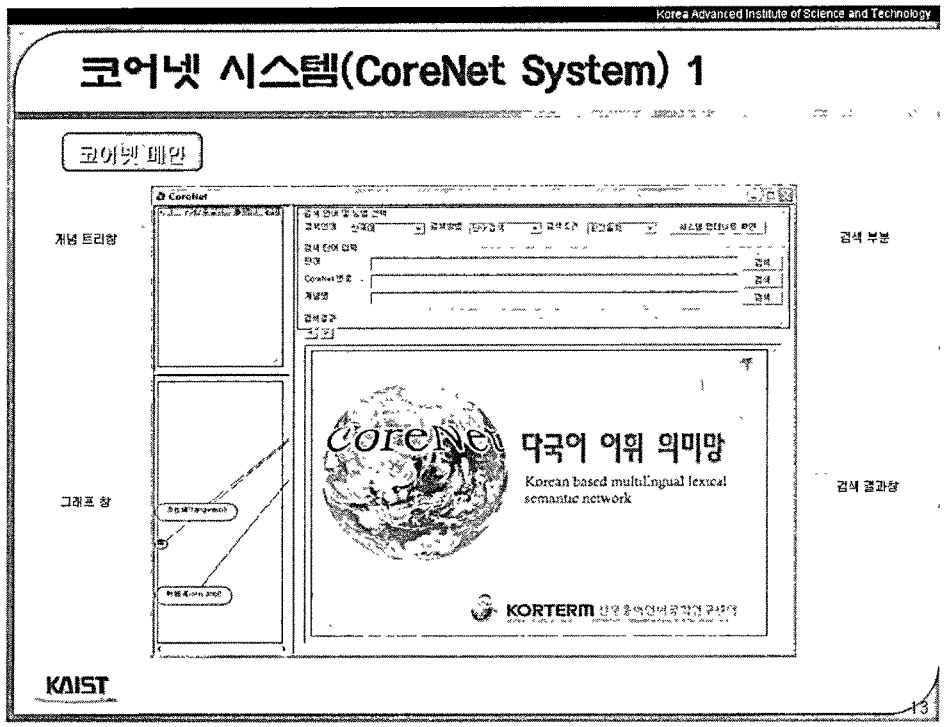
- 상하위 관계 및 관련 어휘를 보여주는 그래픽 에디터 지원
- 개념체계를 나무구조로 보여주는 에디터 지원
- 우리말 큰사전과의 매핑 지원
  (어휘번호 및 의미번호 매핑)

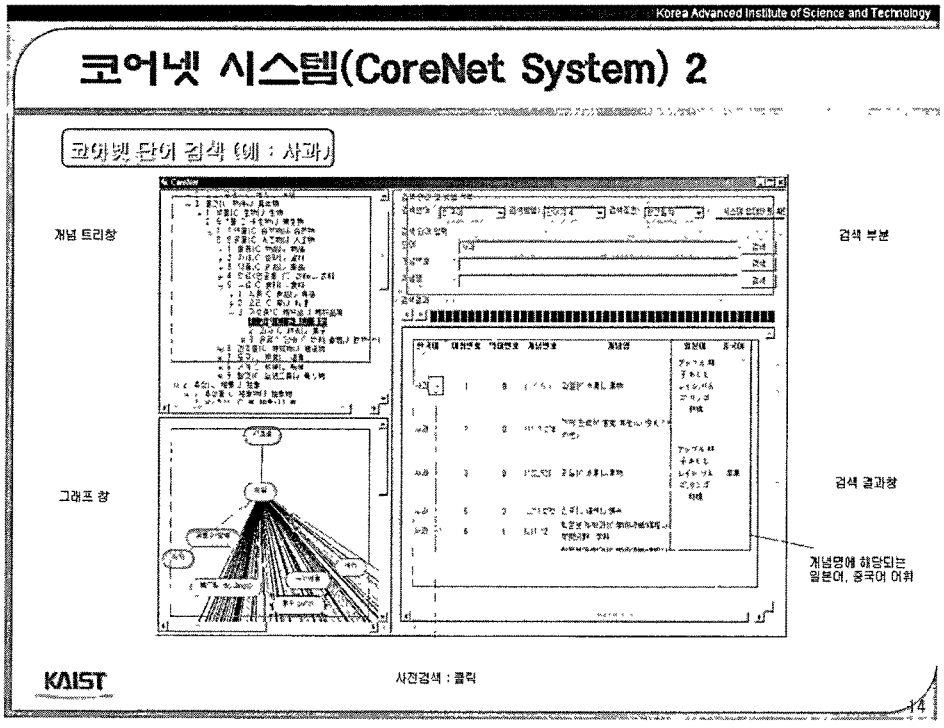
◆동사와 형용사 격틀 검색

- 동사의 개념번호에 따른 논항의 개념명 및 논항의 단어 검색 지원

❑ 브라우저 데모

KAIST 12

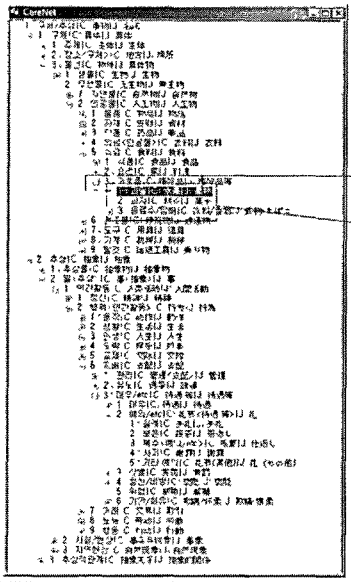

Korea Advanced Institute of Science and Technology
코어넷 시스템(CoreNet System) 1
코어넷 메인
개념 트리창
그래프 창
검색 부분
검색 결과창
CoreNet
다국어 어휘 의미망
Korean based multilingual lexical semantic network
KORTERM
KAIST
13


Korea Advanced Institute of Science and Technology
코어넷 시스템(CoreNet System) 2
코어넷 단어 검색 (예 : 사과)
개념 트리창
그래프 창
검색 부분
검색 결과창
개념명에 해당되는 일본어, 중국어 어휘
KAIST
사전검색 : 클릭
14

Korea Advanced Institute of Science and Technology

코어넷 시스템(CoreNet System) 3

개념 트리창



코어넷 단어 검색(예 : 사과)

1. 개념 트리 탐해가 가능
2. 해당 개념의 어휘를 결과창에서 확인가능
3. 개념 상호간의 관계를 트리형태로 확인가능

상위 노드

동위 노드

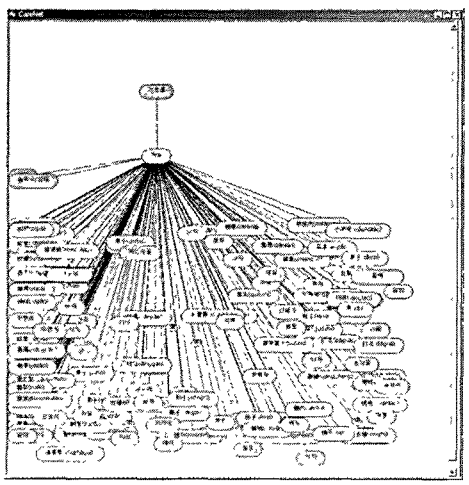
KAIST

15

Korea Advanced Institute of Science and Technology

코어넷 시스템(CoreNet System) 4

그래프 창



코어넷 단어 검색(예 : 사과)

1. 기준노드 – 노란색
2. 부모노드 – 주황색
3. 형제노드 – 녹색
4. 자식노드 – 파란색
5. 관련어휘 – 회색(한국어), 분홍색(중국어)

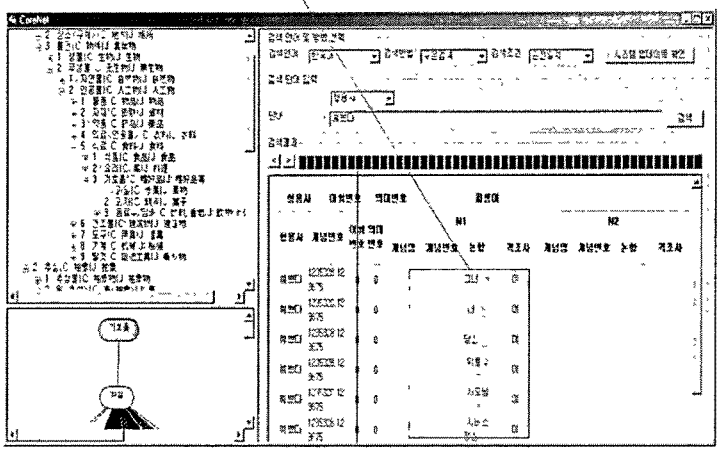
KAIST

16

Korea Advanced Institute of Science and Technology

코어넷 시스템(CoreNet System) 5

"예쁘다" 가질 수 있는 문형과 해당 논항



KAIST

17

Korea Advanced Institute of Science and Technology

KAIST

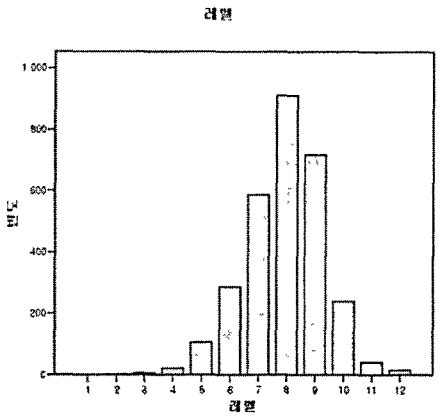
한국어 명사 개념지도

- ❑ 계층별 전체 개념분포 분석
- ❑ 명사의 개념분포 분석
- ❑ 개념명에 따른 명사 어휘 분포

Korea Advanced Institute of Science and Technology

전체 개념체계 계층별 개념분포 분석

계층	개념수	퍼센트	누적퍼센트
1	1	0.034048349	0.034048349
2	2	0.068096697	0.102145046
3	6	0.204290092	0.306435138
4	21	0.715015322	1.02145046
5	106	3.609124957	4.630575417
6	288	9.805924413	14.43649983
7	586	19.95233231	34.38883214
8	912	31.05209397	65.44092612
9	718	24.44871433	89.88784045
10	240	8.171603677	98.05924413
11	41	1.395982295	99.45522642
12	16	0.544773578	100
합계	2937	100	



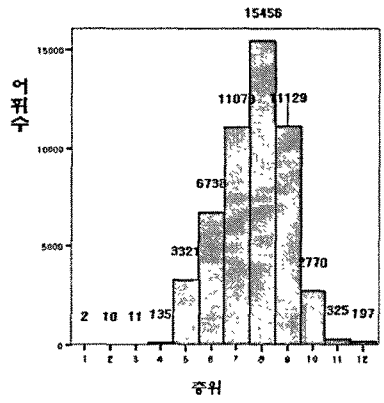
KAIST

19

Korea Advanced Institute of Science and Technology

명사의 계층별 개념 분포 분석

계층	어휘수
1	2
2	10
3	11
4	135
5	3321
6	6738
7	11078
8	15456
9	11129
10	2770
11	325
12	197
합계	51172



KAIST

20

Korea Advanced Institute of Science and Technology

개념명에 따른 명사 어휘 분포 (상위 30위)

코텀번호	명사 갯수	개념명	레벨	백분율	어휘
122323213	429	병의 종류	9	0.8383%	간암,간염,간질,간질병,,편두통,폐결핵,폐렴
12113121	382	인명	8	0.7465%	가명,성,성명,성씨,성함,칭호,휘호
1111336	330	하는 사람	7	0.6449%	가담자,가입자,감독,감독자,감시원,감정가,감사
121142	311	문서류	6	0.6078%	가계부,각서,감사장,감정서,견적서,건의서,결제장부,회의록
122111	284	마음	6	0.5550%	가슴,간담,간장,감심장,격심,격의,결의,경각심,경계심,경쟁심
12325	264	등급	5	0.5159%	각계각층,각층,간부급,갑부,갑,계급,계층,등,등위
1132261I	264	가옥(본체)	8	0.5159%	가건물,가게,가옥,가택,각호,감옥,개집,갱려리,건물
1111132A	260	사람<기타 지위>	7	0.5081%	가판,감찰,검열,경락,경사,경장,고문,고문관,공사관,공인,관찰사
1111131224	254	기타 관료	9	0.4964%	감사,검사장,경,경관,고관,공판,관장,교장
1239221	245	기간(자연/인간활동/etc)	7	0.4788%	간발,개재,개화기,개화기,갱년기,공소시효,과도기
11111212	244	여자	8	0.4768%	계모,계부,계집,계집년,계집애,공주,과부,관기,관기,국모,궁녀,궁인,귀부
121145	233	책(내용)	6	0.4553%	가이드,가정집,간행물,강선,교과서,교본,교전,구약
11222	230	지역(인간활동)	5	0.4495%	각처,각지,거소,거점,거주,거주지,거처,검문소,검사장
1131111	227	포유류	7	0.4436%	강아지,개,갱마,경우,경주마,경찰견,고릴라,고양도치,고양이
1132272222	225	출판물	9	0.4397%	가이드,가정집,경전,고서,고전,고집,관보,교본,교지,국서,국지,그림책
12373	220	단위	5	0.4299%	가구,가닥,가리,가마,가마니,가호,각,각,갈래,갑,갑,개비,거리,건,결음
113221	214	물품	6	0.4182%	가스레인지,고가품,고물,골동품,공물,공법용,군수품,귀중품
1221245	203	스포츠	7	0.3967%	강속구,개인전,게임,결승,결승전,결전,결혼,경마,경보,경주,골,골프
121113E	200	방법	7	0.3908%	강경책,교법,구법,귀납법,묘방,묘법,무법,민간요법,발성법,방도,방법,방수,방
11113191	200	군인	8	0.3908%	감시병,경계병,경비병,계엄사령관,공병,관군,군경,군관,군단장,군사,군사
12113133	196	알	8	0.3830%	각설,각설이,각설이,강안,강화,경구,경어,계사,고명,고어,공안,공치사,과언,구
1132275	186	부기	7	0.3635%	개머리판,건,검,경기,경선,경전,경탄,공기총,공포탄,관용,기관총,기포
111121	182	신분	6	0.3557%	관등보살,구세주,구주,군신,군신,귀신,대장군,도신,독신,동신
1132421	181	의복(본체)	8	0.3537%	가사,갑옷,겹옷,겨울옷,경구,공복,관복,교복,군복,기경복
1132231	179	약품류(의료용)	7	0.3498%	각성제,감초,강심제,강장제,구급약,구단,구충제,구충제,기살,기약,기제,
1211112	178	학문분야/학과	7	0.3478%	가사,가정과,경영학과,경제,경제학,고고학,고대사,고학,과학,광학
112132	174	기타 거주시설	6	0.3400%	가건물,가게,가옥,가택,객점,고층건물,곡장,골방,곳간,공동주택,기숙사
KAIST 1111	170	상<사람<지위>>	7	0.3322%	가장,감,감독,감사,감사원장,감판장,검찰부장판,검사장,검찰총장
1235113	169	상태/이상	7	0.3303%	가관,가소,걱동,경제,경직,고주망태,고혹,구김살,군통,궁상,긴급

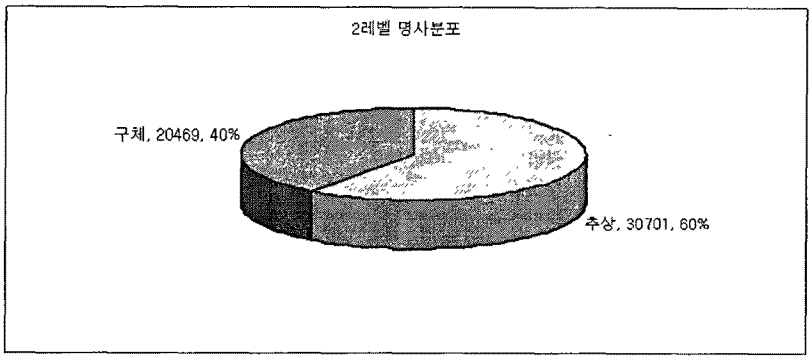
Korea Advanced Institute of Science and Technology

카이스트 명사 어휘망의 계층별 어휘분포

코텀번호	레벨	한국어	누적명사갯수	누적명사백분율
1	1	구체/추상	51172	100.00%
12	2	추상	30701	60.00%
11	2	구체	20469	40.00%
122	3	일<추상>	15507	30.30%
113	3	물건	10335	20.20%
123	3	추상적관계	8814	17.22%
111	3	주체	7166	14.00%
121	3	추상물	6379	12.47%
112	3	장소<구체>	2959	5.78%
1221	4	인간활동	10785	21.08%
1132	4	무생물	7769	15.18%
1111	4	사람	6092	11.90%
1211	4	추상물(정신)	4648	9.08%
1222	4	사실/현상	2671	5.22%
1131	4	생물	2563	5.01%
1223	4	자연현상	2050	4.01%
1212	4	추상물(행위)	1731	3.38%
1239	4	시간	1626	3.18%
1237	4	수량	1522	2.97%
1235	4	상태	1330	2.60%
1121	4	시설	1320	2.58%
1238	4	장소<추상적관계>	1217	2.38%
1234	4	성질	1067	2.09%
1112	4	조직	1067	2.09%
1123	4	자연	833	1.63%
1122	4	지역	806	1.58%
1233	4	관련	737	1.44%
1236	4	형상	570	1.11%
KAIST 1232	4	동류/동계	542	1.06%
1231	4	분재	203	0.40%

Korea Advanced Institute of Science and Technology

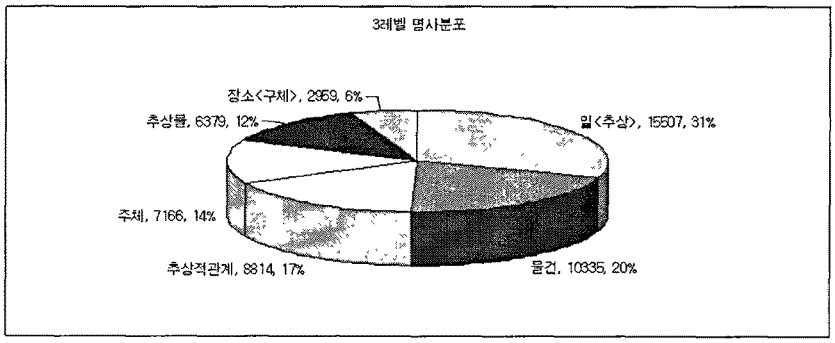
계층별 명사 개념명에 따른 어휘 분포 (2레벨)



KAIST 23

Korea Advanced Institute of Science and Technology

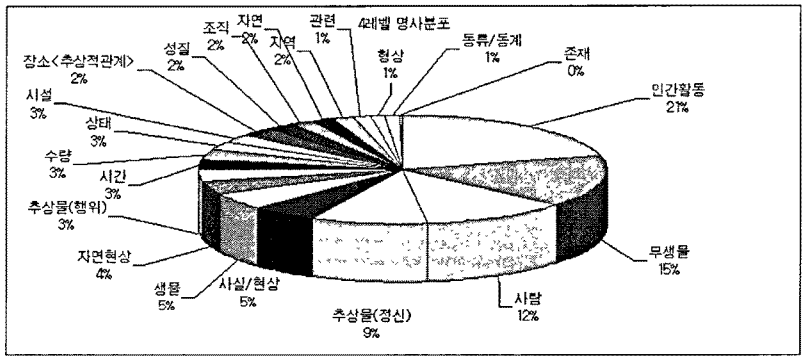
계층별 명사 개념명에 따른 어휘 분포 (3레벨)



KAIST 24

Korea Advanced Institute of Science and Technology

계층별 명사 개념명에 따른 어휘 분포 (4레벨)

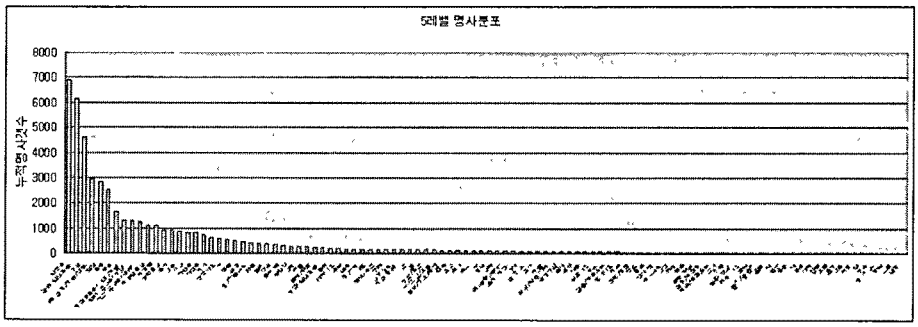


KAIST

25

Korea Advanced Institute of Science and Technology

계층별 명사 개념명에 따른 어휘 분포 (5레벨)



KAIST

26

Korea Advanced Institute of Science and Technology

맺음말

❑계층별 전체 분포 및 명사 분포 8>9>7>6... 레벨 순

❑명사의 레벨별 어휘 분포

- 2레벨: 추상(12)>구상(11)
- 3레벨: 일<추상>(122) > 물건(113)> 추상적 관계(123)> 주체(111)> 추상물(121)> 장소<구체>(112)
- 4레벨: 인간활동(1221) > 무생물(1132) > 사람(1111) > 추상물(정신)(1211) > 사실/현상(1222)> 생물(1131)> 자연현상(1223)> 추상물(행위)(1212)> 시간(1239)> 수량(1237)...
- 5레벨: 인공물 (11322)> 행위<인간활동>(12212)> 정신(12211)> 사람<직업/지위/역할>(11113)> 인간(11111)> 변동(12222)> 동물(11311)> 지적생산물(사고/학습)(12111)> 언어<추상물(정신)>(12113)> 제도<추상물(행위)>(12121)...

KAIST

Korea Advanced Institute of Science and Technology

현 명사 의미체계의 문제점 및 보완점

❑개념명과 어휘관계 미흡

- 개념명 자체의 수정
- 개념과 어휘와의 관계를 고려하여 수정
- 하위 개념 노드의 분화 미흡

❑반의어 처리 미흡

❑어휘 수의 보완 필요

KAIST

Korea Advanced Institute of Science and Technology

향후 과제

- ❑카이스트 어휘망 2차 버전-개념명과 어휘 수정
- ❑영어판의 확장 및 중국어판 보완
- ❑동사 및 형용사의 어휘 보완
- ❑카이스트 기본어휘 등급 부여
- ❑카이스트 어휘망의 활용 가능성 및 사례 연구

KAIST

29