

Text Processing of Korean Using Online Tools

March 21, 2019
CKM Session, Annual CEAL Meeting

Hyoungbae Lee
Korean Studies Librarian
Princeton University Library

Question from a Student

- How can I analyze the text of *Samch'ŏlli* (a Korean literature magazine) to examine the usage and context of certain words and visualize the results?
 - How to analyze:
 - Corpus linguistics
 - Natural language processing
 - Text (data) mining
 - How to visualize
 - Data representation
 - Data visualization

What You Need

- Corpus
 - Raw Corpus / Tagged Corpus
 - Synchronic Corpus / Diachronic Corpus
 - General Corpus / Specialized Corpus

- Text Analysis Tools
 - Frequency, Co-occurrence, Concordance, etc.

- Visualization Tools
 - Integrated in most analysis tools

Raw Corpus vs. Tagged Corpus

- Raw Corpus
 - 코퍼스란 자연언어 연구를 위해 언어의 표본을 추출한 집합이다
- Tagged Corpus
 - 코퍼스/NNG+란/JX 자연/NNG+언어/NNG
연구/NNG+를/JKO 위해/VV 언어/NNG+의/JKG
표본/NNG+을/JKO 추출/NNG+한/XSV
집합/NNG+이/VCP+다/EF

Trend21 Corpus

- Corpus with analysis & visualization tools developed by Korea University
<http://corpus.korea.ac.kr/>
- Texts of all news articles between 2000 and 2013 from *Tonga ilbo*, *Chosŏn ilbo*, *Chungang ilbo*, *Han'gyŏre sinmun*.
- 60 million word segments (語節)
 - 한국 = 1 word 1 segment / 한국의 = 2 words 1 segment
- Corpus Type: Tagged, Synchronic, Specialized



Usage Search

용례검색기

형태소 검색 검색 연도별 전체

1 개의 결과 검색된 문장 48165 개 (1/964 페이지)

keyword ▲	freq
유학/NNG	51378

“유학 (留學)”

영어의 달인 (2) 송창근 삼성라이온스 과장;“영어신도 안될땐 **유학**-연수기도 소용없어요”

해외 **유학**은커녕 연수도 가본 적 없지만, 아침에 일어날 때부터 잠잘 때까지 한국말을 단 한마디도 안한 적도 많았다.

미국 **유학**생활을 오래 하면서 다민족 사회에선 자기 정체성을 갖는 일이 얼마나 중요한가를 깨달았습니다.

국방부는 2일 외국에서 **유학**중인 현역 군인 및 군무원이 관계 당국의 승인 없이 북한인과 접촉하는 등 법률을 위반했을 경우 기무부대도 수사권을 가질 수 있도록 최근 군사법원법을 개정했다고 밝혔다.

유학가는 아버지를 따라 세살 때 미국으로 건너간 김진해(21·국제정치 전공 3년)씨는 지난해 12월 프린스턴 대학의 총학생회장 선거에서 52%의 높은 득표율로 당선됐다.

△나는 한국인이며(신세웅)=좌절과 방황 끝에 영국 옥스퍼드대학에 입학한 신세웅의 이야기. 미국을 선망의 대상으로 삼고 **유학**을 꿈꾸는 청소년들에게 쓰라렸던 자신의 경험을 고백한다.

그는 “**유학**을 가려다 몇년치 기획물을 만들어 놓으면 가족들이 먹고 살 수 있겠거니 하는 생각”에서 출판사를 열었으나 지하 셋방에서 ‘열린책들’의 고행은 막 시작이었다.

차차세대 컴퓨터 연구가 김영식(김영식·38) 서강대 교수는 이제 영어로 ‘먹고 살게’ 됐다. 북아일랜드 벨파스트 퀸즈대학 교수로 가게 된 것. 84년 런던대학 임페리얼 칼리지에 **유학**. 3년 반 만에 석·박사 학위를 마치고 90년 돌아와 서강대 교수로 일해온 그는 꼭 10년 만에 영국 대학으로 ‘역 취업’하게 됐다.

이론 물리학을 전공한 그는 컴퓨터 정보 처리 속도를 지금보다 100만배 이상 빠르게 하는 양자(양자) 전산 전문가다. “**유학** 갈 때 아버지가 딱 두 가지만 하고 오라고 말씀하시더군요. 영어를 잘하게 될 것, 친구를 많이 사귄 것이었습니다.”

유학할 수 있을 정도로 토를 점수는 받았지만, 실제 상황에선 차 한잔도 제대로 시킬 수 없는 상황이었다. “학교 식당에서 ‘어 커바브 티, 플리즈’ 했는데, 도대체 못 알아듣는 거예요. 뒤에 서있던 이가 보다가 안되겠는지, ‘너 차 한잔 마시자는 거냐’ 그래요.”

● 원문보기 ◀ 이전 문장 ▶ 다음 문장

영어의 달인 (2) 송창근 삼성라이온스 과장;“영어신도 안될땐 유학-연수기도 소용없어요”

영어의 달인	영어/NNG+의/JKG
달인	달/NNG+이/VCP+ L/ETM
(2)	(/SS+2/SN+)/SS
송창근	송창근/NNP
삼성라이온스	삼성/NNP+라이온스/NNP
과장;“영어신도	과장/NNG+;/SP+;/SS+영어/NNP+신/NNG+도/JX
안될땐	안될땐/NA
유학-연수기도	유학/NNG+-/SP+연수/NNG+가/JKS+도/JX
소용없어요”	소용없/V/A+어요/EM+;/SS

Co-occurring Words by Year

KW = 유학 (留學)

2000



2013

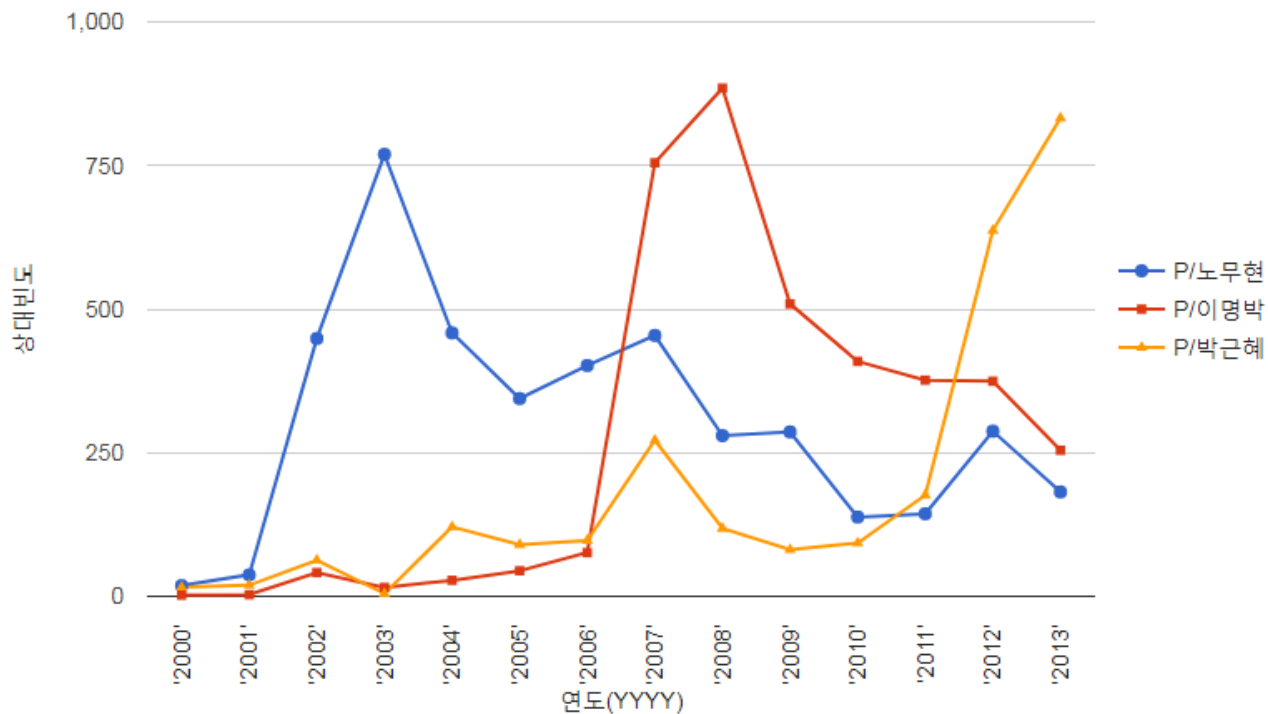


- 미국
- 해외
- 학생
- 조기
- 대학
- 시절
- 영어
- 졸업
- 교육
- 학교
- 외국
- 생활
- 공부
- 한국
- 자녀
- 일본
- 유학생
- 귀국
- 준비
- 중국
- 프로그램

Word Frequency

단어 빈도 차트

P/노무현 P/이명박 P/박근혜



Voyant Tools

- Open-source, web-based application for text analysis developed by Stéfán Sinclair (McGill) and Geoffrey Rockwell (Alberta)
<https://voyant-tools.org/>



- Analysis tools for user-provided corpus



hosis
 n Gregor Samsa woke from
 sformed
 orrible vermin. He lay on his
 could see his brown belly.
 The bedding was hardly ab
 slide
 is many legs, pitifully thin
 helplessly as he looked.
 to me?" he thought. It was
 om
 o small, lay peacefully betw
 lay spread out on the table
 n -
 hung a picture that he had

Trends Document Terms

#	Term	Count	Relative	Trend
<input type="checkbox"/>	1 gregor	298	13,456	
<input type="checkbox"/>	1 room	131	5,915	
<input type="checkbox"/>	1 father	96	4,335	
<input type="checkbox"/>	1 sister	96	4,335	
<input type="checkbox"/>	1 door	87	3,928	
<input type="checkbox"/>	1 mother	82	3,703	



Contexts

Document	Left	Term	Right
1) The ...	It wasn't a dream. His	room	, a proper human room although
1) The ...	His room, a proper human	room	although a little too small
1) The ...	the chief clerk in the	room	on the left. Gregor tried
1) The ...	be heard in the adjoining	room	. From the room on his
1) The ...	the adjoining room. From the	room	on his right, Gregor's
1) The ...	his father now from the	room	to his left, "the chief
1) The ...	forgive the untidiness of your	room	." Then the chief clerk called
1) The ...	it's hanging up in his	room	; you'll see it as soon
1) The ...	No". said Greor. In the	room	on his right there followed

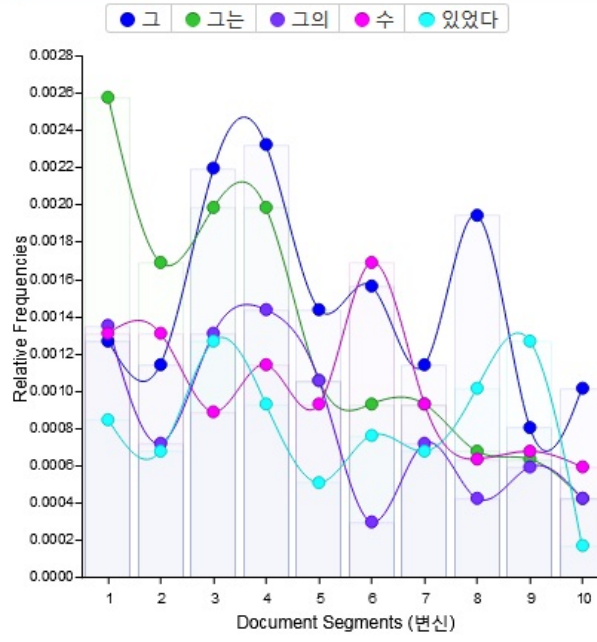
room x 131 context expand

Vocabulary Density: 0.118
 Average Words Per Sentence: 28.1
 Most frequent words in the corpus: gregor (298); room (131); father (96); sister (96); door (87)



변신

어느 날 아침, '그레고르 잠자'가 어수선한 잠에서 깨어났을 때, 자신이 훌쩍한 한 마리 해충으로 변해있는 것을 발견했다. 그것도 그의 침대 위에서.
 그는 무장한 것 같은 등을 대고 누워있었다.
 만약 그가 약간만 자신의 머리를 들었어도, 그는 불룩하게 부풀어 오른 자신의 갈색 배를 볼 수 있었을 텐데. 그 배는 약간의 동형이었고, 딱딱한 마디들로 이어진 아치형구조를 이루고 있었다. 괴상했다.
 그의 이불로는 그 훌쩍한 배를 모두 덮을 수 없이 밝혀졌다. 이불이 어느 순간 그의 배 위에서 스프르크 미끄러져 내렸기 때문이다. 처음부터 미끄러져 내릴 준비를 한 것처럼.
 아, 그의 수많은 발이라니! 비참하게 얇은 발들은 그 수가 너무 많았다. 그의 몸통이에 비해 발이 너무 많고 너무 작았다.
 그가 그 발을 보려고 할 때마다 발들은 요갈 때 없는 것처럼 요동치고 있었다.
 "뭐야! 무슨 일이 나한테 벌어진 거지?"라고 그는 생각했다.
 그러나 그것은 꿈이 아니었다. 그는 자신의 방에 누워있었다. 그 방



Terms:

This corpus has 1 document with 23,694 total words and 8,058 unique word forms. Created now.
 Vocabulary Density: 0.340
 Average Words Per Sentence: 9.7
 Most frequent words in the corpus: 그 (351); 그는 (305); 수 (239); 그의 (197); 있었다 (192)

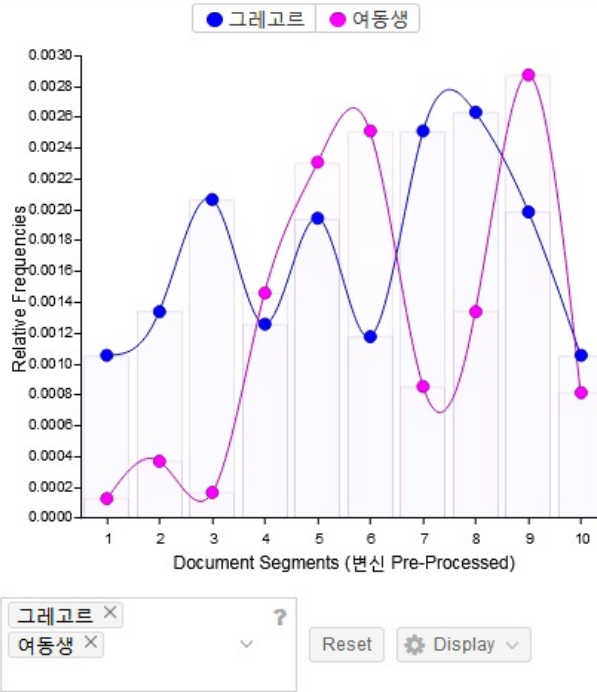
Document	Left	Term	Right
1) 변신	배를 볼 수 있었을 텐데.	그	배는 약간의 동형이었고, 딱딱한 마...
1) 변신	이루고 있었다. 괴상했다. 그의 이...	그	훌쩍한 배를 모두 덮을 수
1) 변신	수많은 발이라니! 비참하게 얇은 발...	그	수가 너무 많았다. 그의 몸통이에
1) 변신	너무 많고 너무 작았다. 그가	그	발을 보려고 할 때마다 발들은
1) 변신	아니었다. 그는 자신의 방에 누워있...	그	방이 아무리 작을지라도 그것은 영...
1) 변신	팔 전체를 들어 올리고 있었다.	그	팔에는 두툼한 포피 머피(방한용
1) 변신	부딪친 것 같은 소리가 들려왔다.	그	소리가 그를 우울하게 만들었다. "...
1) 변신	자신의 다리들 중 하나를 가져가	그	아픈 곳을 찾아보려고 노력했다. 그가
1) 변신	높기로 했다. 애초 깨어났을 때	그	상태로 누웠다. "일찍 일어나는 것은

items:



변신 Pre-Processed

어느 날 아침, '그레고르 잠자'가 어수선한 잠에서 깨어났을 때, 자신이 훌쩍한 한 마리 해충으로 변해있는 것을 발견했다. 그것도 그의 침대 위에서. 그는 무장한 것 같은 등을 대고 누워있었다. 만약 그가 약간만 자신의 머리를 들었어도, 그는 불룩하게 부풀어 오른 자신의 갈색 배를 볼 수 있었을 텐데. 그 배는 약간의 동형이었고, 딱딱한 마디들로 이어진 아치형구조를 이루고 있었다. 괴상했다. 그의 이불로는 그 훌쩍한 배를 모두 덮을 수 없이 밝혀졌다. 이불이 어느 순간 그의 배 위에서 스프룩 미끄러져 내렸기 때문이다. 처음부터 미끄러져 내릴 준비를 한 것처럼. 아, 그의 수많은 발이라니! 비참하게 얇은 발들은 그 수가 너무 많았다. 그의 몸통이에 비해 발이 너무 많고 너무 작았다. 그가 그 발을 보려고 할 때마다 발들은 요갈 때 없는 것처럼 요동치고 있었다. "뭐야! 무슨 일이 나한테 벌어진 거지?"라고 그는 생각했다. 그러나 그것은 꿈이 아니었다. 그는 자신의 방에 누워있었다. 그 방



This corpus has 1 document with 24,718 total words and 7,989 unique word forms. Created now.

Vocabulary Density: 0.323

Average Words Per Sentence: 10.1

Most frequent words in the corpus: 그레고르 (420); 여동생 (316); 아버지 (268); 어머니 (194); 오빠 (82)

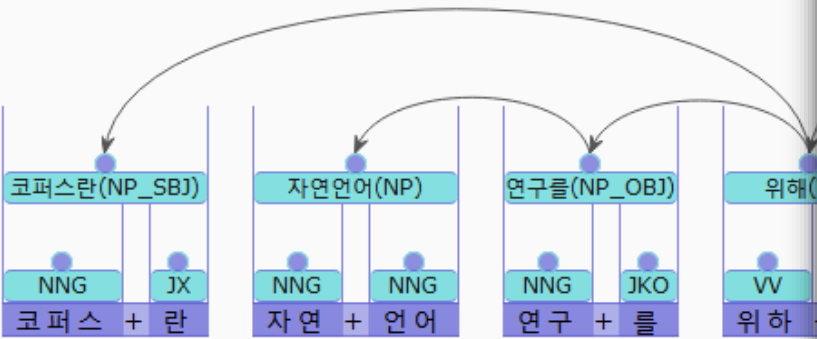
Document	Left	Term	Right
1) 변신 ...	어느 날 아침, '그레고르	잠자'가 어수선한 잠에서 깨어났을	
1) 변신 ...	방한웅 토시)가 둘러있었다. 그때	그레고르	는 창밖으로 고개를 돌렸다. 흐릿한
1) 변신 ...	틀림없어요."라며. 그리고 이렇...	그레고르	씨의 견적물 컬렉션이 아직 도착하지
1) 변신 ...	전적으로 나쁘다고 할 수 있을까?	그레고르	는 오랜 시간을 잤다. 그래도
1) 변신 ...	졌다. 그래도 밀려오는 졸음을 ...	그레고르	는 기분이 아주 괜찮았다. 심지어
1) 변신 ...	는 기분이 아주 괜찮았다. 심지어	그레고르	는 평소보다 훨씬 배고픔을 느끼고
1) 변신 ...	문 쪽에서 조심스러운 노크소리...	그레고르	야!" 누군가가 그를 부르고 있었다
1) 변신 ...	저 부드러운 목소리! 하지만 곧	그레고르	는 충격을 받았다. 그 자신이
1) 변신 ...	들었는지를 자신하지 못하게 만...	그레고르	는 최대한 많이 대답하고 싶었다

Back to the Question

- How can I analyze texts of *Samch'ŏlli* (a Korean literature magazine) to examine the usage and context of certain words and visualize the results?
 - Prepare a raw corpus of *Samch'ŏlli*
 - As of today, no ready-made corpus for *Samch'ŏlli*
 - Find a analysis tool processing user-provided corpora
 - Voyant Tools
 - Make a stop-word list
 - Pre-process the raw corpus
 - ETRI Text Analysis API for Korean for automatic word division and/or tagging (requires programming knowledge)

AI API Data Service Portal

No	Word	형태소	
		단어	태그
0	코퍼스란	코퍼스 란	NNG JX
1	자연언어	자연 언어	NNG NNG



6	추출한	추출 하 ㄴ	NNG XSV ETM	동
7	집합이다	집합 이 다	NNG VCP EF	동

```

{
  request_id: "reserved field",
  result: 0,
  return_type: "com.google.gson.internal.LinkedTreeMap",
  return_object: {
    doc_id: "",
    DCT: "",
    category: "",
    category_weight: 0,
    title: {
      text: "",
      NE: ""
    },
    metaInfo: {},
    paragraphInfo: [],
    sentence: [
      {
        id: 0,
        reserve_str: "",
        text: "코퍼스란 자연언어 연구를 위해 언어의 표본을 추출한 집합이다",
        morp: [
          {
            id: 0,
            lemma: "코퍼스",
            type: "NNG",
            position: 0,
            weight: 0
          },
          {
            id: 1,
            lemma: "란",
            type: "JX",
            position: 9,
            weight: 0.018041
          }
        ]
      }
    ]
  }
}

```

Other Corpora & Tools for Korean

- 현대 한국어 용례 검색기
2 corpora: SJ-RIKS, SJ-RIKS ext.
<http://riksdb.korea.ac.kr/>
- 연세 말뭉치 용례 검색 시스템
4 corpora: 20세기 한국어 말뭉치, 균형 말뭉치, 교육용 말뭉치, 주제별 말뭉치
<https://ilis.yonsei.ac.kr/corpus/>
- 국립국어원 언어정보 나눔터
21-segi Sejong Project Corpus (basic corpus only)
<https://ithub.korean.go.kr>