PRINCETON UNIVERSITY LIBRARY

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

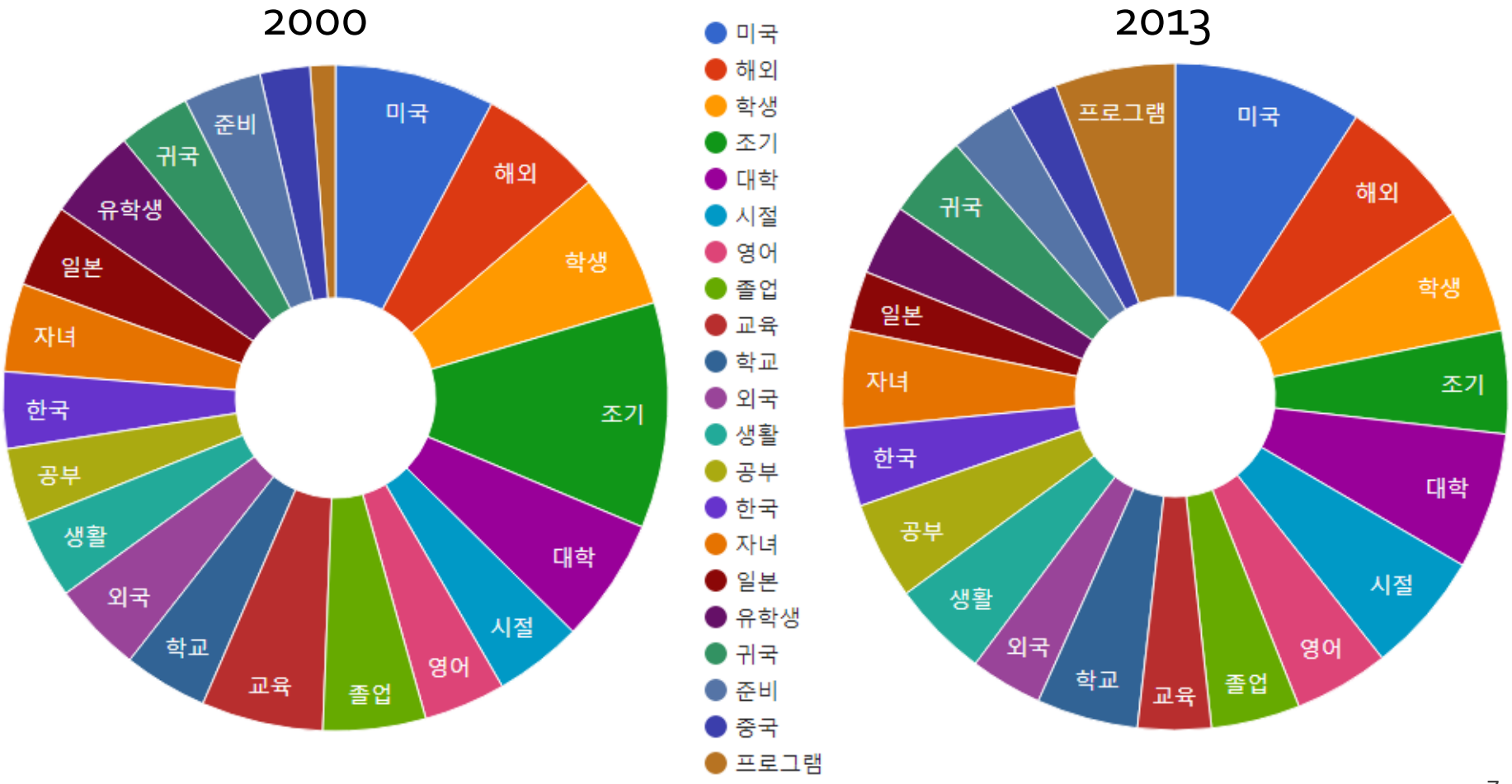
- Corpus with analysis & visualization tools developed by Korea University
  http://corpus.korea.ac.kr/
- Texts of all news articles between 2000 and 2013 from *Tonga ilbo*, *Chosŏn ilbo*, *Chungang ilbo*, *Han'gyŏre sinmun*.
- 60 million word segments (語節)
  - 한국 = 1 word 1 segment / 한국의 = 2 words 1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

Usage Search

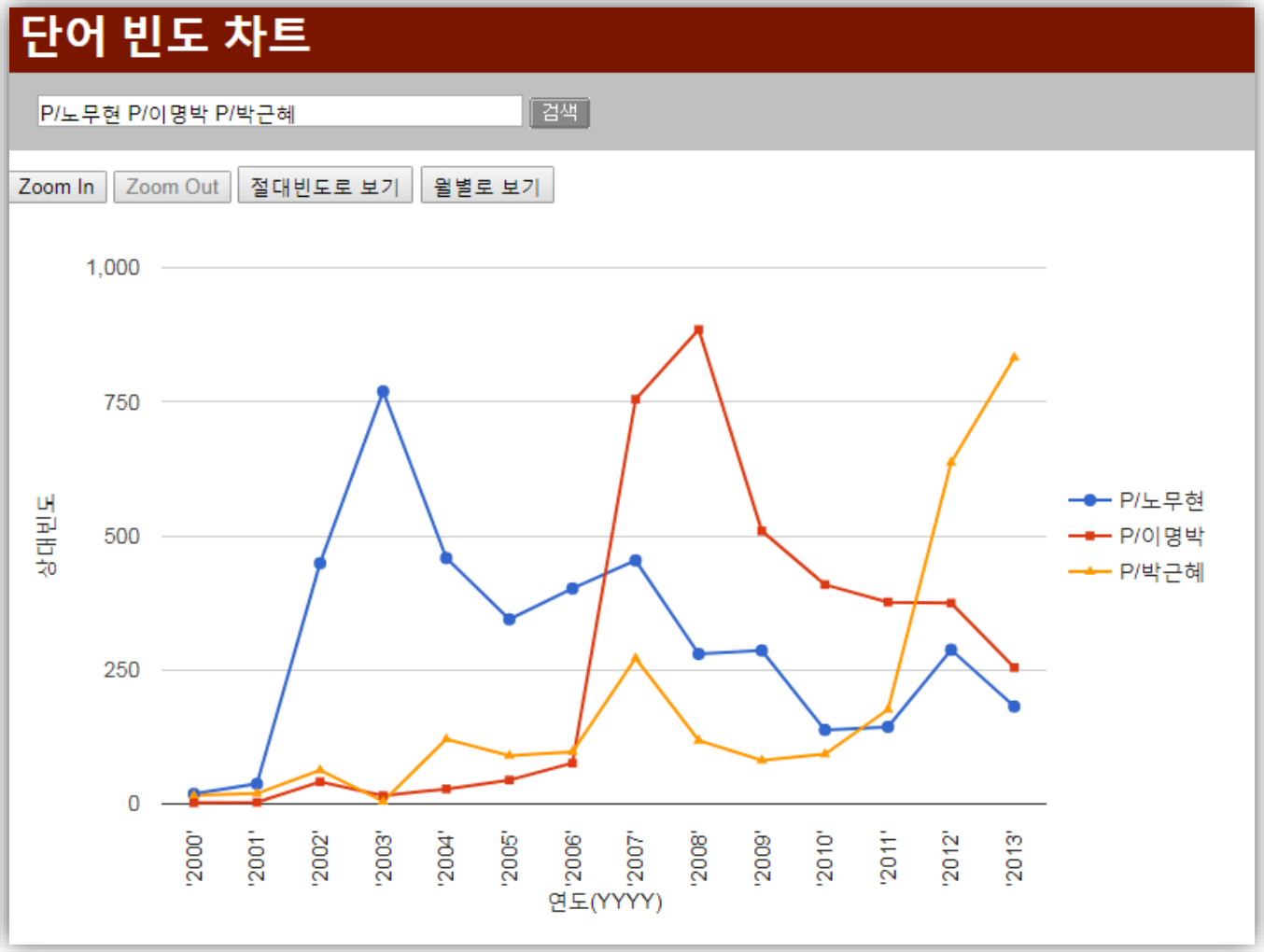
"유학 (留學)"

Co-occurring Words by Year

KW = 유학 (留學)



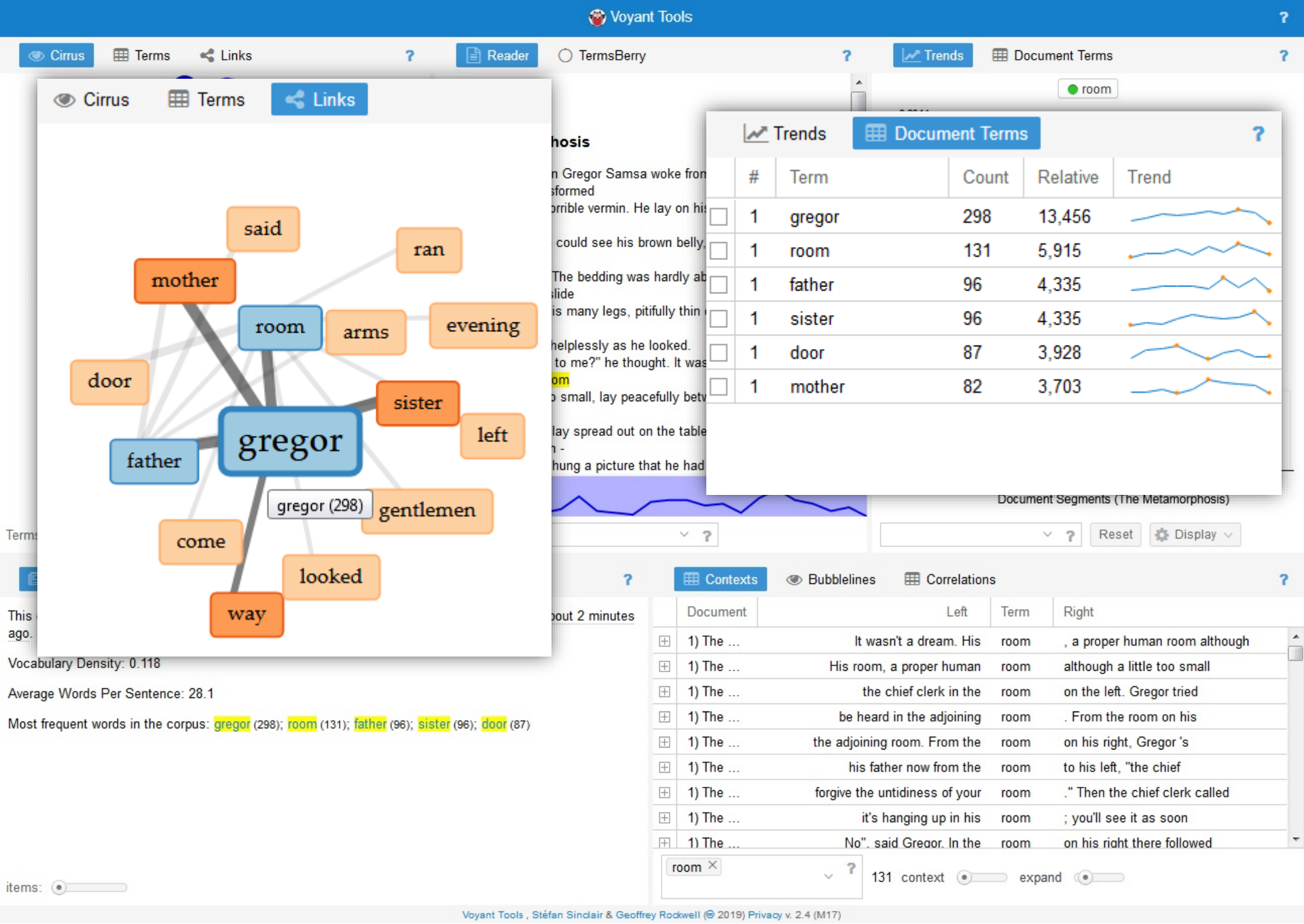
Word Frequency



Voyant Tools

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

- Analysis tools for user-provided corpus


Voyant Tools
Cirrus
Terms
Links
Reader
TermsBerry
Trends
Document Terms
said
ran
mother
room
arms
evening
door
sister
left
gregor
father
gregor (298)
gentlemen
come
looked
way
Term Count Relative Trend
1 gregor 298 13,456
1 room 131 5,915
1 father 96 4,335
1 sister 96 4,335
1 door 87 3,928
1 mother 82 3,703
Document Segments (The Metamorphosis)
Reset
Display
Contexts
Bubblelines
Correlations
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 Gregor. In the room on his right there followed
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)
room
131 context
expand
items:
Voyant Tools , Stéfan Sinclair & Geoffrey Rockwell (© 2019) Privacy v. 2.4 (M17)

Voyant Tools

Cirrus | Terms | Links

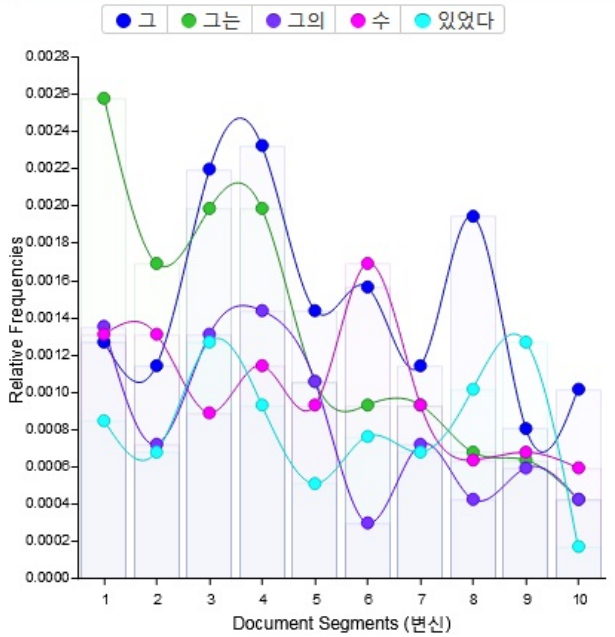
Terms:

Reader | TermsBerry

변신

어느 날 아침, '그레고르 잠자'가 어수선한 잠에서 깨어났을 때, 자신이 흉측한 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 돔형이었고, 딱딱한 마디들로
이어진 아치형구조를 이루고 있었다. 괴상했다.
그의 이불로는 그 흉측한 배를 모두 덮을 수 없이 밝혀졌다. 이불이 어느 순간 그의 배 위에서 스르륵 미끄러져 내렸기 때문이다. 처음부터 미끄러져
내릴 준비를 한 것처럼.
아, 그의 수많은 발이라니! 비참하게 얇은 발들은 그 수가 너무 많았다. 그의 몸뚱이에 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

Trends | Document Terms



Reset | Display

Summary | Documents | Phrases

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)

items:

Contexts | Bubblelines | Correlations

Document	Left	Term	Right
1) 변신	배를 볼 수 있었을 텐데.	그	배는 약간의 돔형이었고, 딱딱한 마...
1) 변신	이루고 있었다. 괴상했다. 그의 이...	그	흉측한 배를 모두 덮을 수
1) 변신	수많은 발이라니! 비참하게 얇은 발...	그	수가 너무 많았다. 그의 몸뚱이에
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 모피 머프(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	눕기로 했다. 애초 깨어났을 때	그	상태로 누웠다. "일찍 일어나는 것은

351 context | expand

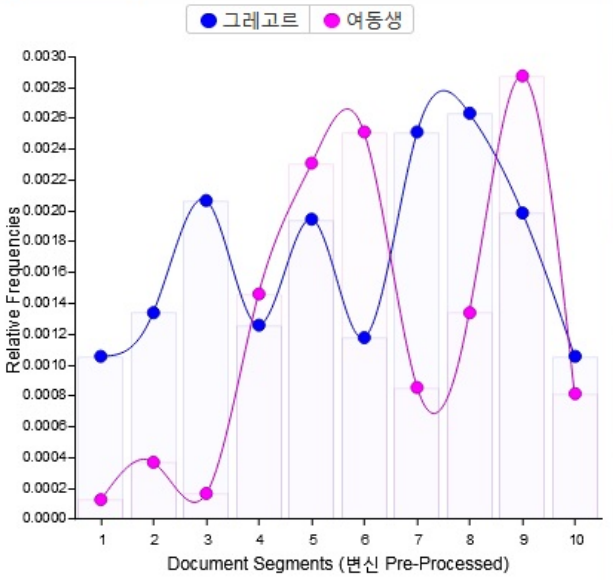
Voyant Tools

Cirrus | Terms | Links | Reader | TermsBerry | Trends | Document Terms

Terms:

변신 Pre-Processed

어느 날 아침, '그레고르 잠자'가 어수선한 잠에서 깨어났을 때, 자신이 흉측한 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 돔형이었고, 딱딱한 마디들로
이어진 아치형구조를 이루고 있었다. 괴상했다.
그의 이불로는 그 흉측한 배를 모두 덮을 수 없이 밝혀졌다. 이불이 어느 순간 그의 배 위에서 스르륵 미끄러져 내렸기 때문이다. 처음부터 미끄러져
내릴 준비를 한 것처럼.
아, 그의 수많은 발이라니! 비참하게 얇은 발들은 그 수가 너무 많았다. 그의 몸뚱이에 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



그레고르 ×
여동생 ×
Reset | Display

Summary | Documents | Phrases

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)

items:

Contexts | Bubblelines | Correlations

Document	Left	Term	Right
1) 변신 ...	어느 날 아침, '	그레고르	잠자'가 어수선한 잠에서 깨어났을
1) 변신 ...	방한용 토시)가 둘러있었다. 그때	그레고르	는 창밖으로 고개를 돌렸다. 흐릿한
1) 변신 ...	틀림없어요."라며. 그리고 이렇...	그레고르	씨의 견직물 컬렉션이 아직 도착하지
1) 변신 ...	전적으로 나쁘다고 할 수 있을까?	그레고르	는 오랜 시간을 잤다. 그래도
1) 변신 ...	잤다. 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 싶었었다

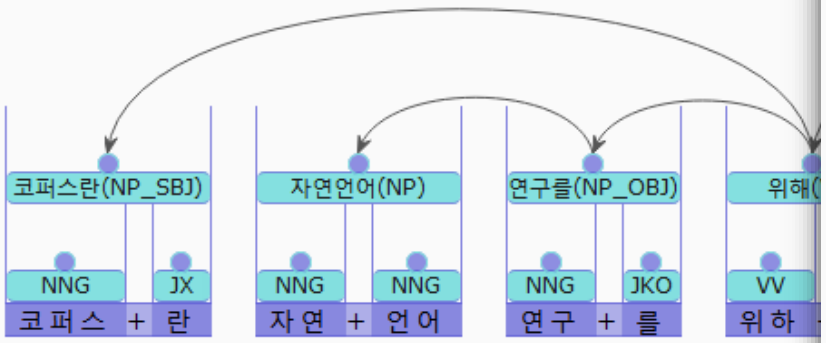
420 context | expand

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




```
{
    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