Corpus-Based Discourse Analysis

Recent Developments and Future Directions

Philipp Heinrich

Computational Corpus Linguistics Group
Friedrich-Alexander University of Erlangen-Nuremberg
http://philipp-heinrich.eu

Seoul

October 1, 2018

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Computational Linguistics

source: Nautilus (Christopher D. Manning)

Corpus Linguistics

- a **corpus**
  - is a **collection of machine-readable texts**
  - can be processed and analyzed using methods from computational linguistics
  - can be a **sample of authentic language data** and can as such be **representative for a language** (variety)

- **corpus linguistics**
  - **creation** and **processing** of corpora
  - analysis and **interpretation** of corpora

- **research questions** in corpus linguistics
  - main goal: research of language **usage**
  - empirical **testing of linguistic hypotheses**
  - language varieties and dialects
  - corpus-based grammars, psycho-linguistics, . . .

1. Introduction
   - Computational Corpus Linguistics
   - Methods in CCL

2. Corpus-Based Discourse Analysis
   - Basic Methodology
   - Case Studies
   - Extensions

3. The Future of CCL
   - Deep Learning and CCL
   - Towards a Hermeneutic Cyborg

Keywords

- **keywords** are words that occur more frequently in a text than what would be expected assuming random variation
- keywords are calculated with respect to a **reference corpus**

Keywords

- **keywords** are words that occur more frequently in a text than what would be expected assuming random variation
- keywords are calculated with respect to a **reference corpus**
- contingency table of observed frequencies for every word w:

	corpus 1	corpus 2
w	k_1	k_2
$\neg w$	$n_1 - k_1$	$n_2 - k_2$

Keywords

- **keywords** are words that occur more frequently in a text than what would be expected assuming random variation
- keywords are calculated with respect to a **reference corpus**
- contingency table of observed frequencies for every word w:

	corpus 1	corpus 2	
w	$O := O_{11}$	O_{12}	$= R_1$
$\neg w$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

Indifference (Independence)

- **association measures** (AMs) provide a quantification of the divergence of observed frequencies from their expected frequencies s. t. independence in contingency table
- **indifference table**:

	corpus 1	corpus 2	
w	$E := E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$	$= R_1$
$\neg w$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$	$= R_2$
	$= C_1$	$= C_2$	$= N$

Statistical Association Measures

	corpus 1	corpus 2	
w	O_{11} vs. E_{11}	O_{12} vs. E_{12}	R_1
$\neg w$	O_{21} vs. E_{21}	O_{22} vs. E_{22}	R_2
	$= C_1$	$= C_2$	$= N$

- t-score $= \frac{O-E}{\sqrt{O}}$
- $LL = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$
- $\chi^2 = \sum_{ij} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
- $PoiL = e^{-E_{11}} \frac{E_{11}^{O_11}}{O_{11}!}$
- Fisher $= \sum_{k=O_{11}}^{\min\{R_1,C_1\}} \frac{\binom{C_1}{k} \cdot \binom{C_2}{R_1-k}}{\binom{N}{R_1}}$
- ...

Collocations

- **distributional hypothesis** (Firth, 1957):
  - "you shall know a word by the company it keeps"
  - "one of the meanings of night is its collocability with dark, and of dark, of course, its collocation with night"

- collocations are based on observed **co-occurrence frequencies of word pairs** (w_1, w_2):

	w_2	$\neg w_2$	
w_1	O_{11}	O_{12}	$= R_1$
$\neg w_1$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

Collocations

- **distributional hypothesis** (Firth, 1957):
  - “you shall know a word by the company it keeps”
  - “one of the meanings of night is its collocability with dark, and of dark, of course, its collocation with night”

- collocations are based on observed **co-occurrence frequencies of word pairs** (w_1, w_2):

	w_2	$\neg w_2$	
w_1	O_{11}	O_{12}	$= R_1$
$\neg w_1$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

- different types of co-occurrence

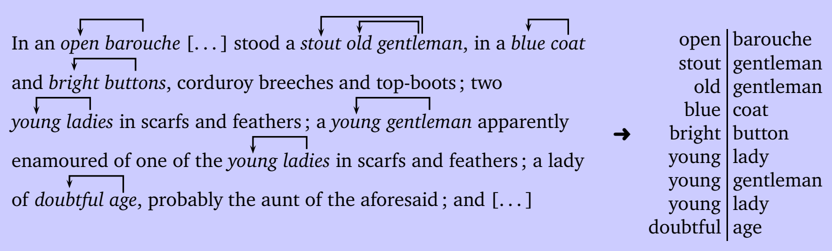
Surface Co-occurrence

A vast deal of coolness and a peculiar degree of judgement, are ⌊requisite in catching a **hat**⌋. A man must not be precipitate, or he runs over it ; he must not rush into the opposite extreme, or he loses it altogether. [. . .] There was a fine gentle ⌊wind, and Mr. Pickwick's **hat** *rolled* sportively before it⌋. The wind puffed, and Mr. ⌊Pickwick puffed, and the **hat** *rolled* over and over⌋ as merrily as a lively porpoise in a strong tide ; and on it might have *rolled*, far beyond Mr. Pickwick's reach, had not its course been providentially stopped, just as that gentleman was on the point of resigning it to its fate.

Textual Co-occurrence

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a hat.	hat	—
A man must not be precipitate, or he runs *over* it ;	—	over
he must not rush into the opposite extreme, or he loses it altogether.	—	—
There was a fine gentle wind, and Mr. Pickwick's hat rolled sportively before it.	hat	—
The wind puffed, and Mr. Pickwick puffed, and the hat rolled *over* and *over* as merrily as a lively porpoise in a strong tide ;	hat	over

Syntactic Co-occurrence



In an *open barouche* [. . .] stood a *stout old gentleman,* in a *blue coat* and *bright buttons,* corduroy breeches and top-boots ; two *young ladies* in scarfs and feathers ; a *young gentleman* apparently enamoured of one of the *young ladies* in scarfs and feathers ; a lady of *doubtful age,* probably the aunt of the aforesaid ; and [. . .]

➜

open	barouche
stout	gentleman
old	gentleman
blue	coat
bright	button
young	lady
young	gentleman
young	lady
doubtful	age

Collocates of *bucket* (noun)

noun	*f*
water	183
spade	31
plastic	36
slop	14
size	41
mop	16
record	38
bucket	18
ice	22
seat	20
coal	16
density	11
brigade	10
algorithm	9
shovel	7
container	10
oats	7
sand	12
Rhino	7
champagne	10

verb	*f*
throw	36
fill	29
randomize	9
empty	14
tip	10
kick	12
hold	31
carry	26
put	36
chuck	7
weep	7
pour	9
douse	4
fetch	7
store	7
drop	9
pick	11
use	31
tire	3
rinse	3

adjective	*f*
large	37
single-record	5
cold	13
galvanized	4
ten-record	3
full	20
empty	9
steaming	4
full-track	2
multi-record	2
small	21
leaky	3
bottomless	3
galvanised	3
iced	3
clean	7
wooden	6
old	19
ice-cold	2
anti-sweat	1

1. Introduction
    - Computational Corpus Linguistics
    - Methods in CCL
2. Corpus-Based Discourse Analysis
    - Basic Methodology
    - Case Studies
    - Extensions
3. The Future of CCL
    - Deep Learning and CCL
    - Towards a Hermeneutic Cyborg

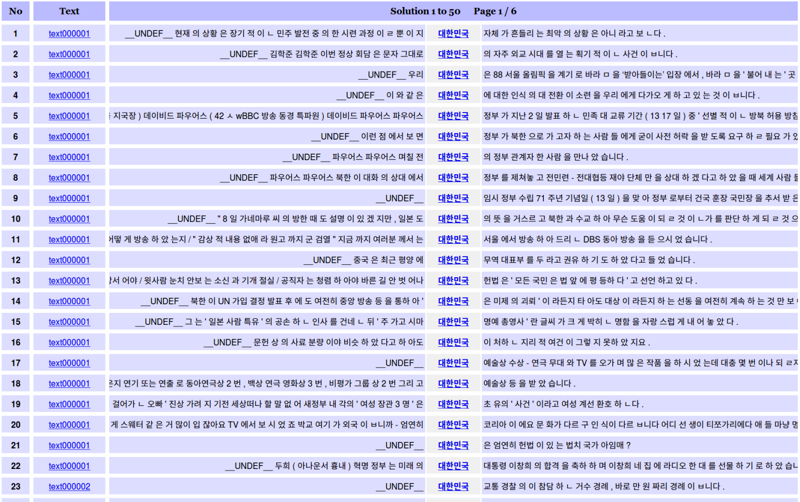
From Text to Discourse

- Foucault (1969): discourses as **statements in conversation**
- interpretation of text means categorizing
  - utterances
  - sentences
  - paragraphs
  - tweets

From Text to Discourse

- Foucault (1969): discourses as **statements in conversation**
- interpretation of text means categorizing
  - utterances
  - sentences
  - paragraphs
  - tweets
- the categories
  - are not known *a priori*
  - must be made up *on the fly* by the hermeneutic interpreter
- CDA is fundamentally different from (statistical) text classification
- ultimate goal of critical discourse analysis: *discover* **what** is said by **whom** (power relations)

Concordances



No	Text	Solution 1 to 50 Page 1 / 6		
1	text000001	__UNDEF__ 현재 의 상황 은 장기 적 이 ㄴ 민주 발전 중 의 한 시련 과정 이 ㄹ 뿐 이 지	대한민국	자체 가 흔들리 는 최악 의 상황 은 아니 라고 보 ㄴ다 .
2	text000001	__UNDEF__ 김학준 김학준 이번 정상 회담 은 문자 그대로	대한민국	의 자주 외교 시대 를 열 는 획기 적 이 ㄴ 사건 이 ㅂ니다 .
3	text000001	__UNDEF__ 우리	대한민국	은 88 서울 올림픽 을 계기 로 바라 ㅁ 을 '받아들이는' 입장 에서 , 바라 ㅁ 을 ' 불어 내 는 ' 곳
4	text000001	__UNDEF__ 이 와 같 은	대한민국	에 대한 인식 의 대 전환 이 소련 을 우리 에게 다가오 게 하 고 있 는 것 이 ㅂ니다 .
5	text000001	지국장) 데이비드 파우어스 (42 ㅅ wBBC 방송 동경 특파원) 데이비드 파우어스 파우어스	대한민국	정부 가 지난 2 일 발표 하 ㄴ 민족 대 교류 기간 (13 17 일) 중 ' 선별 적 이 ㄴ 방북 허용 방
6	text000001	__UNDEF__ 이런 점 에서 보 면	대한민국	정부 가 북한 으로 가 고자 하 는 사람 들 에게 굳이 사전 허락 을 받 도록 요구 하 ㄹ 필요 가
7	text000001	__UNDEF__ 파우어스 파우어스 며칠 전	대한민국	의 정부 관계자 한 사람 을 만나 았 습니다 .
8	text000001	__UNDEF__ 파우어스 파우어스 북한 이 대화 의 상대 에서	대한민국	정부 를 제쳐놓 고 전민련 - 전대협등 재야 단체 만 을 상대 하 겠 다고 하 았 을 때 세계 사람
9	text000001	__UNDEF__	대한민국	임시 정부 수립 71 주년 기념일 (13 일) 을 맞 아 정부 로부터 건국 훈장 국민장 을 추서 받
10	text000001	__UNDEF__ " 8 일 가네마루 씨 의 방한 때 도 설명 이 있 겠 지만 , 일본 도	대한민국	의 뜻 을 거스르 고 북한 과 수교 하 아 무슨 도움 이 되 ㄹ 것 이 ㄴ가 를 판단 하 게 되 ㄹ 것
11	text000001	떻 게 방송 하 았 는지 / " 감상 적 내용 없애 라 원고 까지 군 검열 " 지금 까지 여러분 께서 는	대한민국	서울 에서 방송 하 아 드리 ㄴ DBS 동아 방송 을 듣 으시 었 습니다 .
12	text000001	__UNDEF__ 중국 은 최근 평양 에	대한민국	무역 대표부 를 두 라고 권유 하 기 도 하 았 다고 들 었 습니다 .
13	text000001	서 어야 / 윗사람 눈치 안보 는 소신 과 기개 절실 / 공직자 는 청렴 하 아야 바른 길 안 벗 어나	대한민국	헌법 은 ' 모든 국민 은 법 앞 에 평 등하 다 ' 고 선언 하고 있 다 .
14	text000001	__UNDEF__ 북한 이 UN 가입 결정 발표 후 에 도 여전히 중앙 방송 등 을 통하 아 '	대한민국	은 미제 의 괴뢰 ' 이 라든지 타 아도 대상 이 라든지 하 는 선동 을 여전히 계속 하 는 것 만 보
15	text000001	__UNDEF__ 그 는 ' 일본 사람 특유 ' 의 공손 하 ㄴ 인사 를 건네 ㄴ 뒤 ' 주 가고 시마	대한민국	명예 총영사 ' 란 글씨 가 크 게 박히 ㄴ 명함 을 자랑 스럽 게 내 어 놓 았 다 .
16	text000001	__UNDEF__ 문헌 상 의 사료 분량 이야 비슷 하 았 다고 하 아도	대한민국	이 처하 ㄴ 지리 적 여건 이 그렇 지 못하 았 지요 .
17	text000001	__UNDEF__	대한민국	예술상 수상 - 연극 무대 와 TV 를 오가 며 많 은 작품 을 하 시 었 는데 대충 몇 번 이나 되 ㄹ
18	text000001	지 연기 또는 연출 로 동아연극상 2 번 , 백상 연극 영화상 3 번 , 비평가 그룹 상 2 번 그리 고	대한민국	예술상 등 을 받 았 습니다 .
19	text000001	걸어가 ㄴ 오빠 ' 진상 가려 지 기전 세상떠나 할 말 없 어 새정부 내 각의 ' 여성 장관 3 명 ' 은	대한민국	초 유의 ' 사건 ' 이라고 여성 계선 환호 하 ㄴ다 .
20	text000001	게 스웨터 같 은 거 많이 입 잖아요 TV 에서 보 시 었 죠 박교 여기 가 외국 이 ㅂ니까 - 엄연히	대한민국	코리아 이 에요 문 화가 다르 구 인 식이 다르 ㅂ니다 어디 선 생이 티쪼가리에다 애 들 마냥
21	text000001	__UNDEF__	대한민국	은 엄연히 헌법 이 있 는 법치 국가 아임매 ?
22	text000001	__UNDEF__ 두희 (아나운서 흉내) 혁명 정부 는 미래 의	대한민국	대통령 이창희 의 합격 을 축하 하 며 이창희 네 집 에 라디오 한 대 를 선물 하 기 로 하 았 습
23	text000002	__UNDEF__	대한민국	교통 경찰 의 이 참담 하 ㄴ 거수 경례 , 바로 만 원 짜리 경례 이 ㅂ니다 .

Corpus-Based Discourse Analysis (CDA)

- CDA means analyzing and deconstructing concordance lines
  - concordances are the essence of discourses
- finding **discourses**: **nodes** + **attitudes**
  - (topic) nodes can be defined by *keywords* or (more generally) *corpus queries*
  - attitudes: *collocates* that are retrieved by statistical methods
- examples
  - "refugees as victims" (Baker, 2006)
  - "Fukushima as worst case scenario"

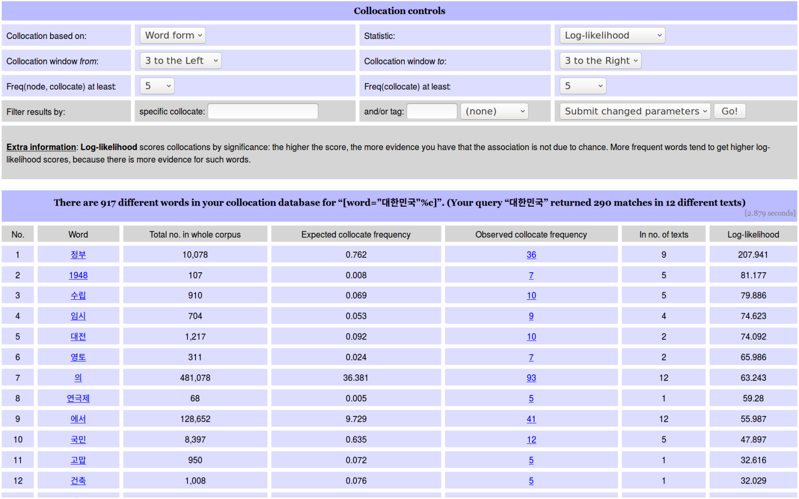
Corpus-Based Discourse Analysis (CDA)

- CDA means analyzing and deconstructing concordance lines
  - concordances are the essence of discourses
- finding **discourses**: **nodes + attitudes**
  - (topic) nodes can be defined by *keywords* or (more generally) *corpus queries*
  - attitudes: *collocates* that are retrieved by statistical methods
- examples
  - "refugees as victims" (Baker, 2006)
  - "Fukushima as worst case scenario"

in practice:

- look at (*n* best) collocates of topic node
- categorize into on-the-fly-groups

Collocations



Collocation controls

Collocation based on:	Word form	Statistic:	Log-likelihood
Collocation window *from*:	3 to the Left	Collocation window *to*:	3 to the Right
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5
Filter results by:	specific collocate:	and/or tag: (none)	Submit changed parameters Go!

Extra information: **Log-likelihood** scores collocations by significance: the higher the score, the more evidence you have that the association is not due to chance. More frequent words tend to get higher log-likelihood scores, because there is more evidence for such words.

There are 917 different words in your collocation database for “[word="대한민국"%c]”. (Your query “대한민국” returned 290 matches in 12 different texts)

[2.879 seconds]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood
1	정부	10,078	0.762	36	9	207.941
2	1948	107	0.008	7	5	81.177
3	수립	910	0.069	10	5	79.886
4	임시	704	0.053	9	4	74.623
5	대전	1,217	0.092	10	2	74.092
6	영토	311	0.024	7	2	65.986
7	의	481,078	36.381	93	12	63.243
8	연극제	68	0.005	5	1	59.28
9	에서	128,652	9.729	41	12	55.987
10	국민	8,397	0.635	12	5	47.897
11	고압	950	0.072	5	1	32.616
12	건축	1,008	0.076	5	1	32.029

Case Studies

- **refugees** (KhosraviNik, 2010)
  - "The representation of refugees, asylum seekers and immigrants in British newspapers: a critical discourse analysis."
  - CDA investigation on discursive strategies employed by various British newspapers between 1996-2006 in the ways they represent refugees, asylum seekers and immigrants.

Case Studies

- **refugees** (KhosraviNik, 2010)
  - "The representation of refugees, asylum seekers and immigrants in British newspapers: a critical discourse analysis."
  - CDA investigation on discursive strategies employed by various British newspapers between 1996-2006 in the ways they represent refugees, asylum seekers and immigrants.
- **gender** (Baker, 2014)
  - "Using Corpora to Analyze Gender"
  - collection of case studies wrt. changes in sexist and non-sexist language use over time, personal adverts, press representation of gay men, and the ways that boys and girls are constructed through language
- **LGBT** (Love and Baker, 2015)
  - "The hate that dare not speak its name?"
  - How have the British Parliamentary arguments against LGBT equality changed in response to decreasing social acceptability of discriminatory language against minority groups?

PhD project *Exploring the Fukushima Effect*

- identification and analysis of the tempo-spatial propagation of **discourses** in the **transnational algorithmic public sphere**
- case study: Fukushima Effect (cf. Gono'i, 2015)
  - attitudes and opinions towards energy sources
- data: mass and social media (German, Japanese)
  - intra- and transmedial and -national
  - "edited mass communication" vs. "mass self-communication"
- further information:
  - `www.linguistik.fau.de/projects/efe/`
  - funded by the **Emerging Fields Initiative** of FAU
  - Team:
    - ★ Chair of Computational Corpus Linguistics
    - ★ Chair of Japanese Studies
    - ★ Chair of Communication Science
    - ★ Chair of Visual Computing

Corpora – Social Media (Twitter)

German Twitter

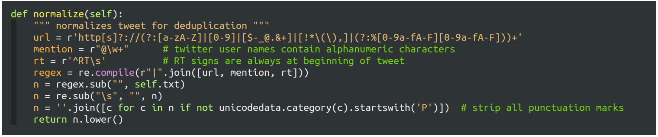
- 10,266,835 original posts
- linguistic annotation:
  - tokenization: SoMaJo (Proisl and Uhrig, 2016)
  - POS-tagging: SoMeWeTa (Proisl, 2018)
  - lemmatization: work in progress

Japanese Twitter

- 411,452,027 original posts
- linguistic annotation:
  - special dictionary: ipadic-neologd (Sato et al., 2017)

Identification of Social Bots (Schäfer et al., 2017)

1. normalization of texts



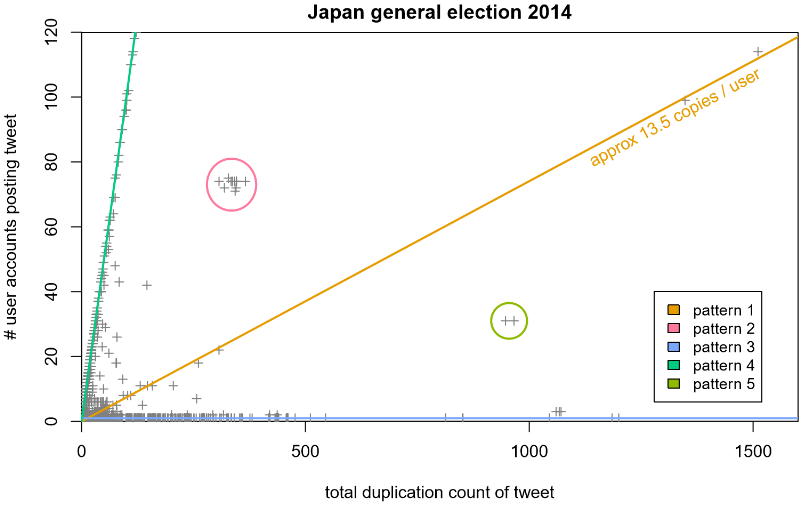
```
def normalize(self):
    """ normalizes tweet for deduplication """
    url = r'http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\(\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+'
    mention = r"@\w+"      # twitter user names contain alphanumeric characters
    rt = r'^RT\s'          # RT signs are always at beginning of tweet
    regex = re.compile(r"|".join([url, mention, rt]))
    n = regex.sub("", self.txt)
    n = re.sub("\s", "", n)
    n = ''.join([c for c in n if not unicodedata.category(c).startswith('P')])  # strip all punctuation marks
    return n.lower()
```

2. mapping of normalized strings onto tweet ids
3. extension: hierarchical clustering based on Levenshtein distance

Footprint of a Social Bot net

$$\frac{\text{number of near duplicates}}{\text{number of user accounts}}$$

Identification of Social Bots during the Japanese General Election of 2014

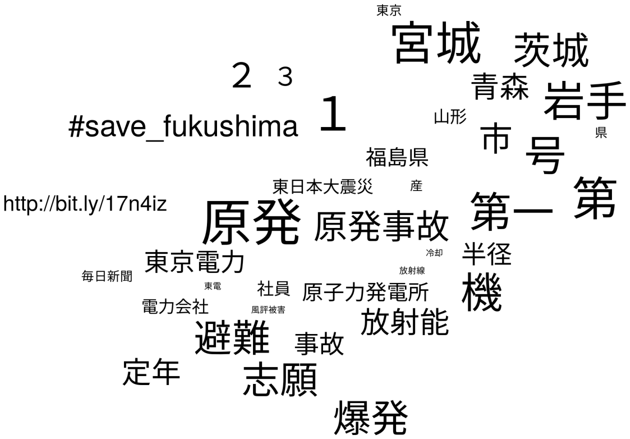


Visualization

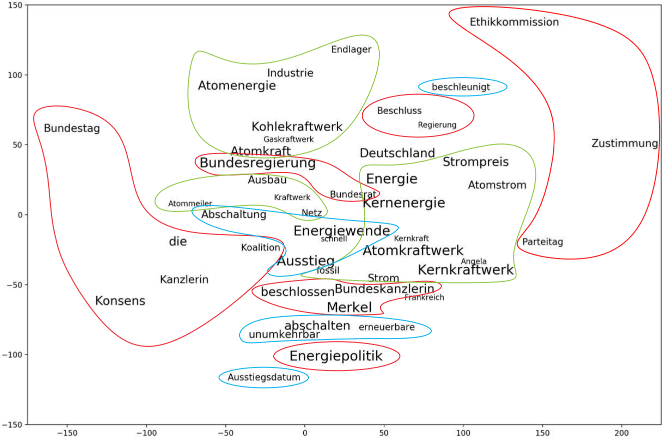
- high-dimensional word embeddings (Word2Vec) (Mikolov et al., 2013)
  - based on shallow, two-layer neural networks
  - capturing co-occurrence information of words in 50–1000 dimensions
- t-distributed stochastic neighbour-embedding (t-SNE) (van der Maaten and Hinton, 2008)
  - project high-dimensional embeddings onto two-dimensional plane
  - semantically similar items are pre-grouped together
- size of lexical items represents association strength towards (topic) node (Evert, 2008)
  - different AMs retrieve different sets of collocates and sizes
- see Heinrich et al. (2018); Heinrich and Schäfer (2018)

Visualizing Collocational Profiles (node: *Fukushima*)

2011.03.12 – 2011.03.19 node: 5121.9 tw.p.m (29425/5744937)



Visualizing Collocational Profiles (node: *Nuclear Phase-Out*)



Higher-Order Collocates

1. discourse collocates
   - straightforward generalization with respect to textual co-occurrence
   - look at co-occurrence frequencies of tweets that were identified to be part of the discourse at hand (topic + attitude)
   - collocates represent lexical items that are particularly important for the **discourse**

Higher-Order Collocates

1. discourse collocates
   - straightforward generalization with respect to textual co-occurrence
   - look at co-occurrence frequencies of tweets that were identified to be part of the discourse at hand (topic + attitude)
   - collocates represent lexical items that are particularly important for the **discourse**

2. second-order topic-collocates
   - look at co-occurrence frequencies of one set of lexical items c in tweets that are about a certain topic t
   - for all w: compare co-occurrence frequencies of w with c among tweets that contain t with marginal frequencies of w in all tweets that contain t
   - collocates of c that are particularly important for the **topic** t

Second-Order Collocates

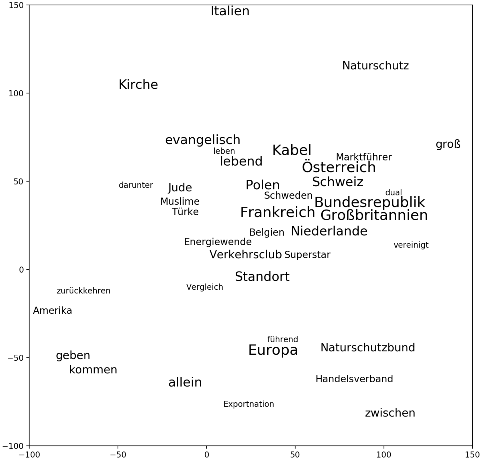


Figure: Paragraph-collocates of *Germany* in the FAZ corpus.

Second-Order Collocates

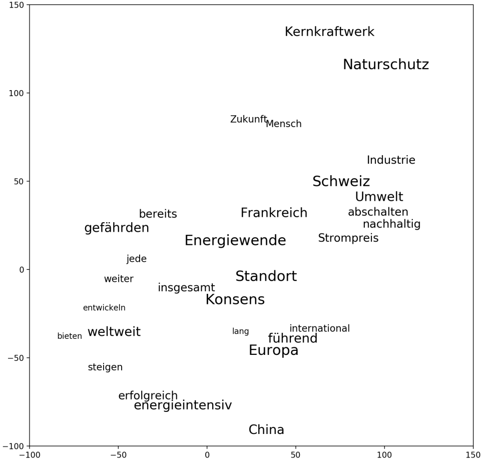


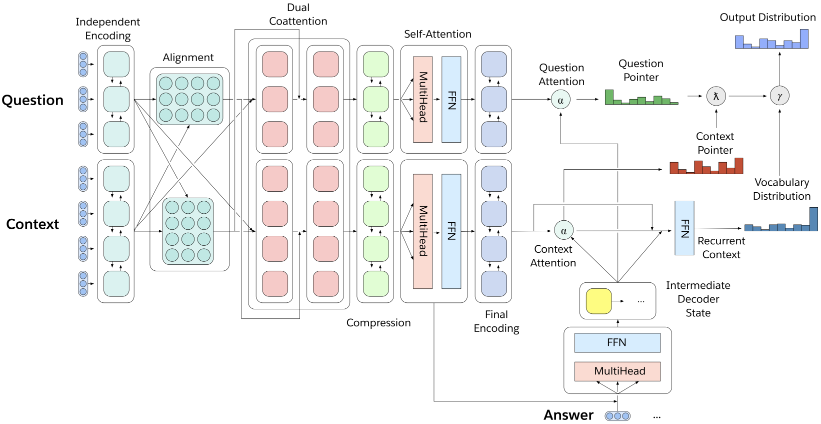
Figure: Collocates of *Germany* in energy-transition paragraphs.

1. Introduction
    - Computational Corpus Linguistics
    - Methods in CCL
2. Corpus-Based Discourse Analysis
    - Basic Methodology
    - Case Studies
    - Extensions
3. The Future of CCL
    - Deep Learning and CCL
    - Towards a Hermeneutic Cyborg

Deep Learning and AI

- artificial neural networks
  - general end-to-end ML algorithms
  - origins in 1950s
  - recent hype due to improvements in processing power
- amazing performance in
  - visual object recognition
  - OCR
  - text categorization
  - machine translation
  - strategic games (Go)
  - simulating humans (Google assistant)

Deep Learning

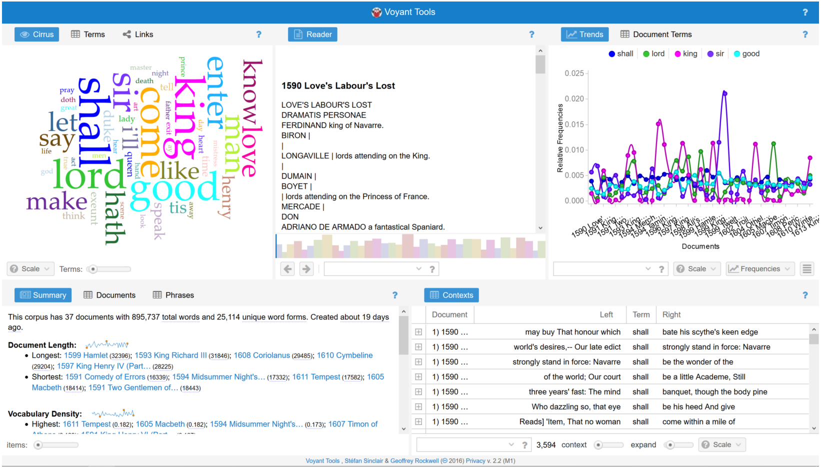


source: einstein.ai

Will human input become irrelevant?

- standard toolbox of corpus linguistics:
  - concordancing
  - frequencies and frequency comparison
  - collocations
- these techniques have been around for 50 years!
- AI techniques outperform humans when it comes to real-world applications
  - even the creation of gold-standard data (manual annotation) becomes less and less important
  - why bother with rule-based systems?

Digital Humanities



1590 Love's Labour's Lost

LOVE'S LABOUR'S LOST
DRAMATIS PERSONAE
FERDINAND king of Navarre.
BIRON |
|
LONGAVILLE | lords attending on the King.
|
DUMAIN |
BOYET |
| lords attending on the Princess of France.
MERCADE |
DON
ADRIANO DE ARMADO a fantastical Spaniard.

This corpus has 37 documents with 895,737 total words and 25,114 unique word forms. Created about 19 days ago.

Document Length:

- Longest: 1599 Hamlet (32396); 1593 King Richard III (31846); 1608 Coriolanus (29485); 1610 Cymbeline (29204); 1597 King Henry IV (Part… (28225)
- Shortest: 1591 Comedy of Errors (16339); 1594 Midsummer Night's… (17332); 1611 Tempest (17582); 1605 Macbeth (18414); 1591 Two Gentlemen of… (18443)

Vocabulary Density:

- Highest: 1611 Tempest (0.182); 1605 Macbeth (0.182); 1594 Midsummer Night's… (0.173); 1607 Timon of

Document	Left	Term	Right
1) 1590 …	may buy That honour which	shall	bate his scythe's keen edge
1) 1590 …	world's desires,-- Our late edict	shall	strongly stand in force: Navarre
1) 1590 …	strongly stand in force: Navarre	shall	be the wonder of the
1) 1590 …	of the world; Our court	shall	be a little Academe, Still
1) 1590 …	three years' fast: The mind	shall	banquet, though the body pine
1) 1590 …	Who dazzling so, that eye	shall	be his heed And give
1) 1590 …	Reads] 'Item, That no woman	shall	come within a mile of

3,594

Voyant Tools , Stéfan Sinclair & Geoffrey Rockwell (© 2016) Privacy v. 2.2 (M1)

source: Voyant Tools

Towards a Hermeneutic Cyborg

1. interoperability
   - query tool → quantitative data → visualization
   - exchange quantitative results and manual grouping across systems

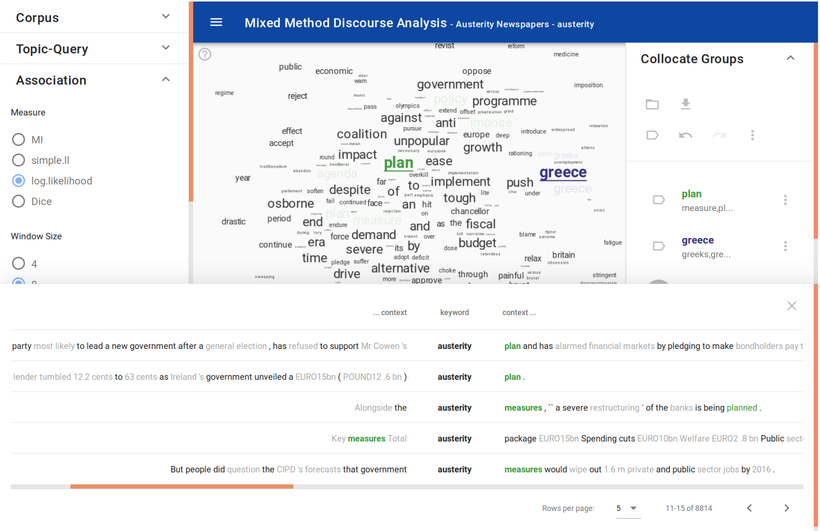
Towards a Hermeneutic Cyborg

1. interoperability
   - query tool → quantitative data → visualization
   - exchange quantitative results and manual grouping across systems
2. interactivity
   - integrate larger part of workflow into corpus software
   - maintain connection to concordances
   - implement visualization components in analysis tools

Towards a Hermeneutic Cyborg

1. interoperability
   - query tool → quantitative data → visualization
   - exchange quantitative results and manual grouping across systems
2. interactivity
   - integrate larger part of workflow into corpus software
   - maintain connection to concordances
   - implement visualization components in analysis tools
3. integration
   - key challenge: how to feed back information from manual grouping into quantitative procedures?
   - applied to CDA: how to update discourse embeddings?

Mixed-Methods Discourse Analysis



Measure

○ MI

○ simple.ll

● log.likelihood

○ Dice

Window Size

○ 4

● 8

plan
measure,pl...

greece
greeks,gre...

... context	keyword	context ...
party most likely to lead a new government after a general election , has refused to support Mr Cowen 's	austerity	plan and has alarmed financial markets by pledging to make bondholders pay t
lender tumbled 12.2 cents to 63 cents as Ireland 's government unveiled a EURO15bn (POUND12 .6 bn)	austerity	plan .
Alongside the	austerity	measures , `` a severe restructuring '' of the banks is being planned .
Key measures Total	austerity	package EURO15bn Spending cuts EURO10bn Welfare EURO2 .8 bn Public sect
But people did question the CIPD 's forecasts that government	austerity	measures would wipe out 1.6 m private and public sector jobs by 2016 .

Rows per page: 5

11-15 of 8814

Thanks for listening.

Questions?

P. Baker. *Using Corpora to Analyze Gender*. Bloomsbury Publishing, 2014.

Paul Baker. *Using Corpora in Discourse Analysis*. Continuum, London, 2006.

Stefan Evert. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, chapter 58. Mouton de Gruyter, Berlin, 2008.

J.R. Firth. *Papers in linguistics, 1934-1951*. Oxford University Press, 1957.

Michel Foucault. *L'Archéologie du savoir*. Éditions Gallimard, Paris, 1969.

Ikuo Gono'i. 2015-nen ANPO, Minshushugi wo futatabi hajimeru wakamono-tachi (ANPO in 2015. The Youth that is restarting Democracy), 2015.

Philipp Heinrich and Fabian Schäfer. Extending corpus-based discourse analysis for exploring japanese social media. In *Proceedings of the Asia Pacific Corpus Linguistics Conference 2018*, 2018.

Philipp Heinrich, Christoph Adrian, Olena Kalashnikova, Fabian Schäfer, and Stefan Evert. A Transnational Analysis of News and Tweets about Nuclear Phase-Out in the Aftermath of the Fukushima Incident. In Andreas Witt, Jana Diesner, and Georg Rehm, editors, *Proceedings of the LREC 2018 "Workshop on Computational Impact Detection from Text Data"*, Paris, 2018. ELRA.

Majid KhosraviNik. The representation of refugees, asylum seekers and immigrants in british newspapers : a critical discourse analysis. *Journal of Language and Politics*, 9(1):1–28, 2010.

Robbie Love and Paul Baker. The hate that dare not speak its name? *Journal of Language Aggression and Conflict*, 3(1):57–86, October 2015.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

Thomas Proisl. SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*, 2018.

Thomas Proisl and Peter Uhrig. SoMaJo: State-of-the-art tokenization for German web and social media texts. In Paul Cook, Stefan Evert, Roland Schäfer, and Egon Stemle, editors, *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin, 2016. Association for Computational Linguistics.

Toshinori Sato, Taiichi Hashimoto, and Manabu Okumura. Implementation of a word segmentation dictionary called mecab-ipadic-neologd and study on how to use it effectively for information retrieval (in japanese). In *Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing*, pages NLP2017–B6–1. The Association for Natural Language Processing, 2017.

Fabian Schäfer, Stefan Evert, and Philipp Heinrich. Japan's 2014 General Election: Political Bots, Right-Wing Internet Activism and PM Abe Shinzō's Hidden Nationalist Agenda. *Big Data*, 5:1 – 16, 2017.

L.J.P van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.