

# Shifting discourse-semantics of risk in US newspapers, 1987–2014

Daniel McDonald   Jens Zinn

@interro\_gator

This slideshow is available at: <http://git.io/vBfbw>

23rd November 2015

- Context of the investigation: risk theory
- Data and research questions
- Linguistic approaches to risk
- Our methods and linguistic findings
- Sociological significance of the results

This slideshow is available at: <http://git.io/vBfbw>

I'm presenting work from closely related projects:

- ① Risk words in the NYT, 1963, 1987–2014
- ② Risk words in NYT health articles
- ③ Risk words in six US newspapers, 1987–2014

All investigations involve making longitudinally structured, parsed corpora and looking at how risk words behave.

Risk as concept is sociologically important:

- New global risks (Beck, 1992)
- Calculative technologies (Dean, 1999)
- Individualisation—from tradition to decision (Beck, 1992)
- Technologies of the Self (Dean, 1998)
- Risk-taking (Luhmann, 1993)

- Risk can be nominal, verbal, adjectival, adverbial
- Risk as lexical item is increasingly frequent in print journalism (Zinn 2011)
- Risk as a lexical item in naturalistic text may behave contrary to expectations (Hamilton, Adolphs, & Nerlich, 2007)
- Meaning of risk moves toward *threat/danger*

Risk as participant is more closely related to negative outcomes than risk as process:

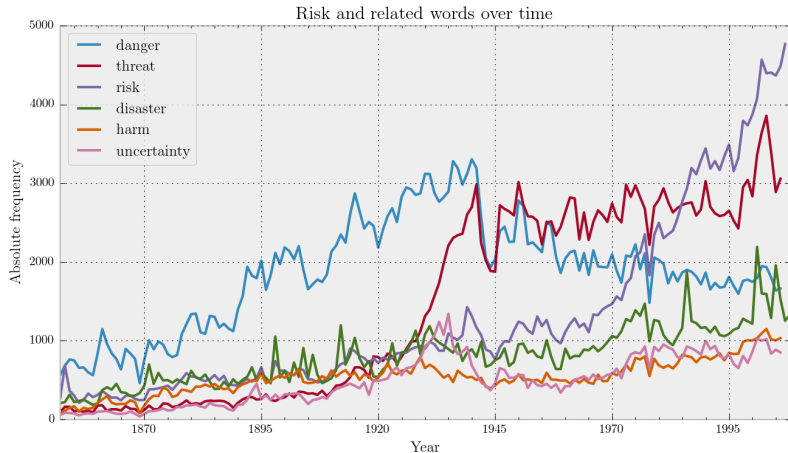
- Process: *“Only those who will risk going too far can possibly find out how far one can go”*
- Participant: *risks/rewards, risk-to-benefit-ratio*

New kinds of data and tools make it possible to empirically analyse risk language in new ways:

- Digitisation of newspapers means we have large, well-structured datasets
- Parsing makes it possible to search for lexical and grammatical features in tandem
- Programmatic approaches to social science research facilitate:
  - ▶ Automation
  - ▶ Reproducibility
  - ▶ Transparency
  - ▶ Ability to deploy methods on new data
  - ▶ *Objectivity?*

- ① *NYT Annotated Corpus*: 1.8 million articles, 1987–2007 (Sandhaus, 2008)
- ② *ProQuest Newsstand* for NYT 2007–2014
- ③ *ProQuest Newsstand* for five other newspapers, 1987–2014
  - ① *Washington Post*
  - ② *Tampa Bay Times*
  - ③ *USA Today*
  - ④ *Chicago Tribune*
  - ⑤ *Wall Street Journal*
- ④ 54,288,152 words
- ⑤ 1,031,208 risk words
- ⑥ 43GB when parsed

# Increasing frequency of *risk* lemma





We span the sociological, linguistic and computational. Examples:

- ① How do risk words behave longitudinally at the lexicogrammatical and discourse-semantic strata?
- ② What kinds of tools or methods are needed for this kind of (digital humanities) research?
- ③ How does the institutionalisation of new societal practices manifest linguistically in the change of risk discourses and the use of risk language?

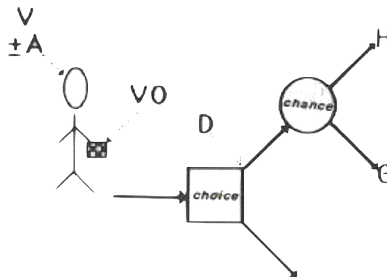
Frame semantics: risk as a cognitive schema (Fillmore & Atkins, 1992)

- Conceptualises risk mostly as experiential Process/Event
  - ▶ *What kind of participants and circumstances occur when risk is the Process?*
- Problem: risk often takes less prominent experiential roles
  - ▶ Is the risk frame actually invoked when the word is used?
  - ▶ Example:

*Mr. Tepfer noted that Mr. Douglas, who was in the neighborhood when the body was found and was interviewed by the police at the time, 'preyed on at-risk women, on prostitutes, and he engaged in sex and strangled them to death.'*

# The risk frame (Fillmore & Atkins)

H = Harm, G = Goal,  
D = Deed, VO = Valued Object,  
V = Victim, A = Actor.



Corpus linguistics: risk as token (Hamilton et al., 2007)

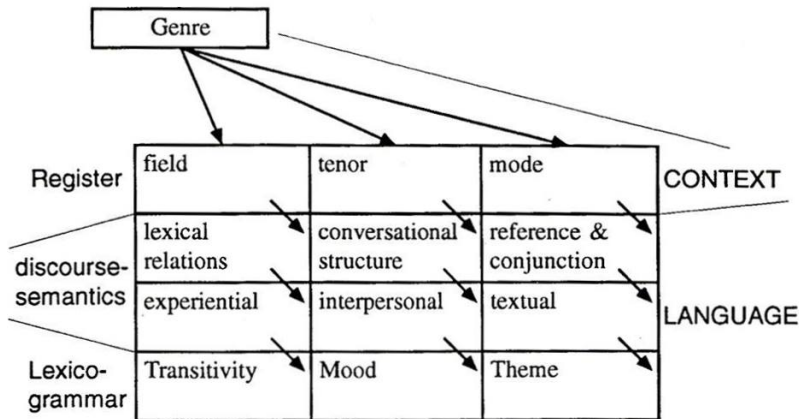
- Topics and text-types in which risk tokens appear
- Collocates of risk tokens
- Risk appears a lot in discussions of health
- Use of risk words is different to invented examples

Shortcomings:

- Smaller corpus size, heterogeneity of samples
- No parsing, lemmatisation
- No systematic connection of lexicogrammatical patterns to discourse/meaning

- Get all paragraphs containing \brisk in all 1987–mid 2014 articles
- Annotate/parse the data with full *Stanford CoreNLP* suite with embarrassingly parallel HPC
- Develop **corpk**it, a toolkit for searching the corpus and communicating results
- Interrogate the corpus according to notions from systemic functional grammar
- Connect to sociological theory

- *Systemic*: lexis as delicate grammar
- *Functional*: focus on language as a tool for the performance of functions
  - ① Interpersonal: negotiating relationships
  - ② **Experiential: representing the world**
  - ③ Textual: reflexive organisation into meaningful sequences



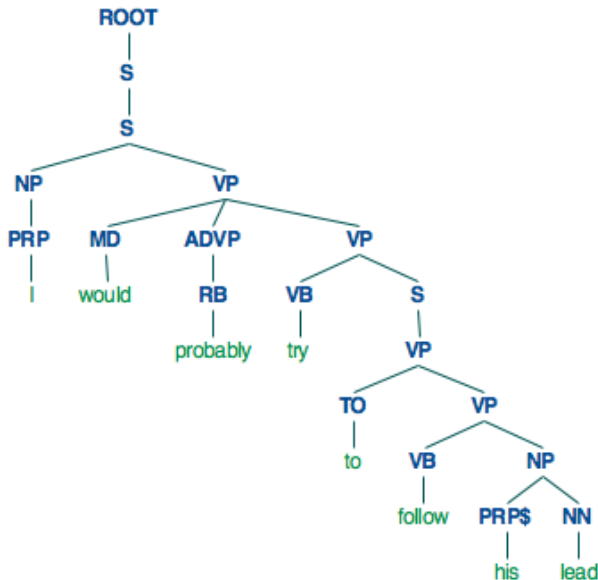
- Focus on the clause as a unit of analysis
- Centre on the *process* (i.e. rightmost verb in VP)
- Processes *select* participants (i.e. arguments of the verb)
- PPs and RBs are typically *circumstances*



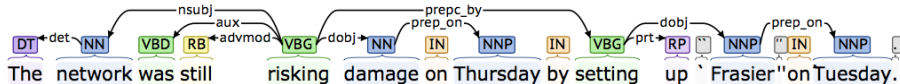
# SF transitivity analysis

<i>But</i>	<i>the bang of the gavel</i>	<i>can hold</i>	<i>risk</i>	<i>for novices</i>
	Participant: Carrier	Process: Relational attributive	Participant: Attribute	Circumstance: Extent

# Constituency grammar



# Dependency grammar



# The controversial question

The question: *Can we get systemic functional information from constituency and dependency parses?*

The answer: *Yep, quite a lot.*

How to investigate this huge dataset, and make the investigation transparent/reproducible?

- **corpkit**: a Python module designed for parsed and structured corpora
  - ▶ **interrogator()**: search for lexicogrammatical phenomena in each subcorpus, tally results, output Pandas objects
  - ▶ **editor()**: edit results, calculate keyness, linear regression
  - ▶ **plotter()**: visualise via *matplotlib*
  - ▶ **conc()**: concordance via parses
- Scriptable, multiprocessing, handles arbitrary data, open-source
- Systemic-functionally aware
- More recently, a GUI, aimed at corpus linguists

corpkit: lanc\_demo

Build **Interrogate** Edit Visualise Concordance

Corpus: client-a-stripped-parsed

Search: Trees

Query: /NN. 2/ >># @NP

Return:  
☐ Token ☒ Lemma ☐ POS ☐ Tree ☐ Count  
☐ Index ☐ Distance ☐ Function ☐ Governor ☐ Dependent

Preset query: Off Spelling: Off

Ngrams: Size Split contractions: No

☐ Lemmatise ☐ Multiword results  
☐ Filter titles ☐ Case sensitive

Blacklist:

Function filter:

POS filter:

Result word class (for lemmatisation):

☒ Speakers: ALL  
 CLIENT  
 THERAPIST

Dependency type: CC-processed

Interrogation name: untitled

Interrogate

Interrogation results: subheads

	01	02	03	04	05	06	07	08	09	10	11	12	13	Total
thing	34	24	21	25	16	29	22	34	31	14	30	18	34	332
person	25	28	17	17	26	23	12	18	15	12	32	11	33	269
something	13	10	13	25	26	8	26	23	16	30	29	27	19	265
time	29	20	18	13	15	15	17	28	14	29	26	21	16	261
way	11	8	8	16	8	19	13	11	13	6	11	16	22	162
kind	21	5	9	8	10	14	16	8	14	10	9	5	3	132
situation	1	9	8	21	6	6	9	7	3	5	23	15	11	124
family	9	17	4	14	40	28	0	0	0	0	1	0	1	114
dad	19	14	10	11	8	7	6	7	9	2	4	4	8	111
someone	7	5	5	8	10	9	7	3	3	12	11	6	23	109
lot	16	8	6	2	6	6	10	14	7	6	14	1	10	108
day	9	13	4	2	2	1	7	5	0	27	9	3	21	103
mom	20	11	7	10	26	12	2	1	4	0	2	0	3	96
parent	12	16	16	9	4	7	5	0	17	1	6	0	2	95
anything	10	19	3	8	8	4	5	4	8	4	9	1	7	90
conversation	6	7	7	5	10	0	4	33	8	2	2	2	1	89
feeling	1	5	5	4	4	10	8	6	21	0	6	8	8	86
part	4	8	11	9	11	7	3	7	6	0	3	1	4	74
week	3	2	0	0	1	8	4	5	0	22	7	1	13	66
problem	0	5	12	1	6	0	0	4	5	11	5	9	6	64
bit	5	2	5	1	4	4	4	7	11	9	2	5	4	63
relationship	8	7	7	9	6	2	7	3	2	2	3	0	2	58
unach	1	2	0	1	6	10	6	7	8	6	4	7	2	NA
Total	796	653	442	470	586	515	558	672	520	614	674	362	671	7535

Previous Next Save as dictionary Update interrogation

14:07:04: Log saved to "logs/log-00.txt".

corpkit: risk

Build Interrogate Edit Visualise **Concordance**

24	WAP-2013-01.txt.xml	ayer isn't worried about the new rental securities creating	systemic	risk, because the institutional pr
33	WAP-2013-02.txt.xml	ka 's true financial condition and make it difficult to spot	systemic	riska, says Anat Admati, a finance
29	WAP-2013-01.txt.xml	o officials said policy makers haven't fully dealt with the	systemic	risk posed by large Wall Street ba
25	brainstorm	ncial system had "expose -LSB- d -RSB- the system to large	systemic	shocks "from sudden changes in th
22	brainstorm	But the risks that will have the most impact are	systemic	financial failure, government debt
18	brainstorm	keep a close eye on riskier product lines but not under the	systemic	designation, which could alter the
15	brainstorm	leton Community Bank, added, "this legislation will reduce	systemic	risk, protect taxpayers and put ou
14	brainstorm	ve for investors than exchange trading, it is also prone to	systemic	risks. "
9	brainstorm	ose same scholars are also trying to improve how we measure	systemic	risk.
7	brainstorm	"It's a more	systemic	way to determine whether or not a
6	brainstorm	system, which encompasses identifying potential sources of	systemic	risk - a pretty broad mandate.
49	WAP-2013-02.txt.xml	The remedy for bubbles and panics, if any, lies in	systemic	reform, an objective that the case
9	brainstorm	keep a close eye on riskier product lines but not under the	systemic	designation, because the label cou
5	colour	s were given an spid for pain, but in most, there were no	systematic	checks for withdrawal symptoms or
28	WAP-2013-02.txt.xml	The failure of some schools to provide a more	systematic	education is saddening and inconsi
11	WAP-2013-01.txt.xml	at the University of Michigan, to conduct one of the first	systematic	studies of Iranian Internet censor
4	colour	t or assault, and realize that our only defense lies in the	systematic	reduction of risk factors -- which
19	WAP-2013-02.txt.xml	You must want the	systematic	breath and depth in order to get
8	WAP-2013-02.txt.xml	relevant time, with news of the National Security Agency's	systematic	surveillance of citizens ' phone
3	colour	those settings, you're faced with a choice : Chart your own	systematic	breath and depth, or risk piece-
38	WAP-2013-02.txt.xml	t or assault, and realize that our only defense lies in the	systematic	reduction of risk factors - which
1	misc	But what exactly will this	sysadmin-vaporizing	pixie dust look like ?
48	scheme	rescribed between the 1960s and the early 2000s -- supplies	synthetic	versions of the lost estrogen and
13	scheme	Daimler's electric cars are coming with a	synthetic	room to thwart the risks of silen
24	scheme	alone or estrogen and progesterone together, found that the	synthetic	hormones increased the risk of hea
12	scheme	BlueMountain says that in 2011 it bought a portfolio of	synthetic	CCOs and CDOs from Credit Agricole
36	WAP-2013-01.txt.xml	rely unmixd, and so it is with the development of powerful	synthetic	or semi-synthetic opioid analgesics - painkillers such as f
27	WAP-2013-01.txt.xml	rely unmixd, and so it is with the development of powerful	synthetic	or semi-synthetic opioid analgesics -- painkillers such as
30	WAP-2013-01.txt.xml	If you use	synthetic	fertilizers, keep the granules away from the plant stem.

WAP-parsed

2013

Preset query

Trees

CC-processed

/JJ.?! < /~^y/

☐ Speakers:
 ☐ Split sentences
 ☐ Show trees

Limit results to function:

nsubj (pass) ?

M1

☒ Index
 ☒ Filenames
 ☒ Scheme
 ☐ Speakers
 60

Stored concordances

Key: 0

Key: 1

Key: 2

Key: 3

Key: 4

Key: 5

Key: 6

Key: 7

Key: 8

Key: 9

Coding scheme

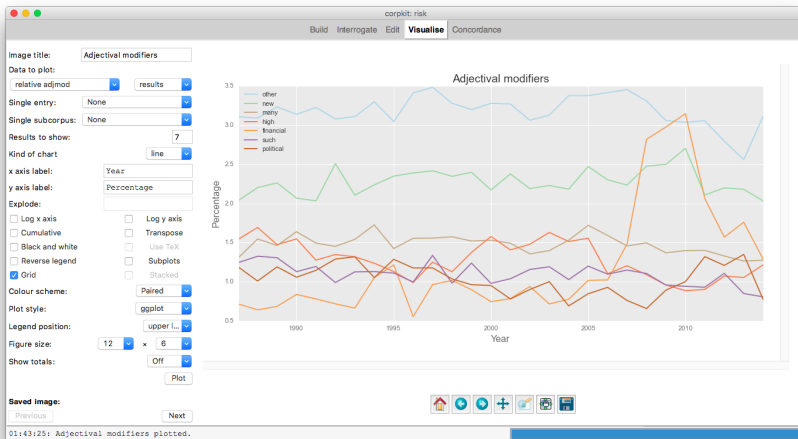
brainstorm

colour

scheme

misc

01:57:34: Concordancing done: 50 results.





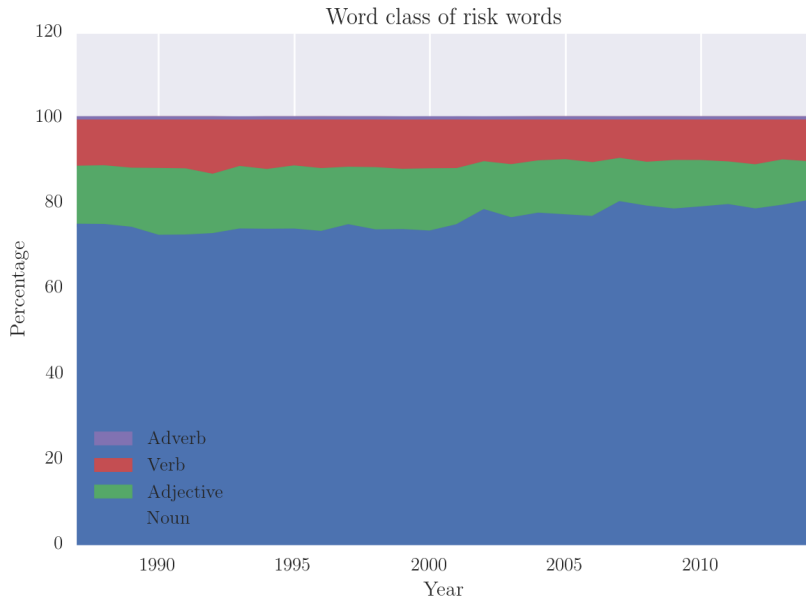
```
# import module and set data path
>>> from corpkit import *
>>> corpus = 'data/NYT-parsed'

# get pos of risk words, show word class
>>> res = interrogator(corpus, 'words', r'\brisk',
...     show = ['p'], lemmatise = True)

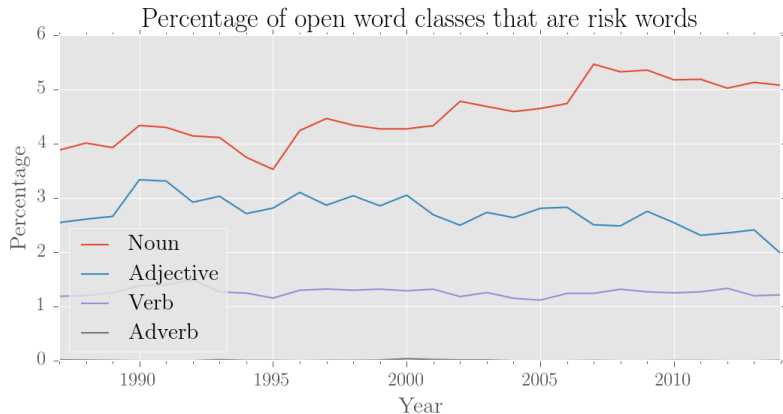
# get relative frequency
>>> rel = editor(res.results, '%', res.totals, keep_top = 4)

# visualise
>>> plotter('Word class of risk words', rel.results,
...     kind = 'area', style = 'seaborn-talk')
```

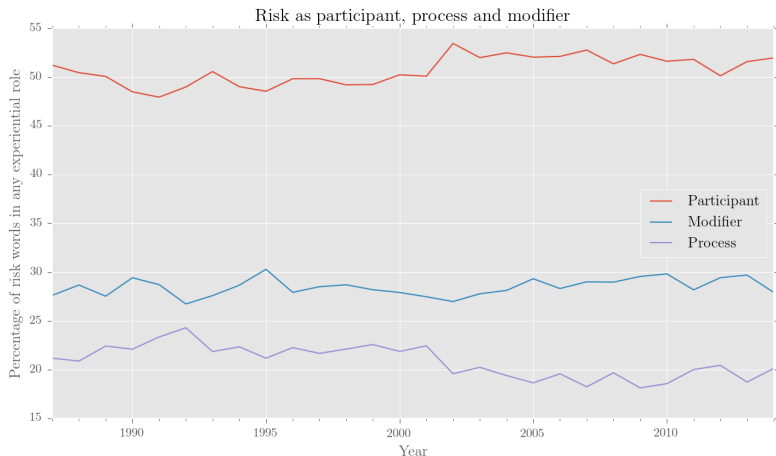
# Example output



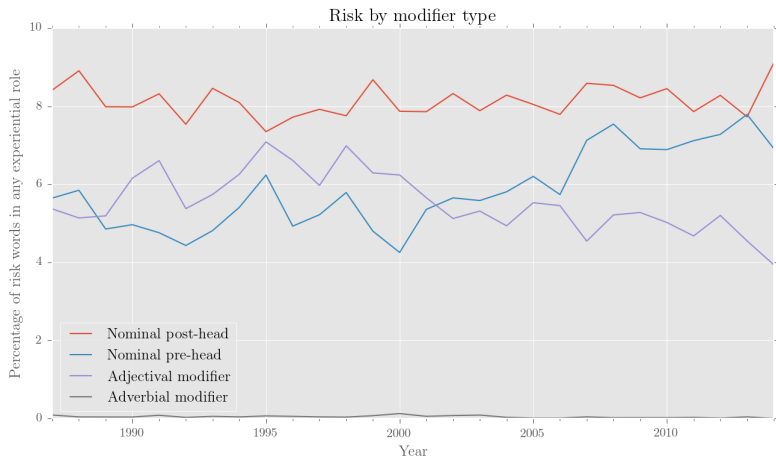
# Nominalisation of risk in the NYT



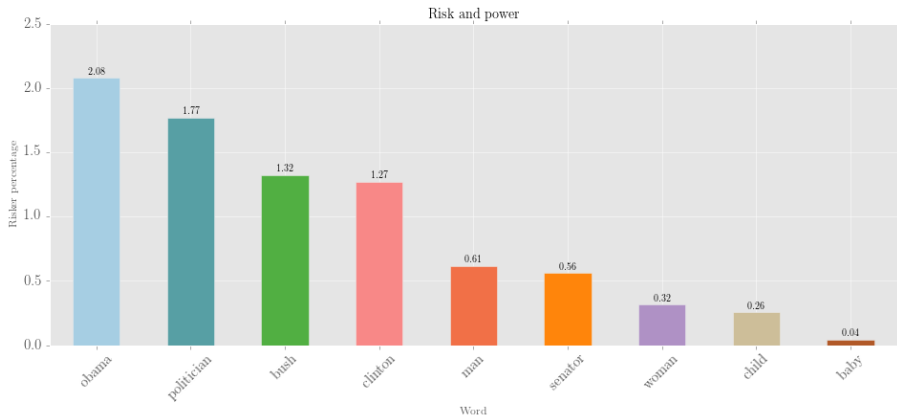
# Experiential roles of risk words



*They risked their life → It was a risk*

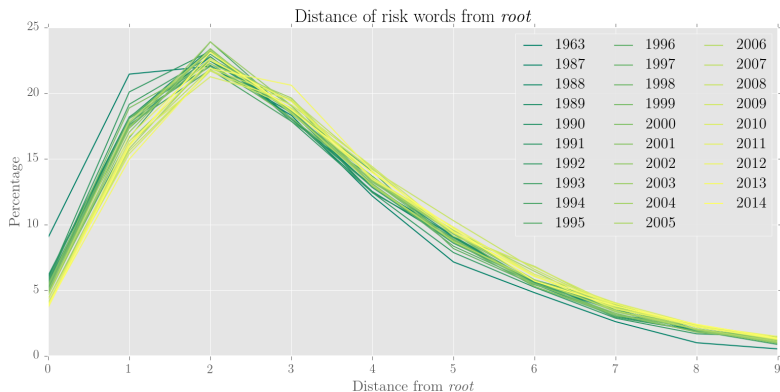


*Risky decision → risk assessment*



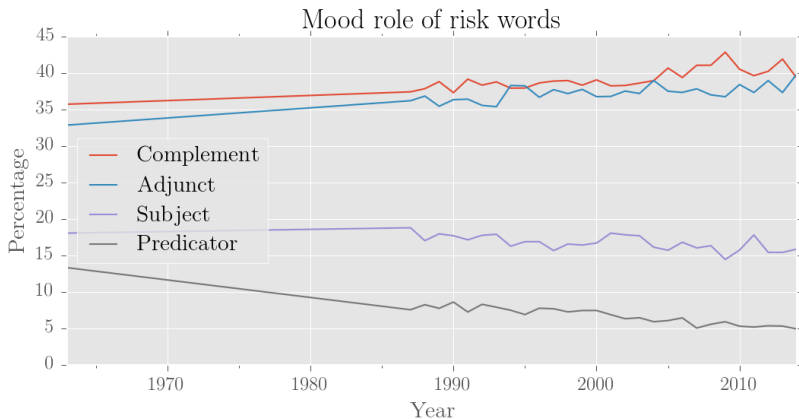
→ Powerful and influential people *do* risking

# Distance of risk word from *root*



→ looked promising, but seems to be a general phenomenon.

# Mood role of risk words (NYT)



Arguable → inarguable

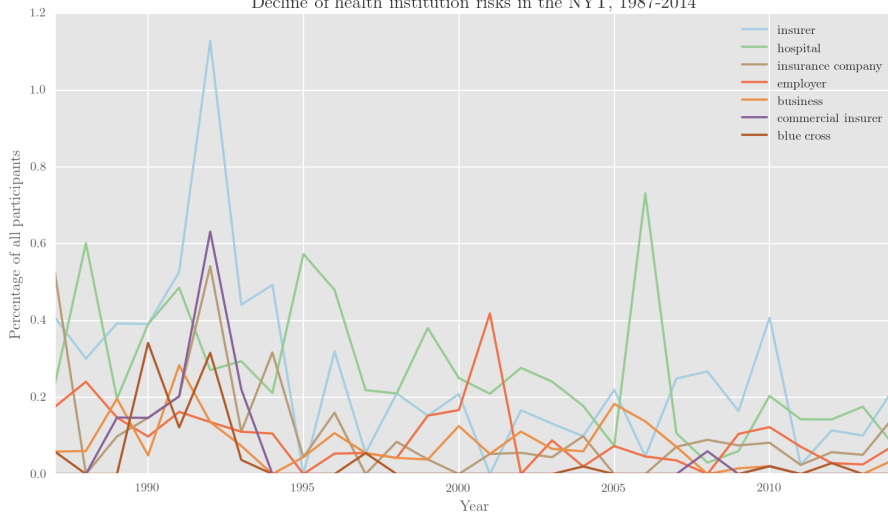


- Nominalisation and *participation*: synonymy of risk and negative outcome
  - ▶ risking harm → risk assessment
  - ▶ Meaning of risk expanding beyond the *risk frame*
- Risk words becoming more implicit
  - ▶ Routinisation of the management of risk
  - ▶ Risk as increasingly present, but decreasingly debated
- More everyday exposure to risk, but less risking
- Neoliberal conceptualisations of agency: institutional expectation to take risk, absolution of responsibility for institutions themselves

- As earlier studies have shown, risk words often occur in health domains.
- The NYT Annotated Corpus had some manually added topic tags
- We created a subcorpus of health articles

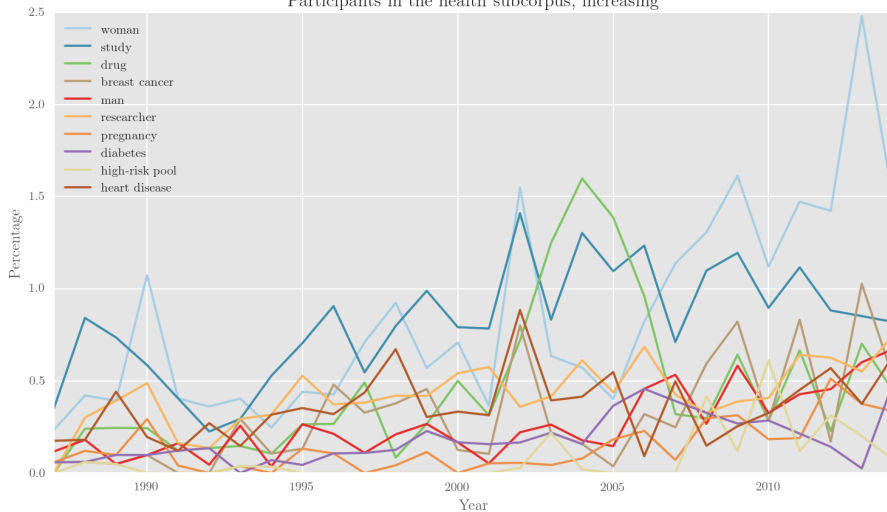
# Decreasing participants in health discourse

Decline of health institution risks in the NYT, 1987-2014



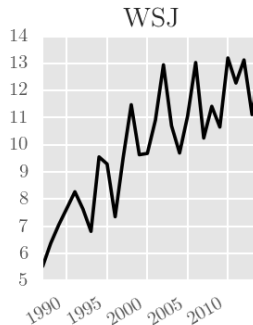
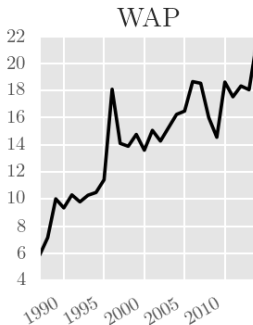
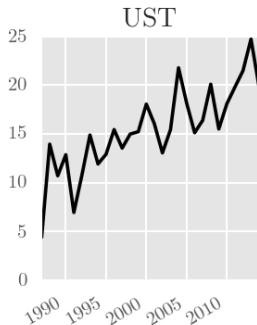
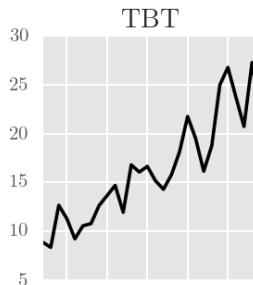
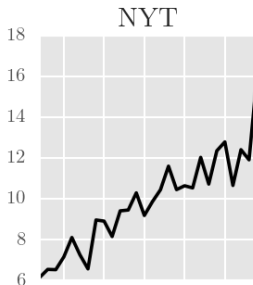
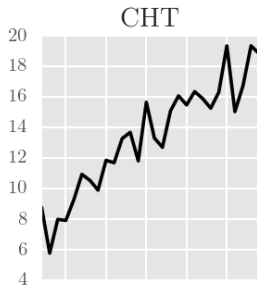
# Increasing participants in health discourse

Participants in the health subcorpus, increasing

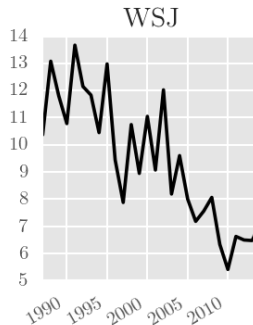
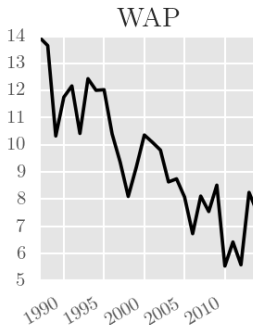
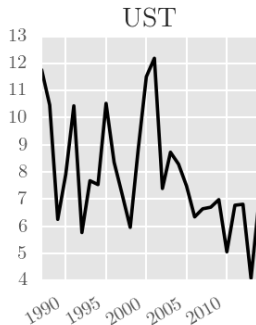
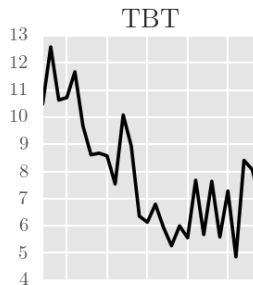
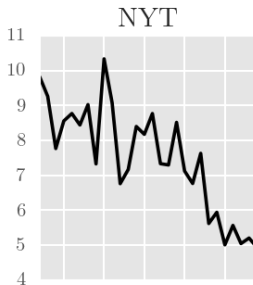
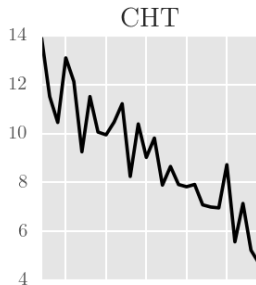


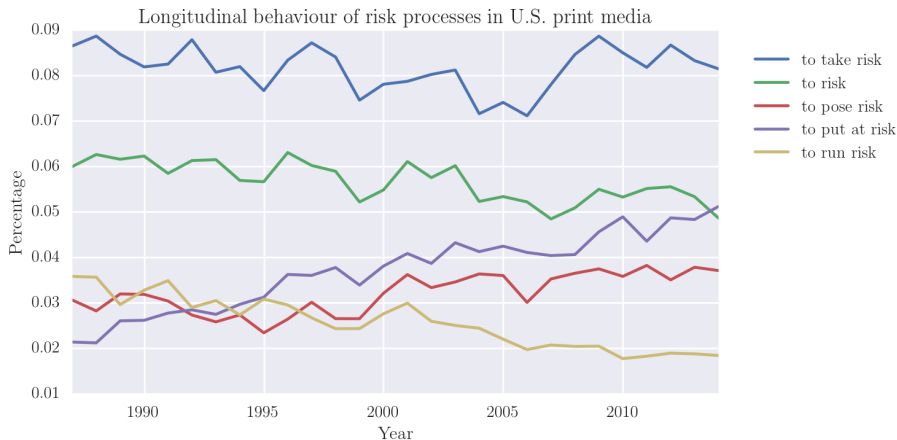
- We've only just started interrogating the six newspaper corpus
- First, we'd like to check if the NYT findings are generalisable to other publications.
- Then, understanding the reason for differences and similarities would be nice
- Would love help on dealing with the complexity of the data structure!

# *To put at risk*



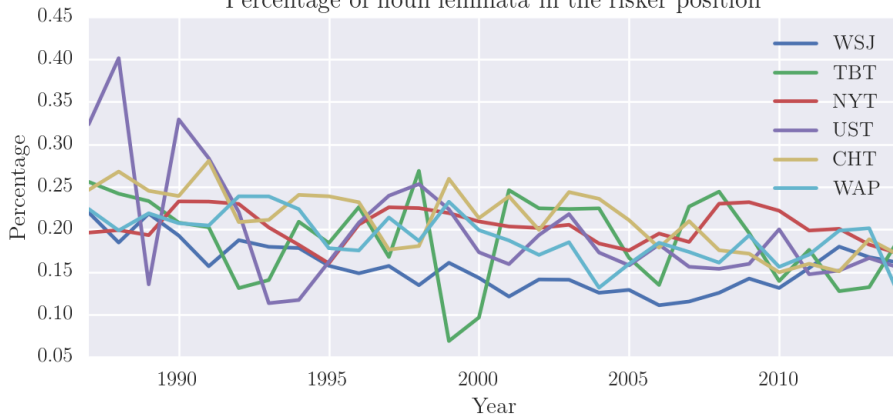
# *To run risk*



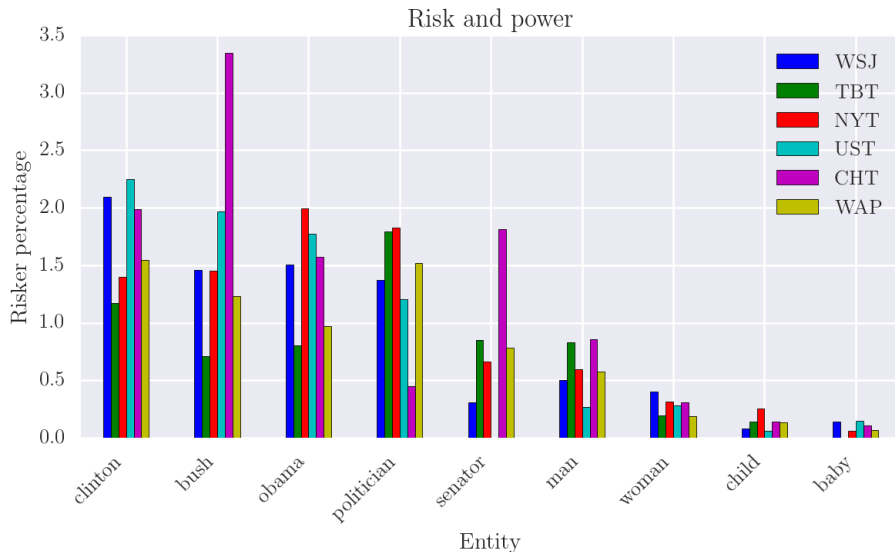




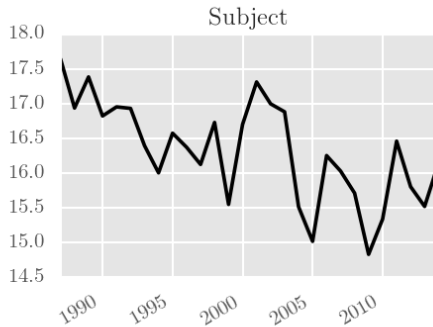
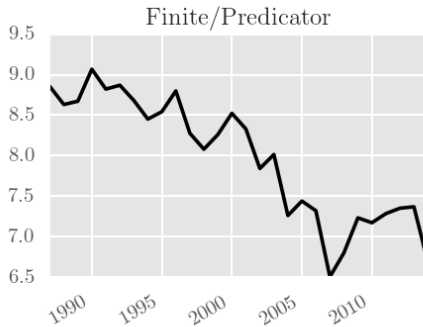
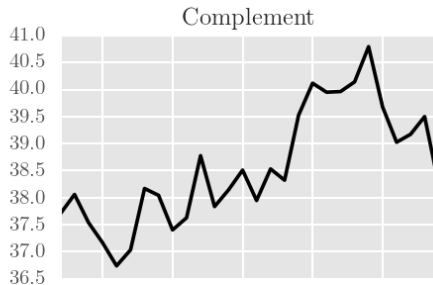
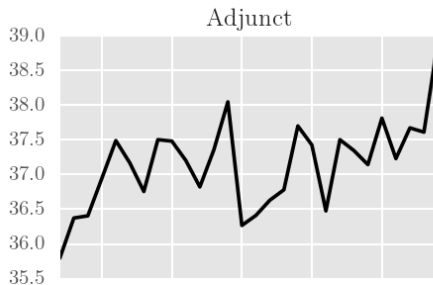
Percentage of noun lemmata in the risker position



# Risk and power across publications



# Mood role of risk words



- Many phenomena generalisable
- Some newspaper specific constructions: *risk appetite* in the WSJ
- Fewer grammatical riskers, but risk characterising more participants and processes
- Hints of influence of newspaper's politician position

- SFL proves a useful means of dividing up and investigating the behaviour of a given word
- Systemic categories are *sometimes* more telling than formal/constituency/dependency labels
- Though theoretical orientations are different, much of the grammar (esp. at group/phrase levels) are actually very similar

- This is a study of risk words, not risk
- Congruent realisations are analysed at the expense of the incongruent
- Little concordancing, close reading of individual texts
- Parser accuracy
- Lack of reference corpus to compare related words/general language

Data and tools are available for reuse:

- <https://www.github.com/interrogator/risk>
- <https://www.github.com/interrogator/corpkrit>

Findings are presented dynamically in an Jupyter Notebook:

- NYT: <http://git.io/vIM2W>
- All: <http://git.io/vBTHI>

Project report:

- <http://git.io/vZ7yh>

This slideshow:

- <http://git.io/vBfbw>

- Beck, U. (1992). *Risk society: Towards a new modernity*. Sage.
- Dean, M. (1998). Risk, calculable and incalculable. *Soziale Welt*, 25–42.
- Dean, M. (1999). *Governmentality: Power and rule in modern society*. SAGE Publications, Inc.
- Fillmore, C. J., & Atkins, B. T. (1992). Toward a frame-based lexicon: The semantics of RISK and its neighbors. *Frames, fields, and contrasts: New essays in semantic and lexical organization*, 103.
- Hamilton, C., Adolphs, S., & Nerlich, B. (2007, March). The meanings of ‘risk’: a view from corpus linguistics. *Discourse & Society*, 18(2), 163–181.
- Luhmann, N. (1993). *Risk: A Sociological Theory*. New York: Walter de Gruyter.
- Sandhaus, E. (2008). *The New York Times Annotated Corpus LDC2008T19*. Linguistic Data Consortium.