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

Overview

THE UNIVERSITY OF MELBOURNE

- 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`

The project(s)

THE UNIVERSITY OF MELBOURNE

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.

Context: sociological risk theory

THE UNIVERSITY OF MELBOURNE

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)

Context: risk as word

- 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*

Context: new methodologies

THE UNIVERSITY OF MELBOURNE

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

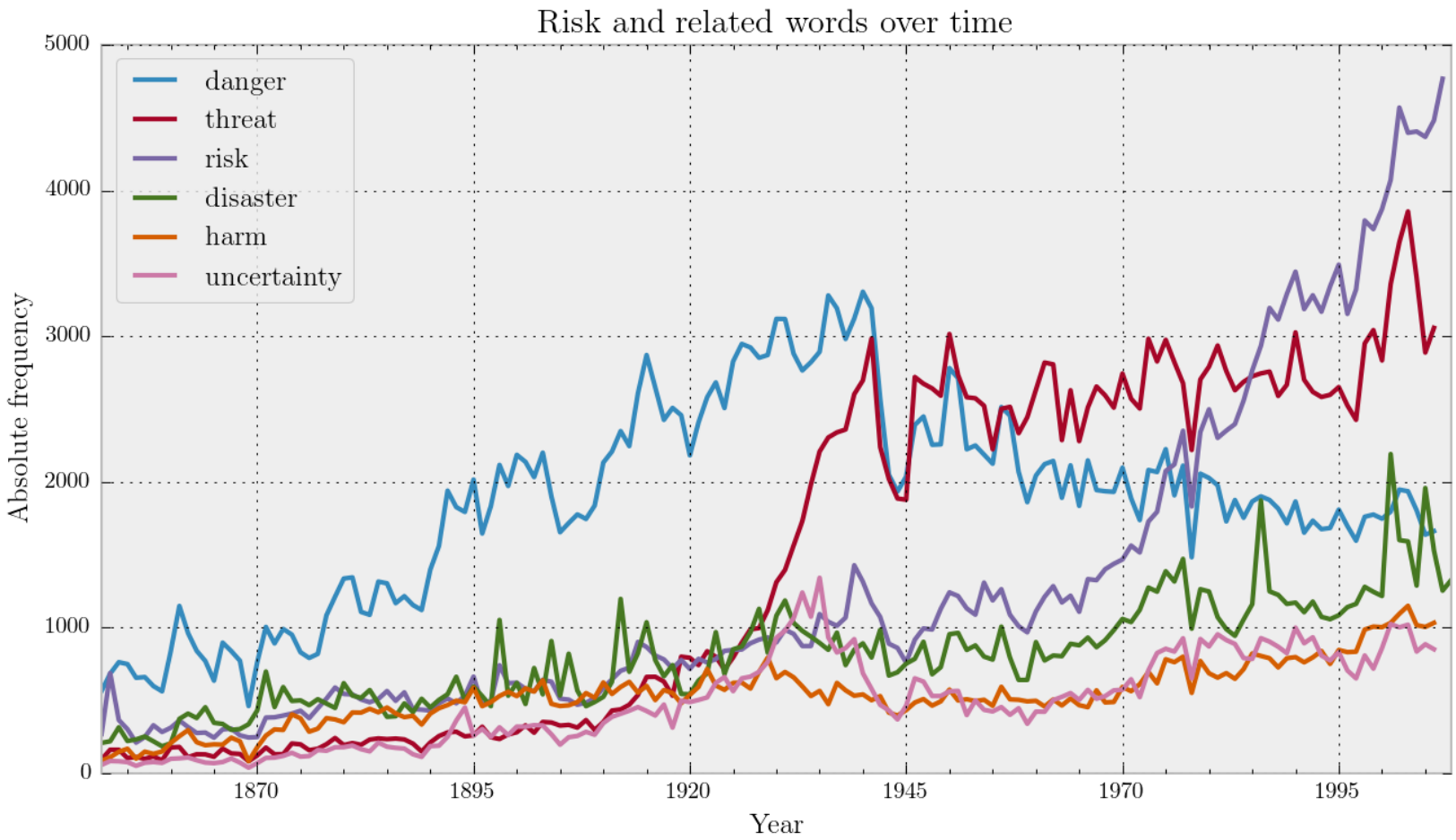
- Digitisation of newspapers means we have large, well-structured datasets
- Parsing makes it possible to search for lexical and grammatical features in tandem
- Programmatic approaches to social science research facilitate:
  - Automation
  - Reproducibility
  - Transparency
  - Ability to deploy methods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

The data

THE UNIVERSITY OF MELBOURNE

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* (Zinn 2011)



Research questions

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?

Linguistics: frame semantic approach

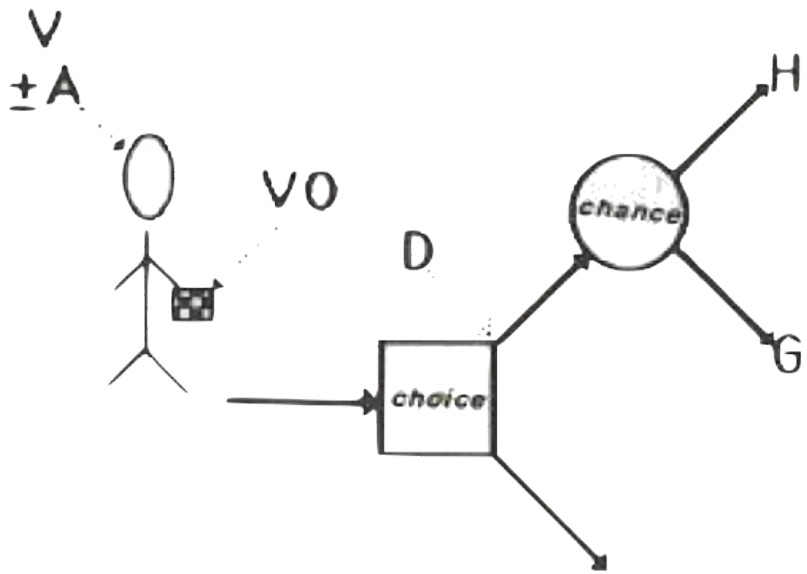
Frame semantics: risk as a cognitive schema (Fillmore & Atkins, 1992)

- 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, 1992)

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



Corpus linguistic approach

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

Our methods

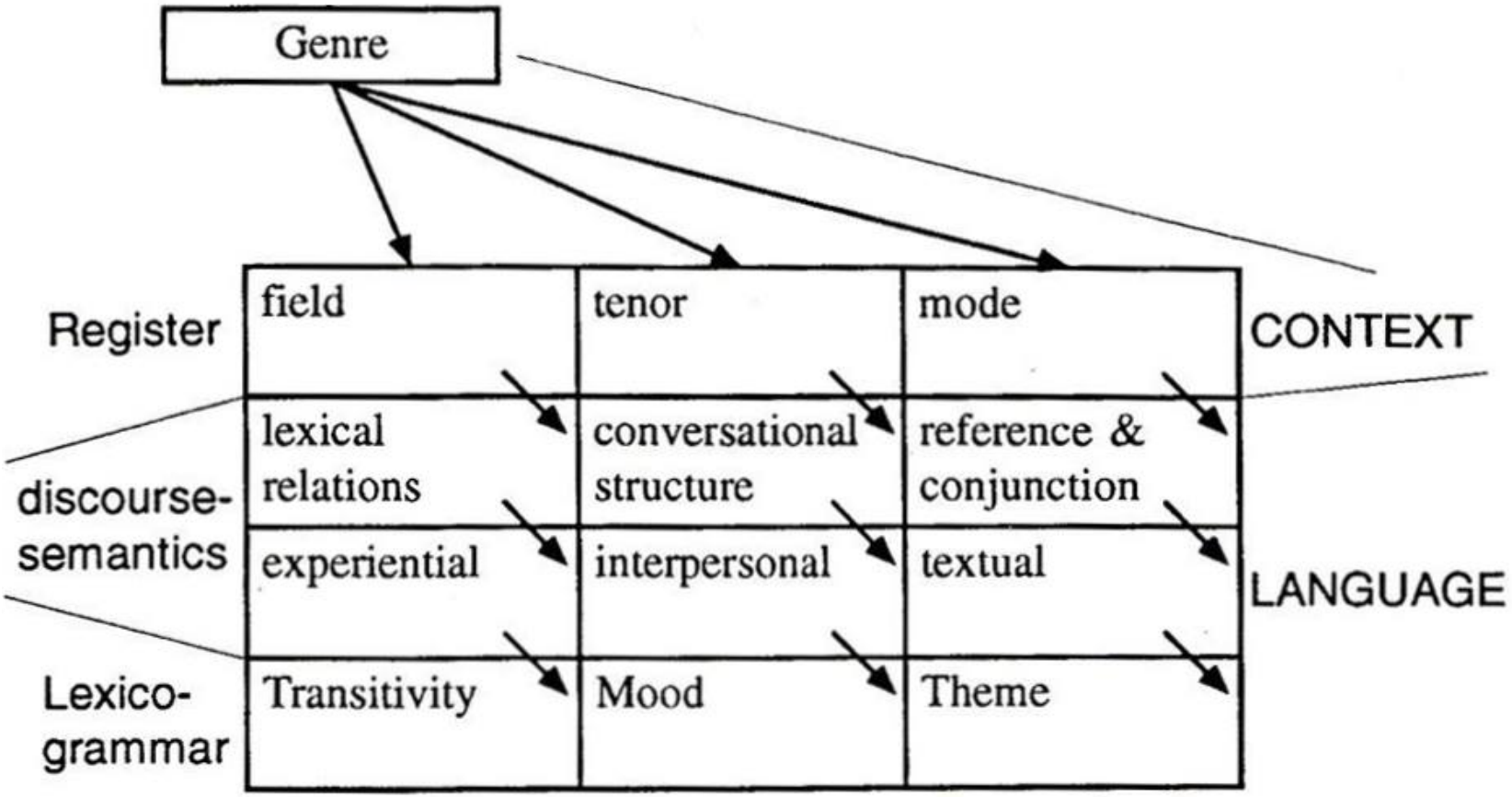
THE UNIVERSITY OF MELBOURNE

- 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 `corpkit`, a toolkit for searching the corpus and communicating results
- Interrogate the corpus according to notions from systemic functional grammar
- Connect to sociological theory

SFL: the basics

- *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

Overview of SFL



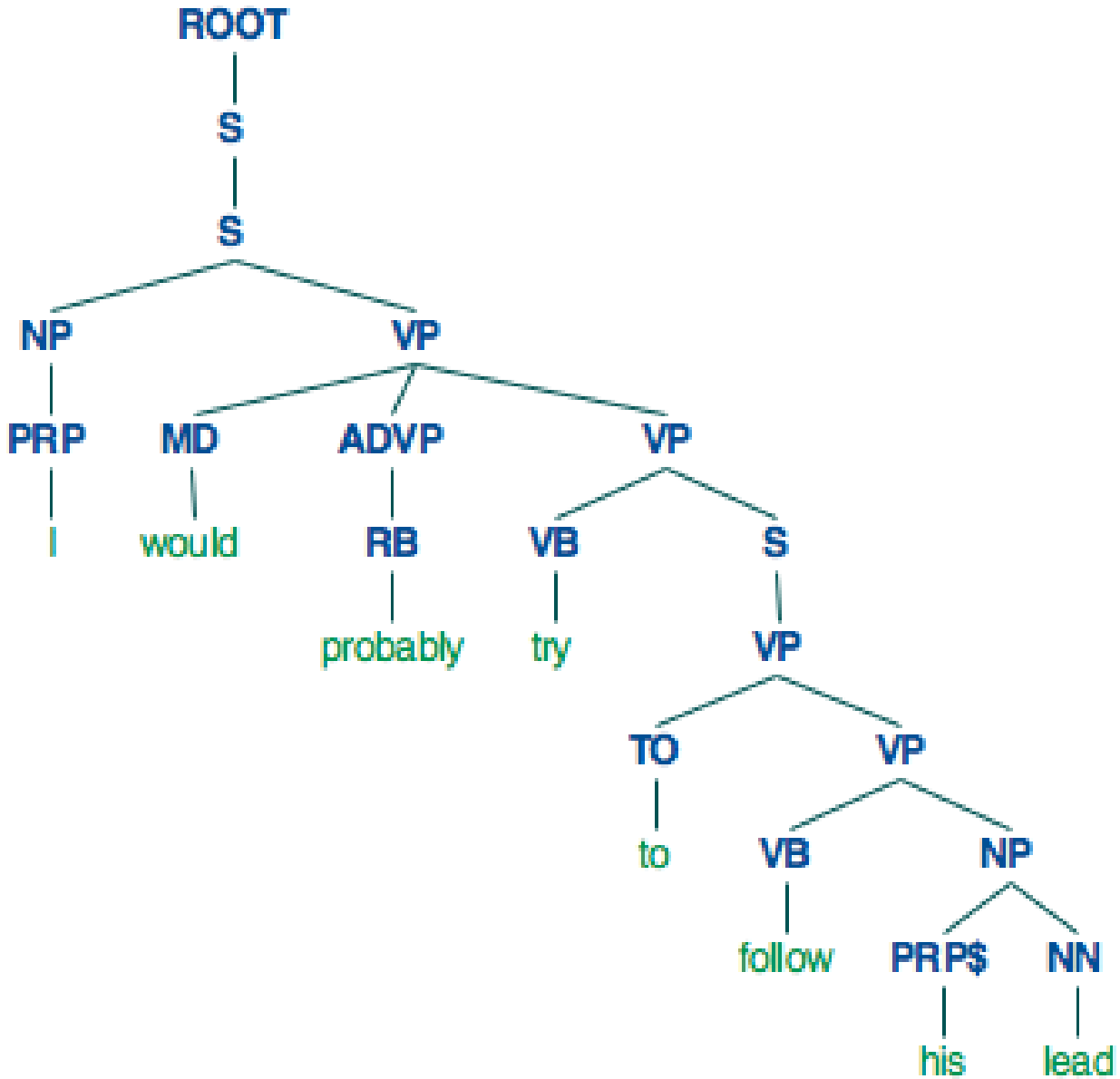
Transitivity system

- 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*

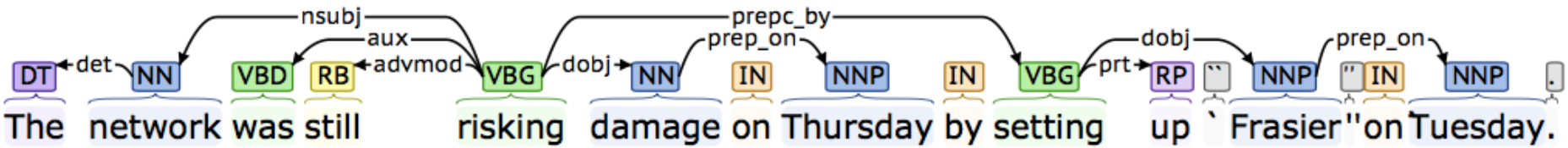
SF transitivity analysis

But	*the bang of the gavel*	*can hold*	*risk*	*for novices*
	Participant: Carrier	Process: Relational attributive	Participant: Attribute	Circumstance: Extent

Constituency grammar



Dependency grammar



The controversial question

THE UNIVERSITY OF MELBOURNE

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

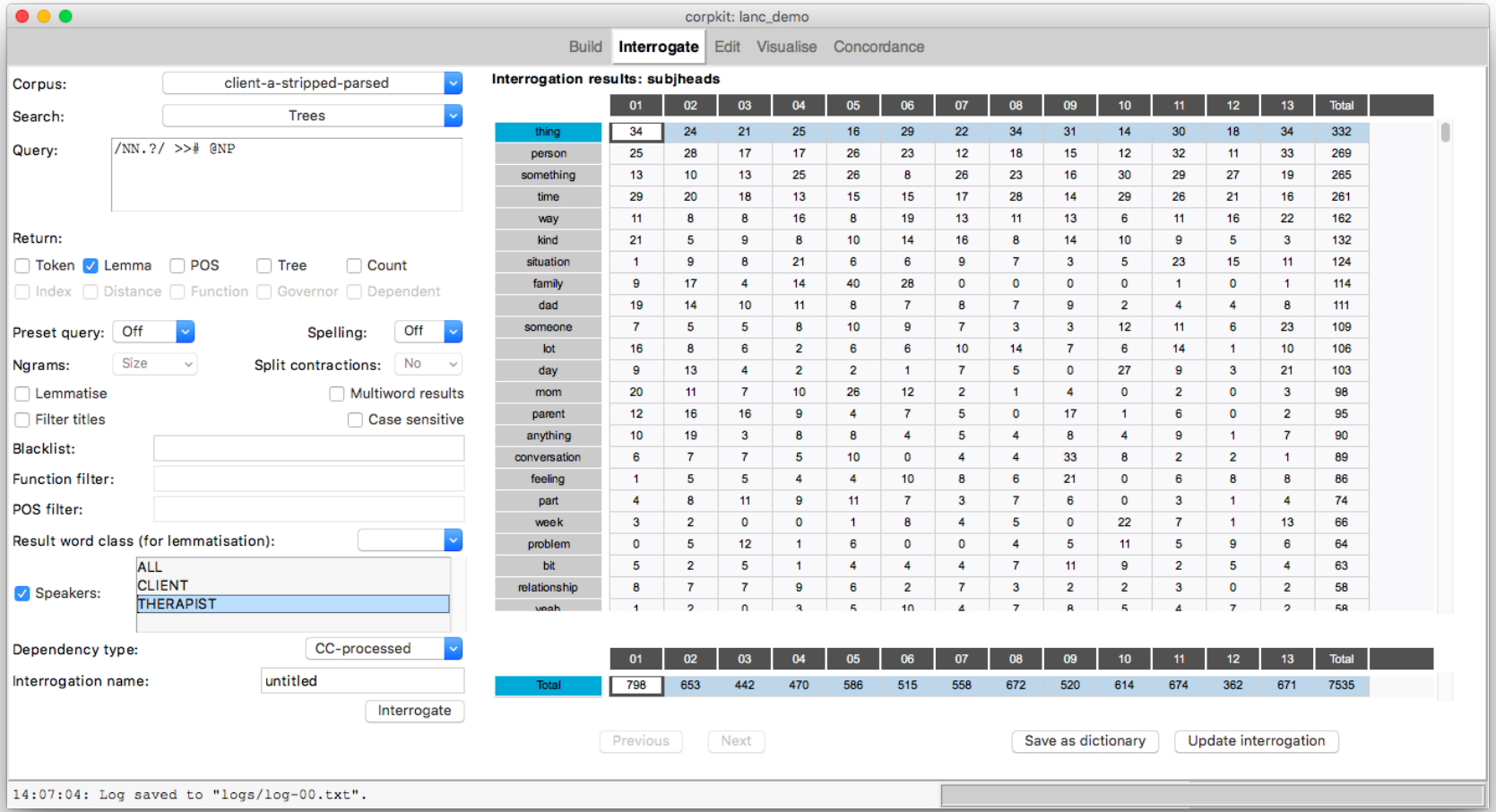
The answer: *Yep, quite a lot.*

Developing tools

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

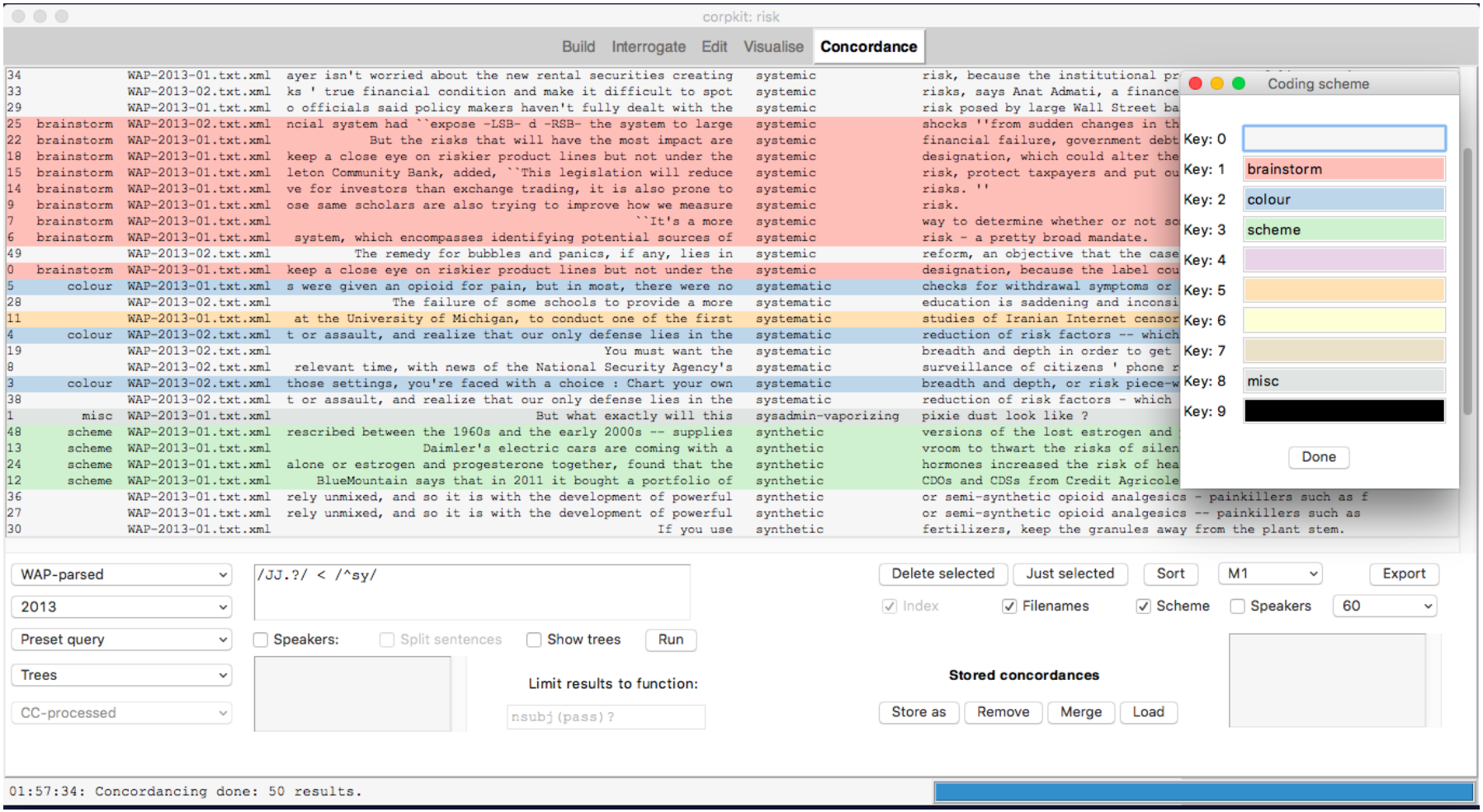
corpkit GUI: interrogating



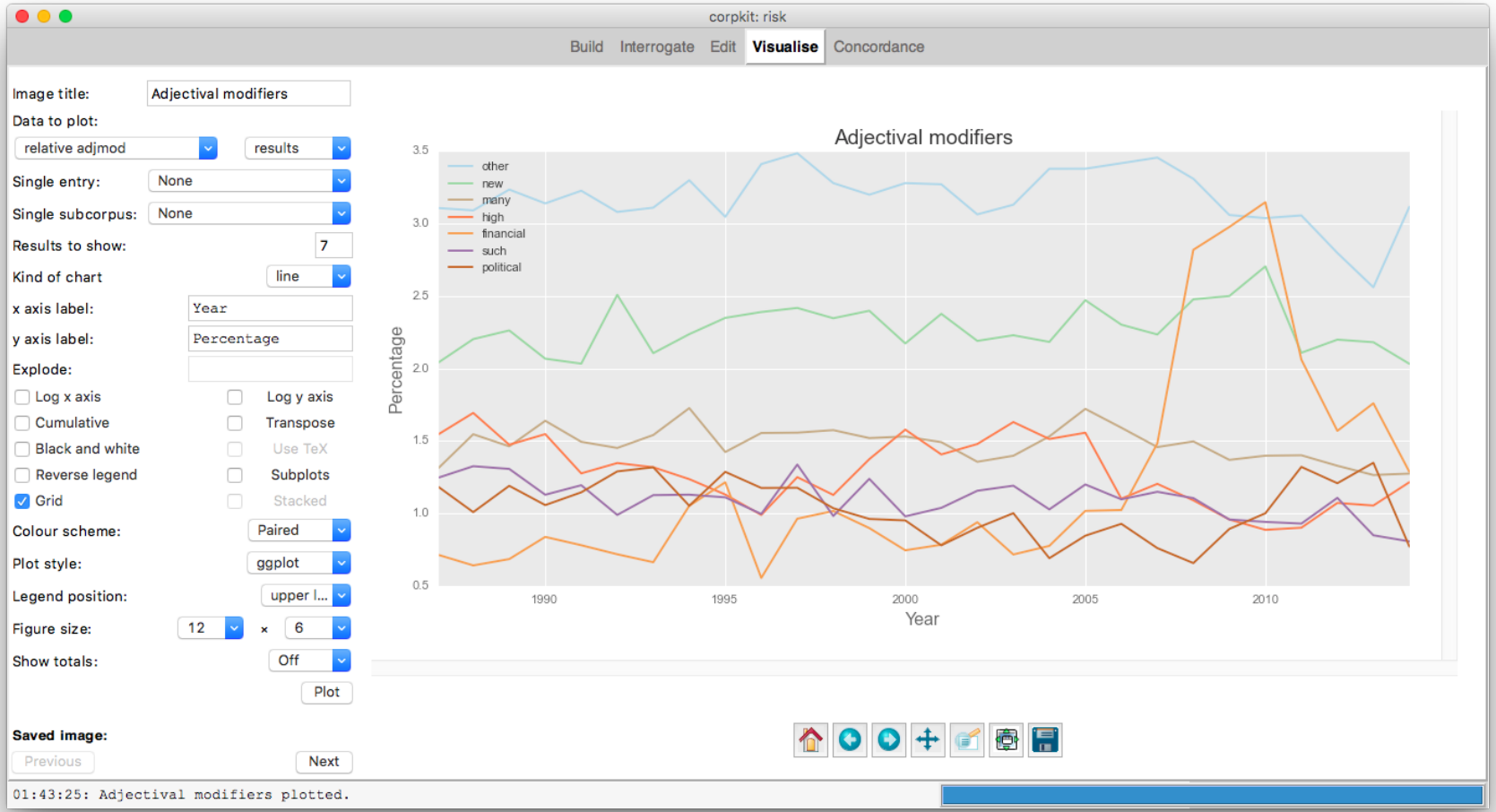
	01	02	03	04	05	06	07	08	09	10	11	12	13	Total
thing	34	24	21	25	16	29	22	34	31	14	30	18	34	332
person	25	28	17	17	26	23	12	18	15	12	32	11	33	269
something	13	10	13	25	26	8	26	23	16	30	29	27	19	265
time	29	20	18	13	15	15	17	28	14	29	26	21	16	261
way	11	8	8	16	8	19	13	11	13	6	11	16	22	162
kind	21	5	9	8	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	8	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	106
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	98
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	4	33	8	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
yeah	1	2	0	3	5	10	4	7	8	5	4	7	2	58

	01	02	03	04	05	06	07	08	09	10	11	12	13	Total
Total	798	653	442	470	586	515	558	672	520	614	674	362	671	7535

corpkit: concordancer



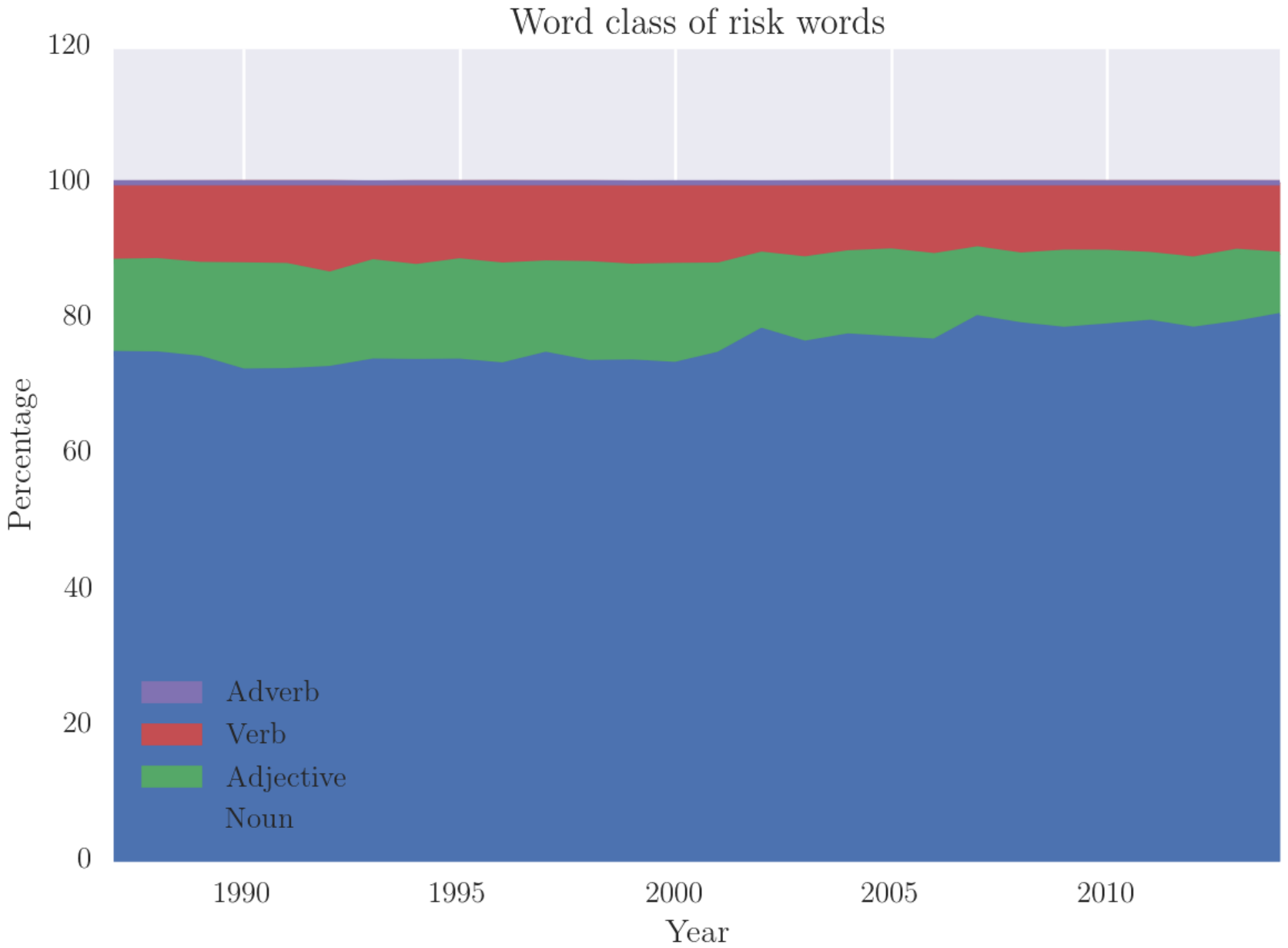
corpkit: plotting



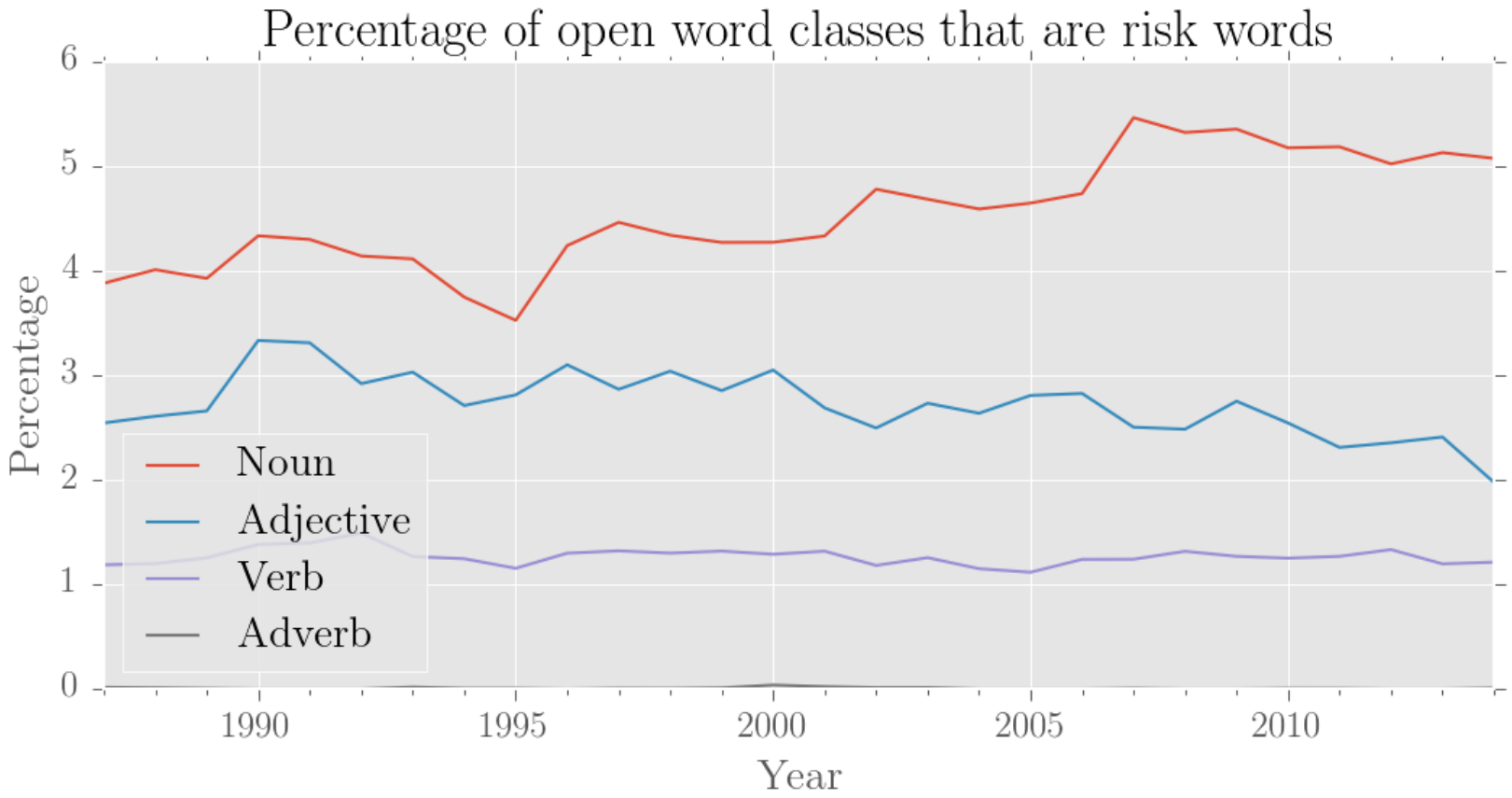
corpkit: example code

```
# 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')
```

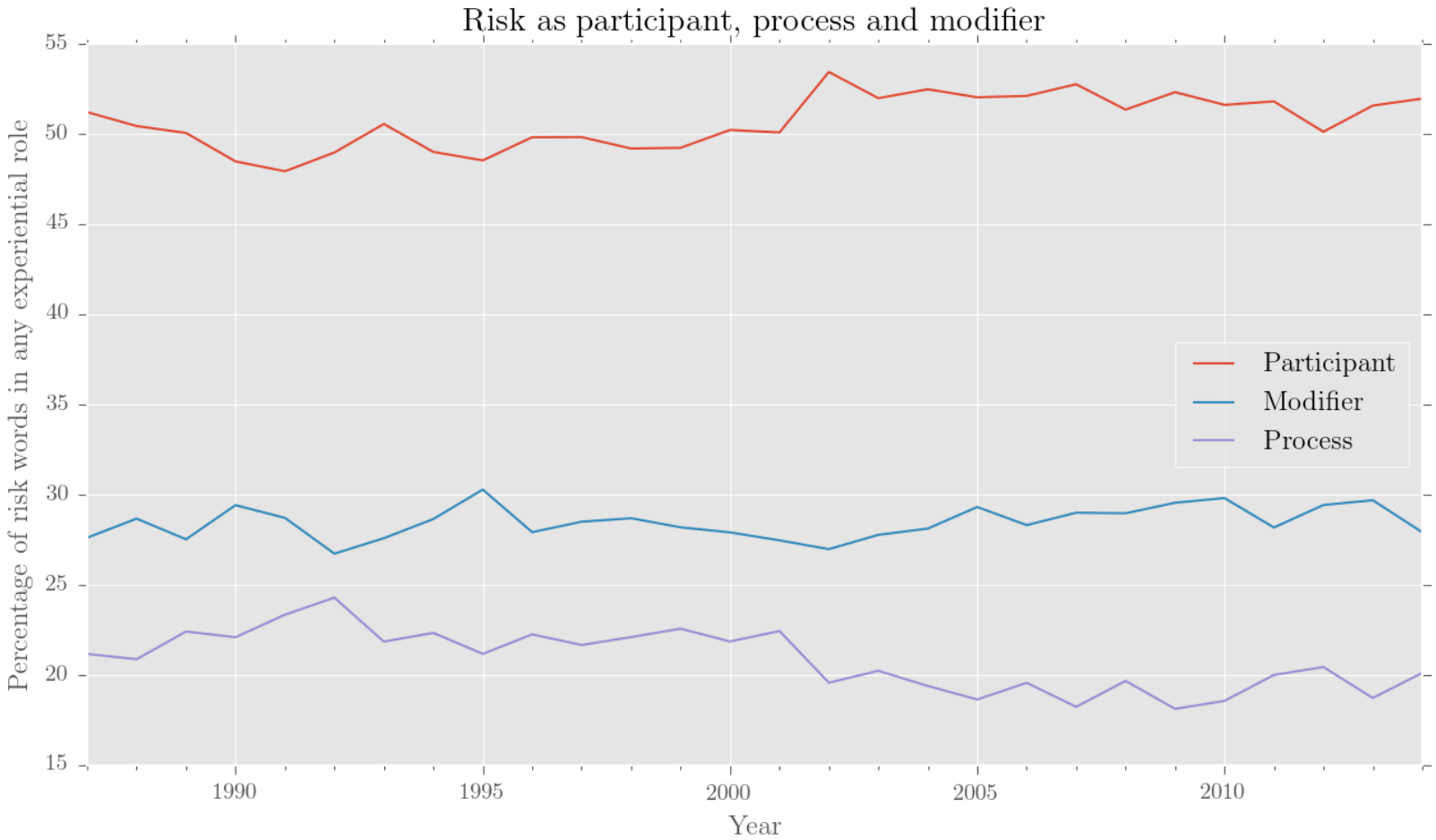
Example output



Nominalisation of risk in the NYT

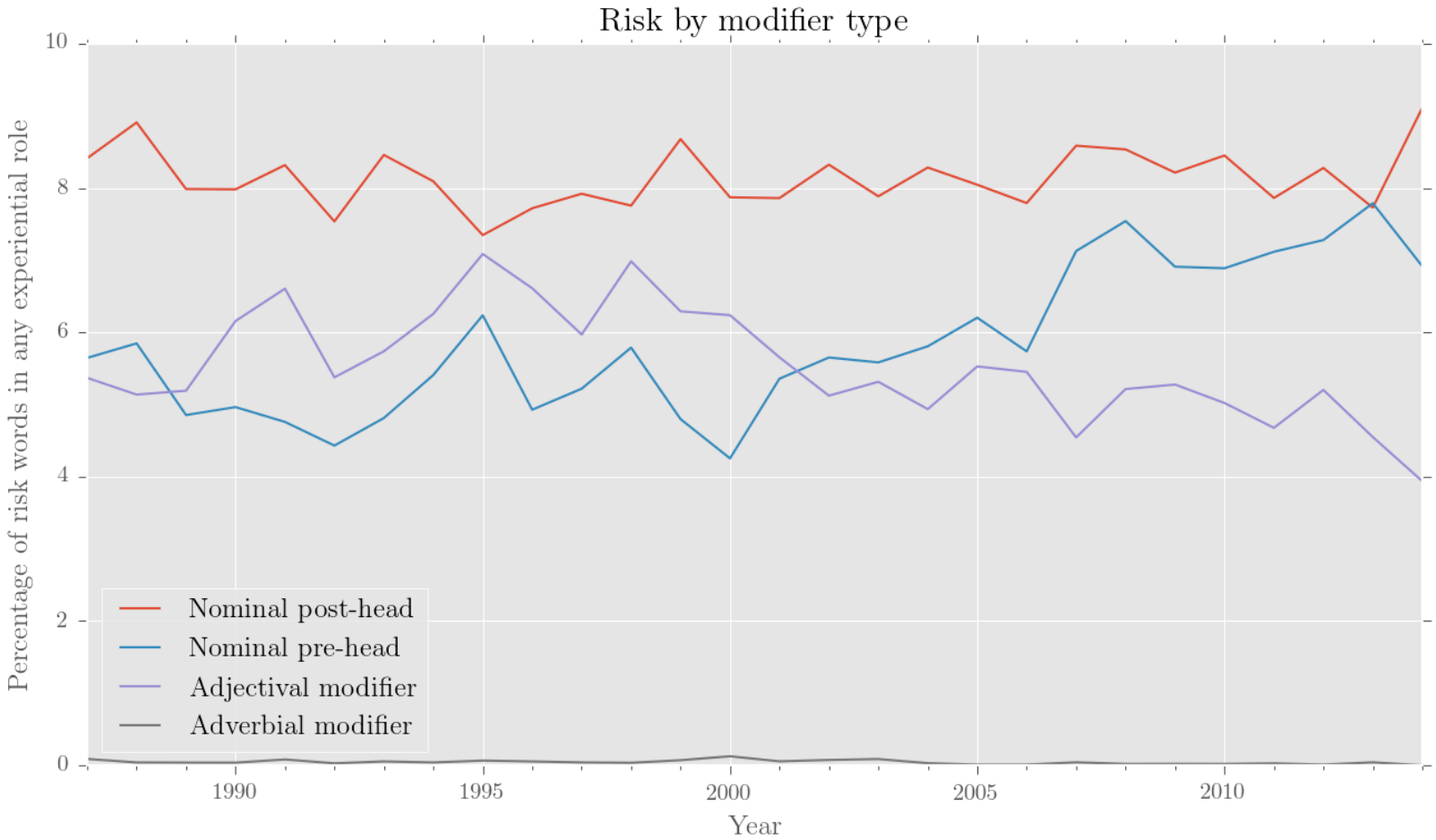


Experiential roles of risk words



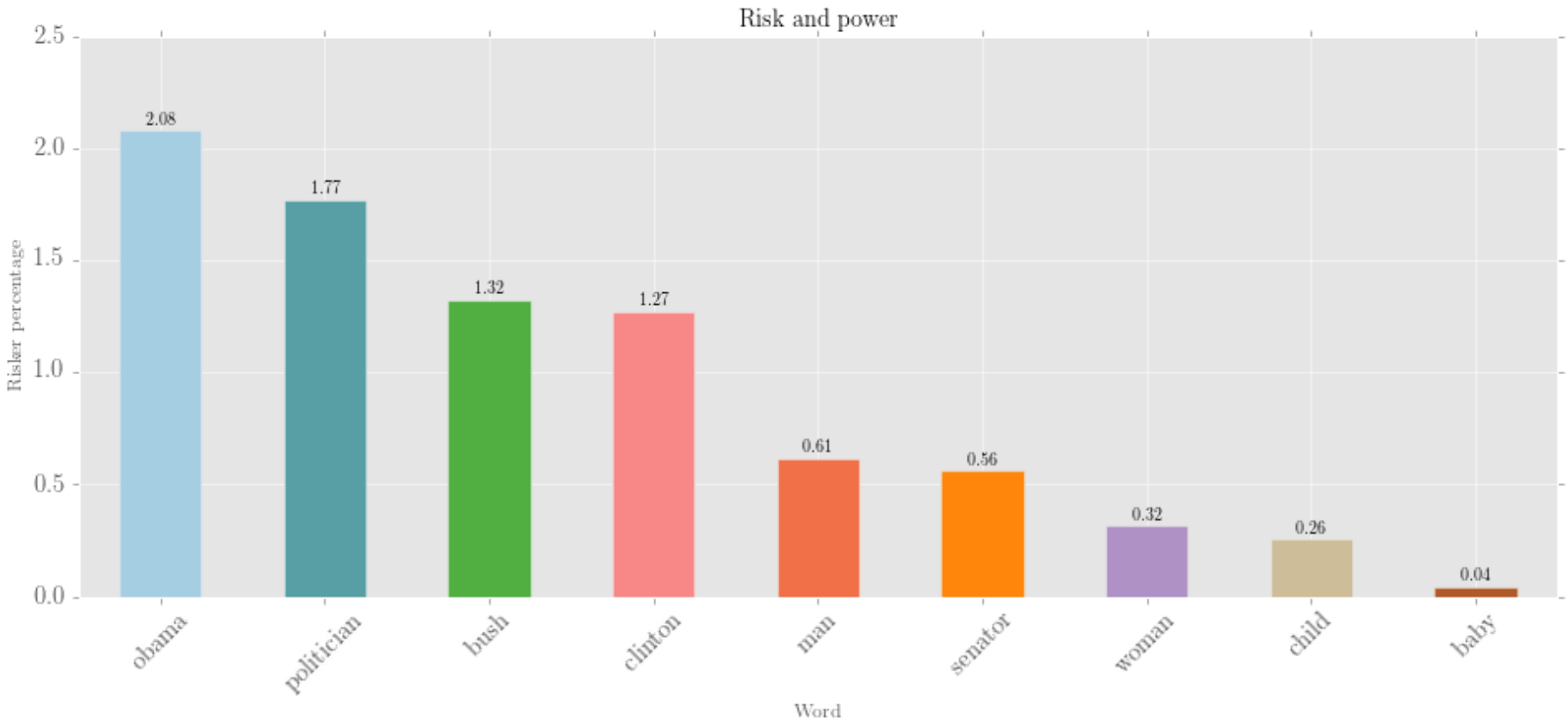
They risked their life → It was a risk

Risk as modifier



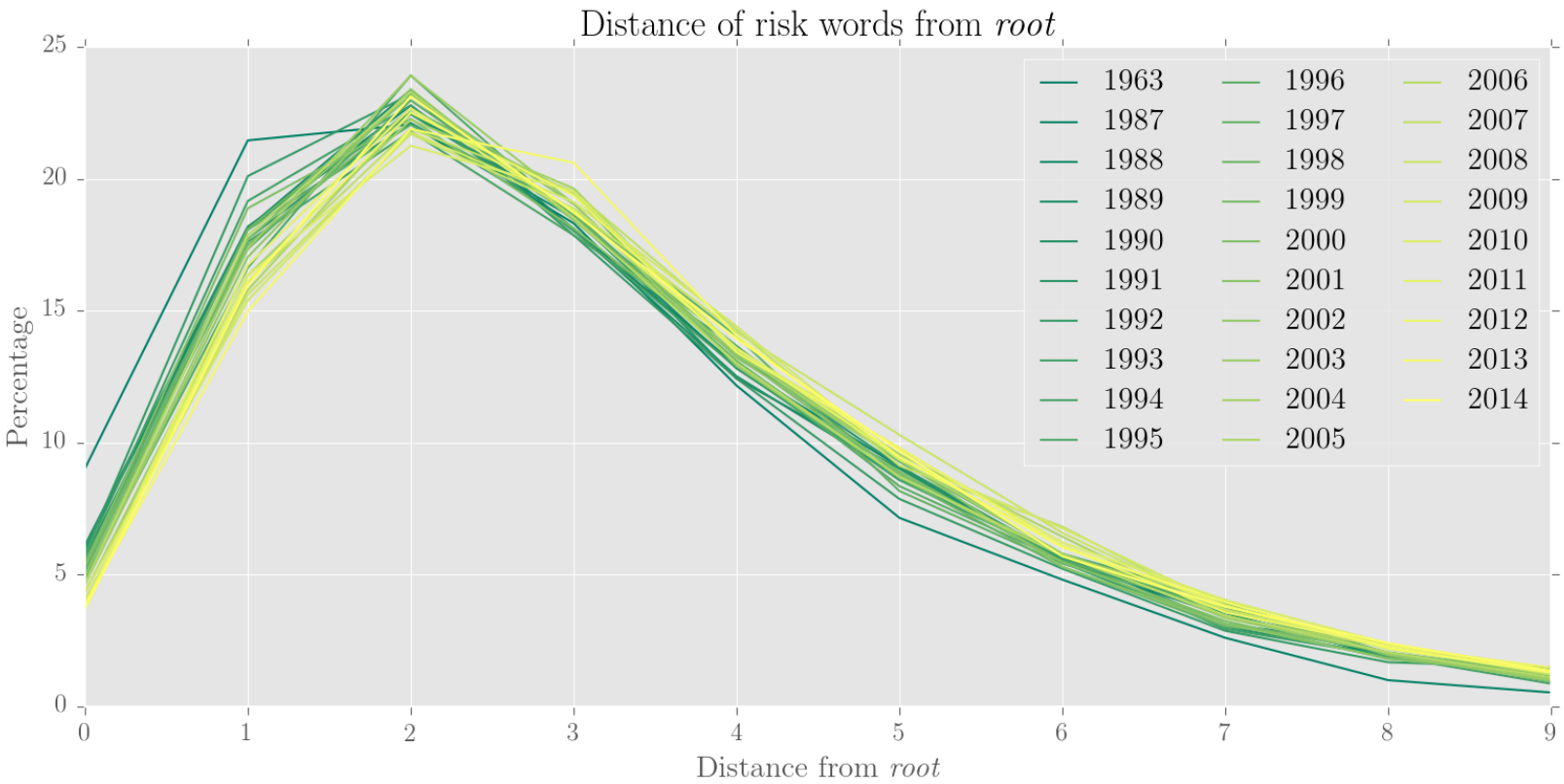
Risky decision → risk assessment

Risk and power



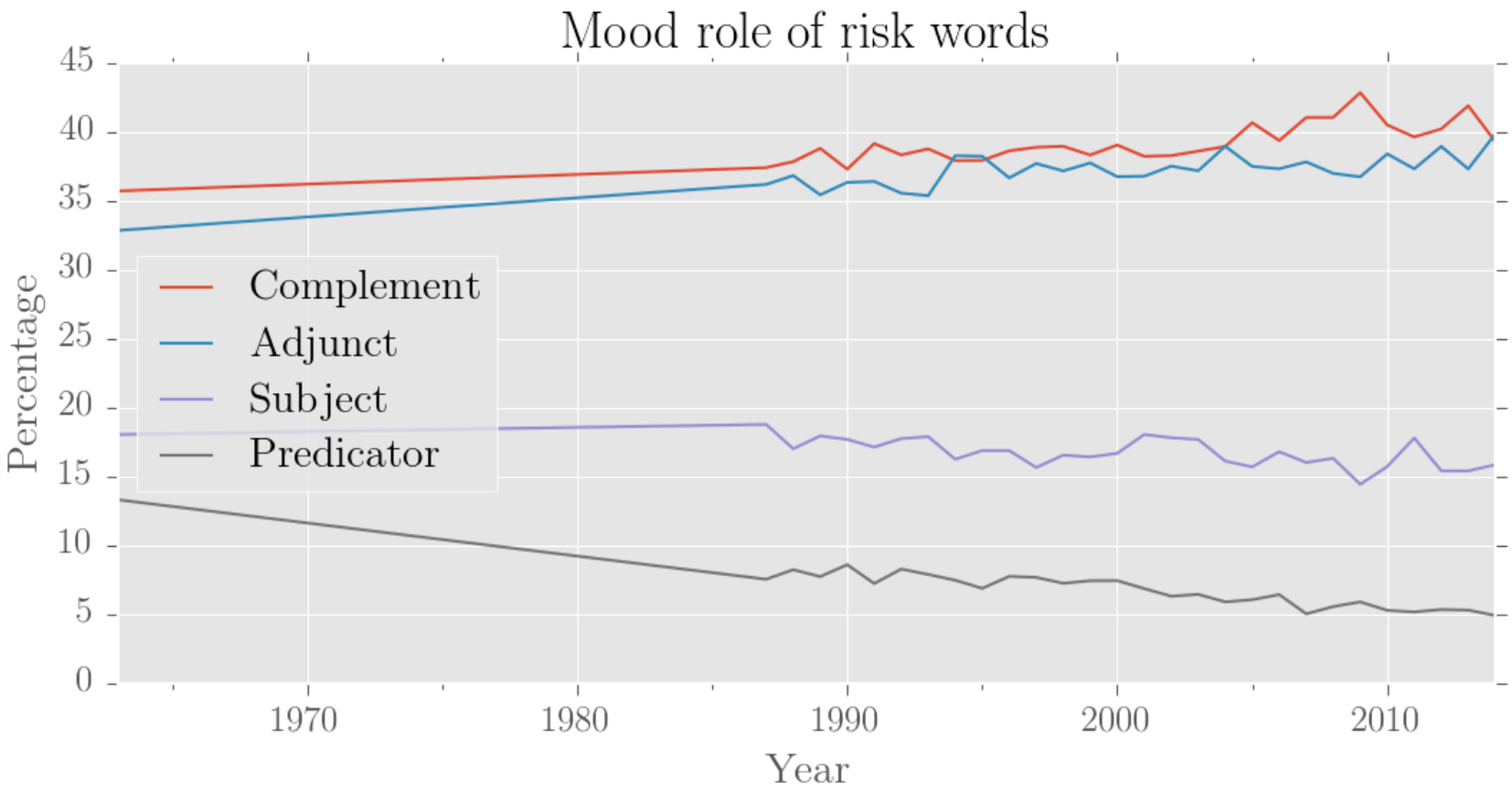
$\rightarrow$ Powerful and influential people *do* risking

Distance of risk word from *root*



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

Mood role of risk words (NYT)



Arguable → inarguable

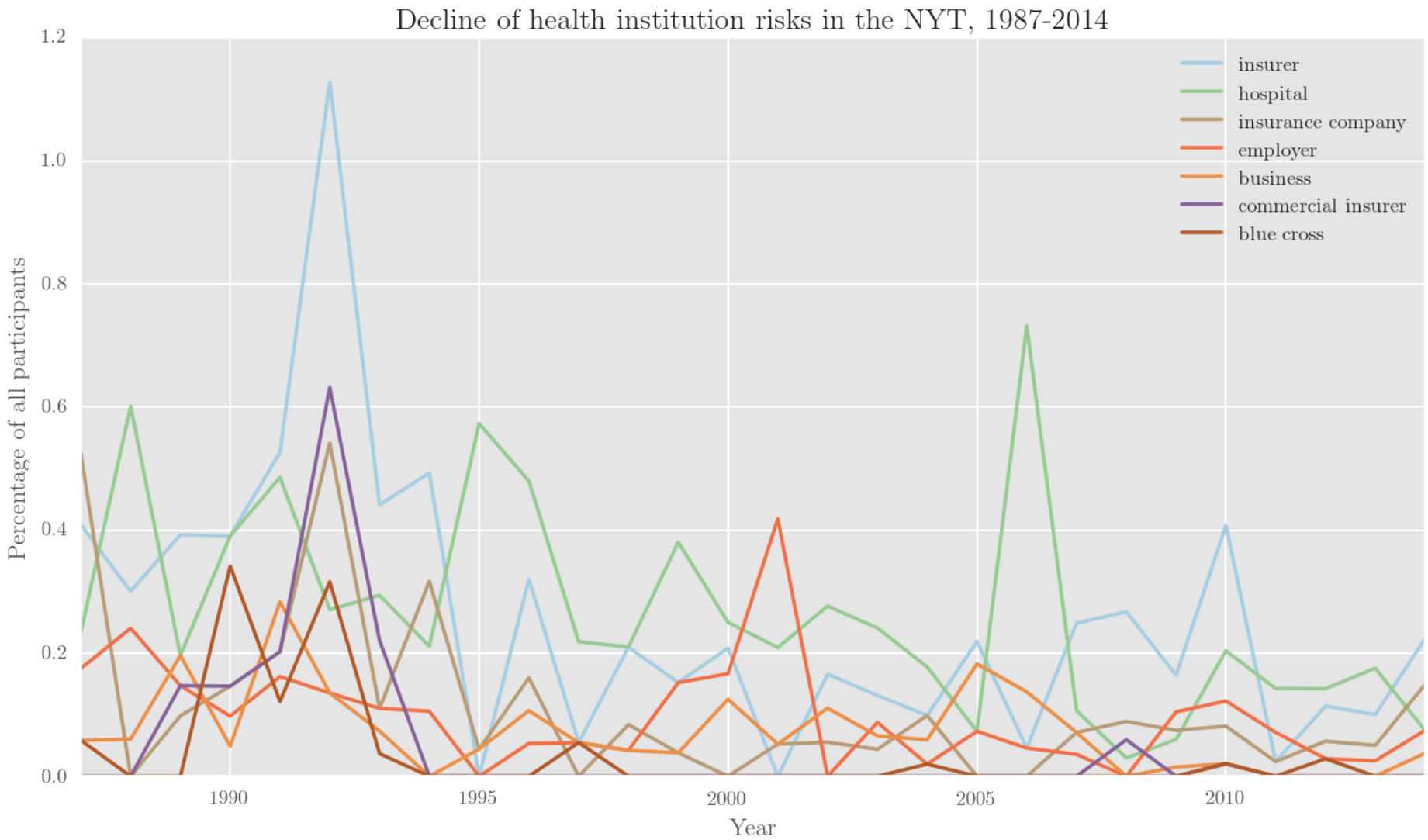
First investigation: key findings

- Nominalisation and *participantification*: 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

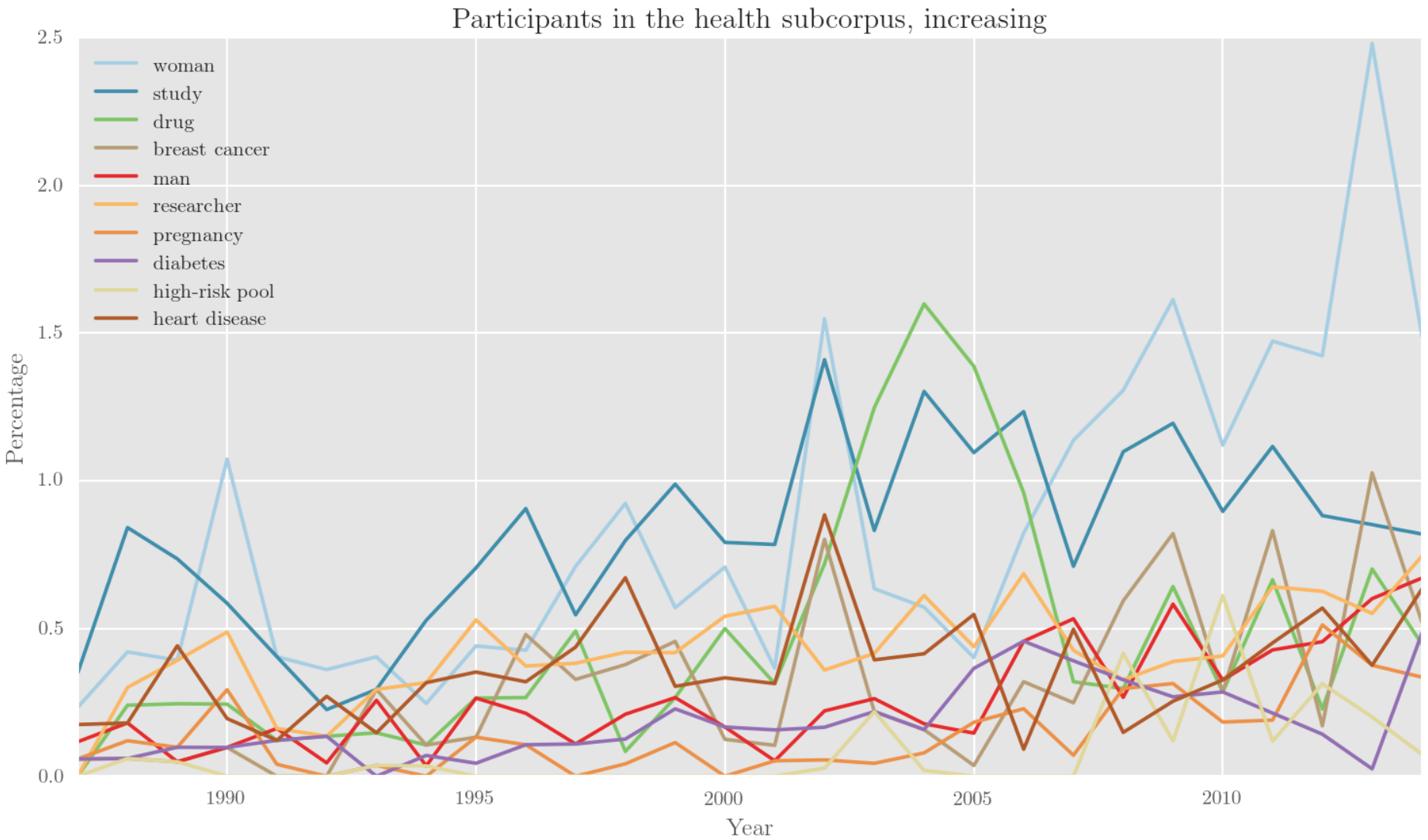
Health risk

- 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



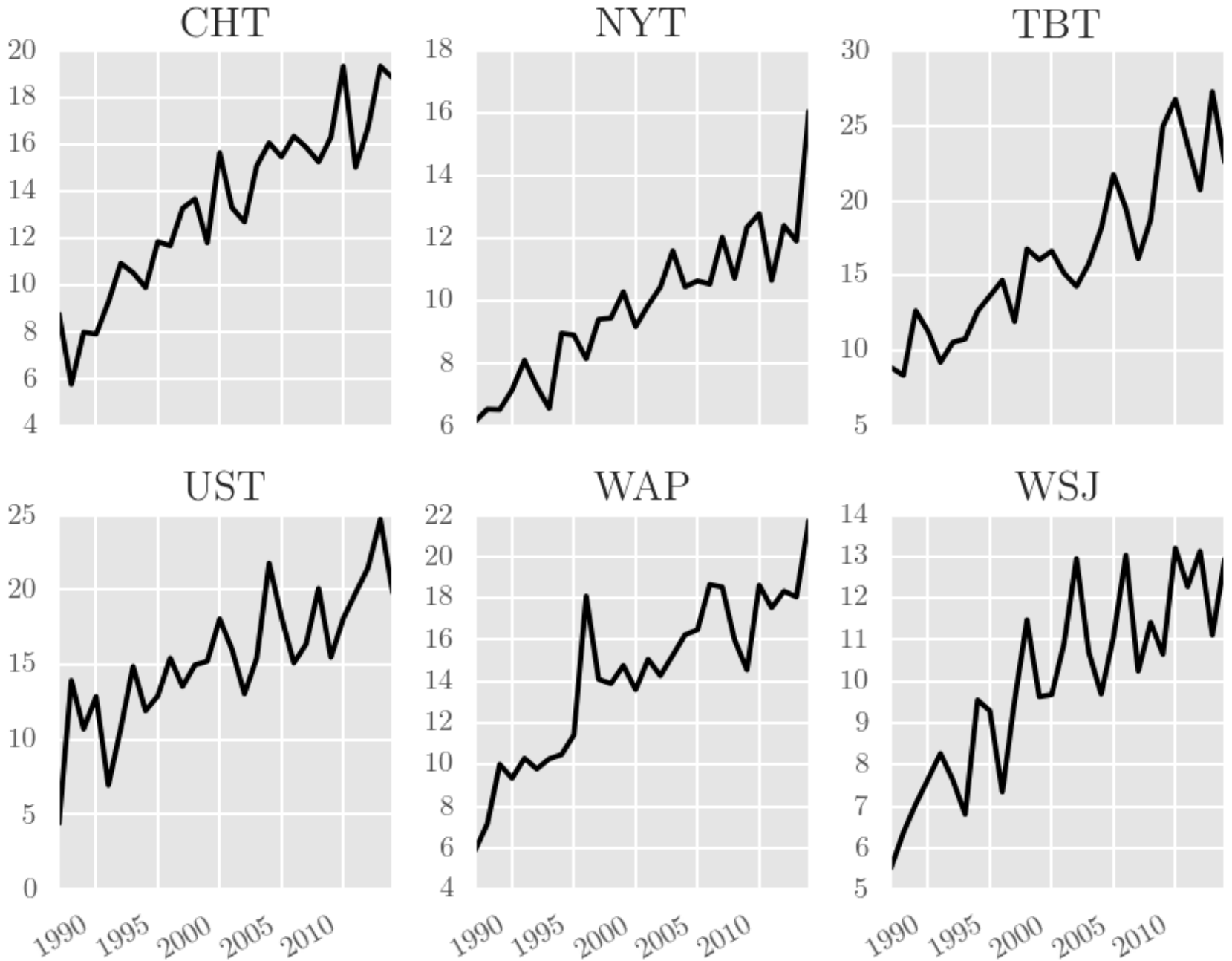
Increasing participants in health discourse



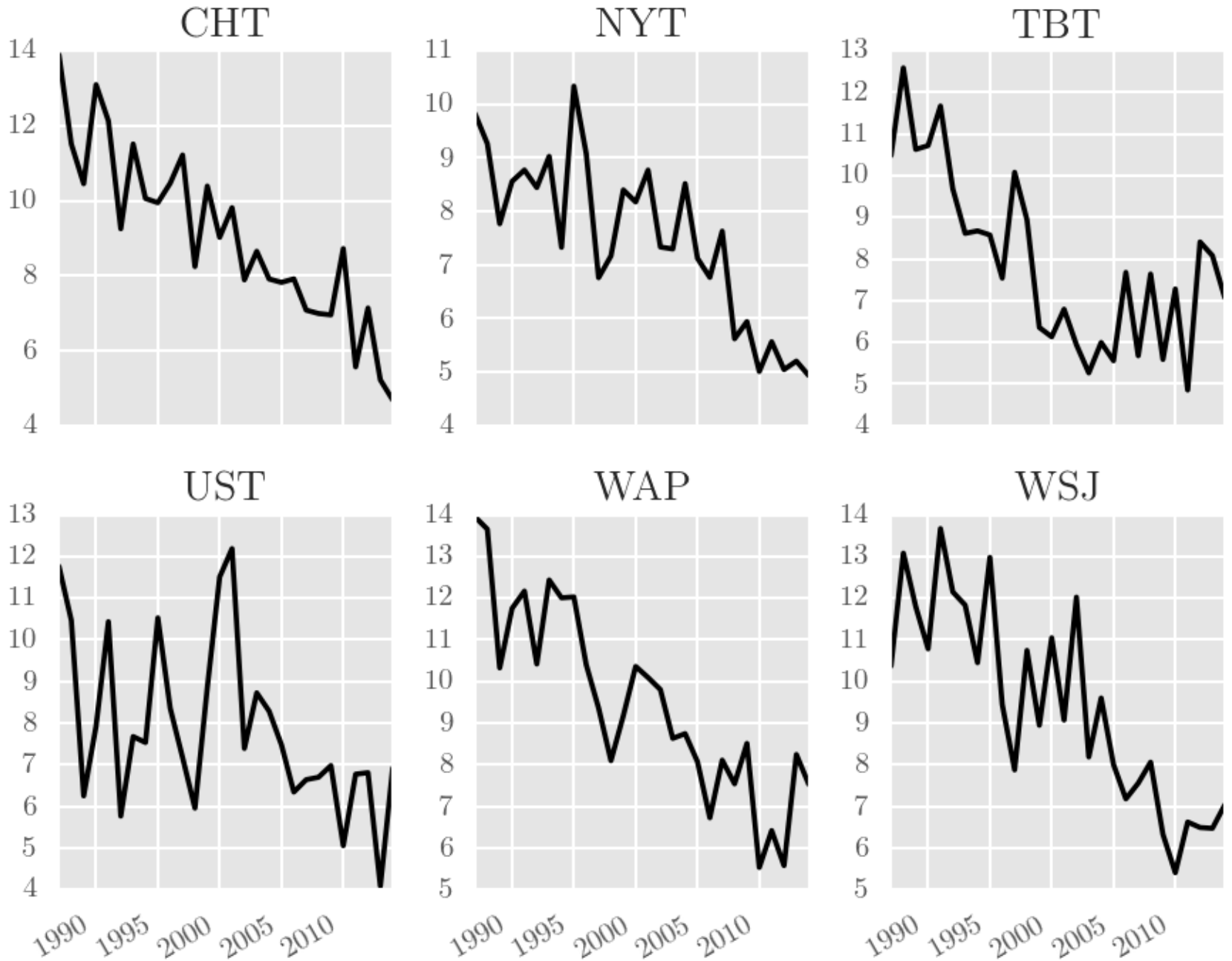
Six newspapers

- 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!

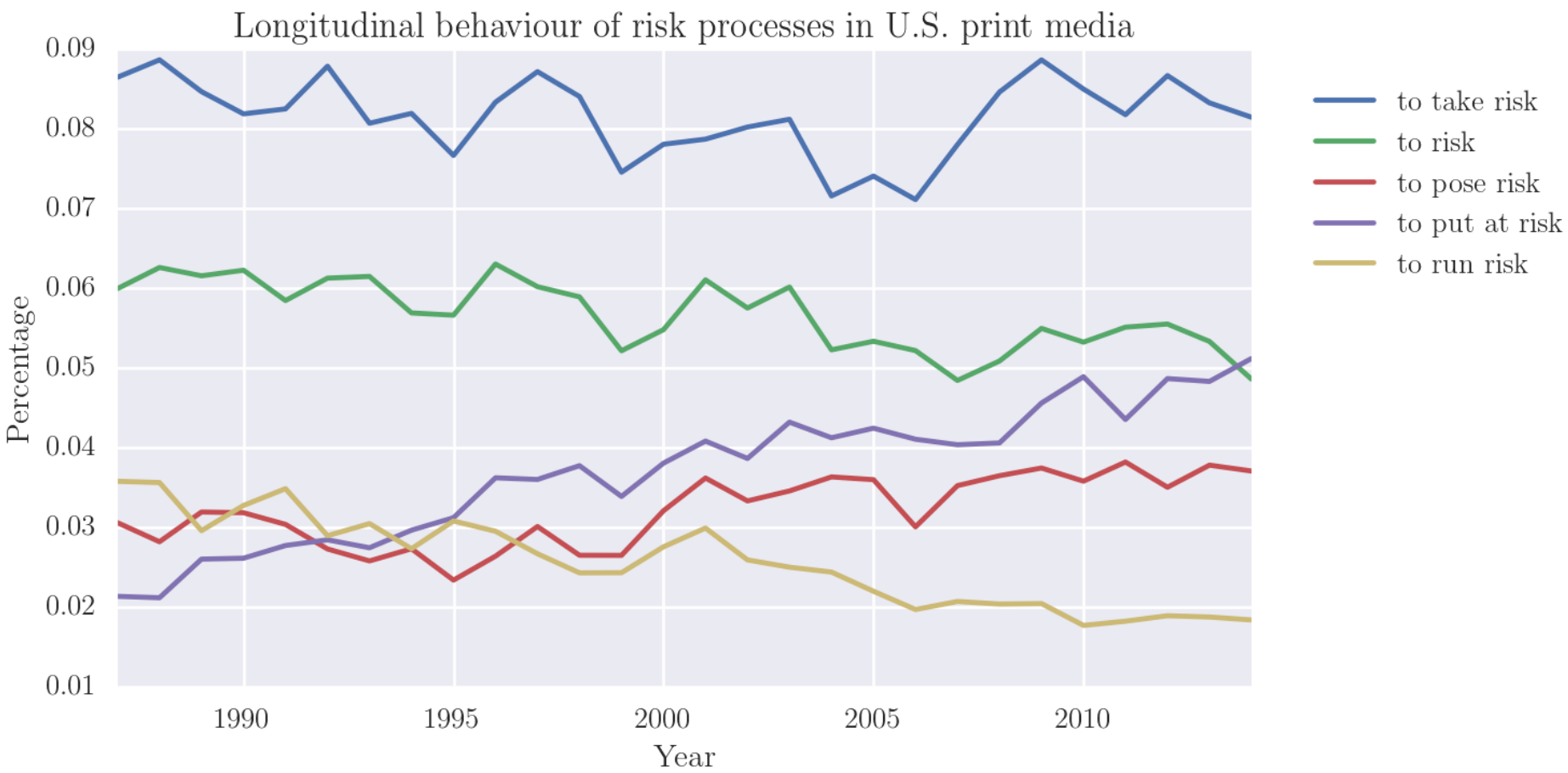
To put at risk



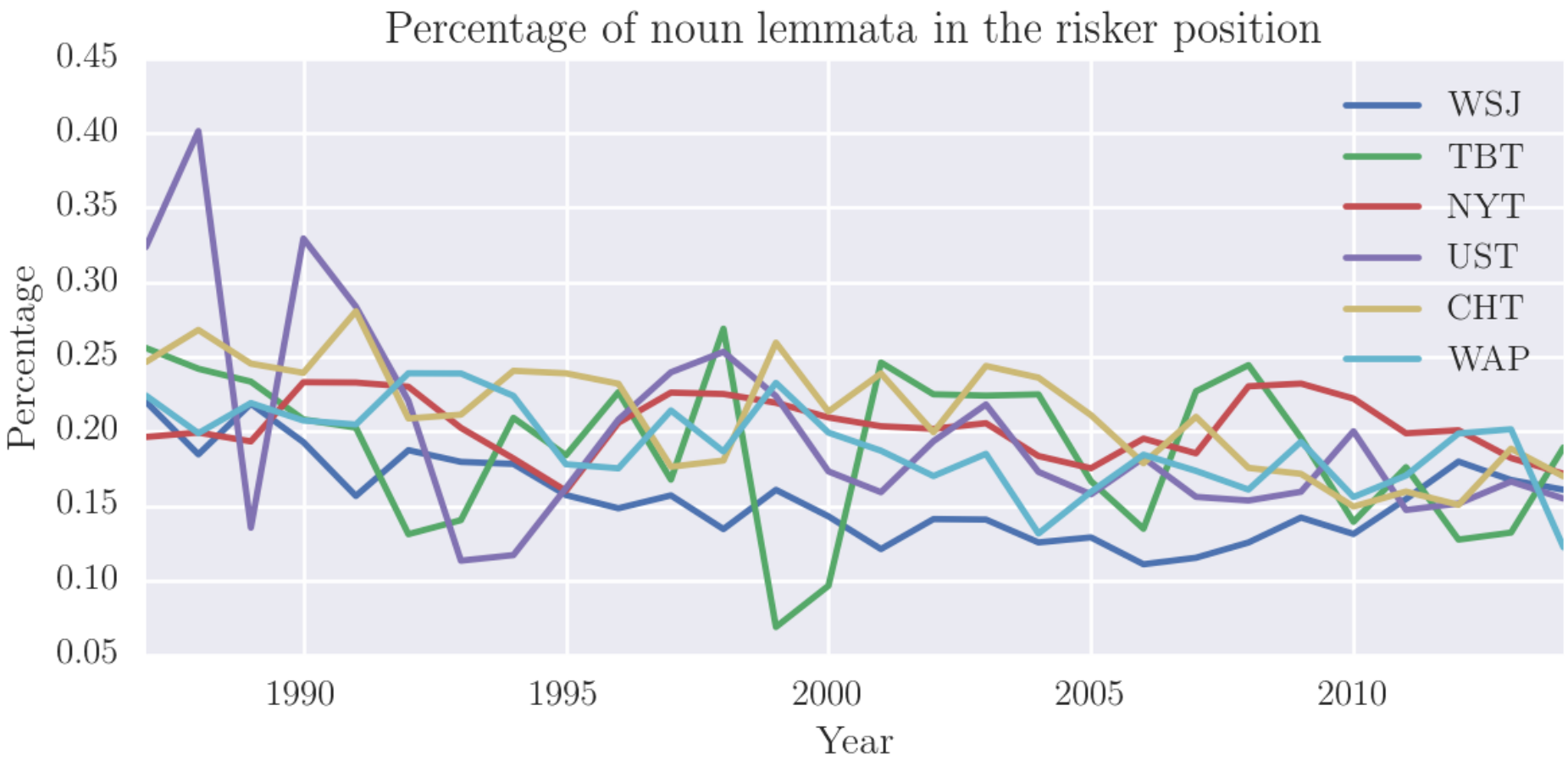
To run risk



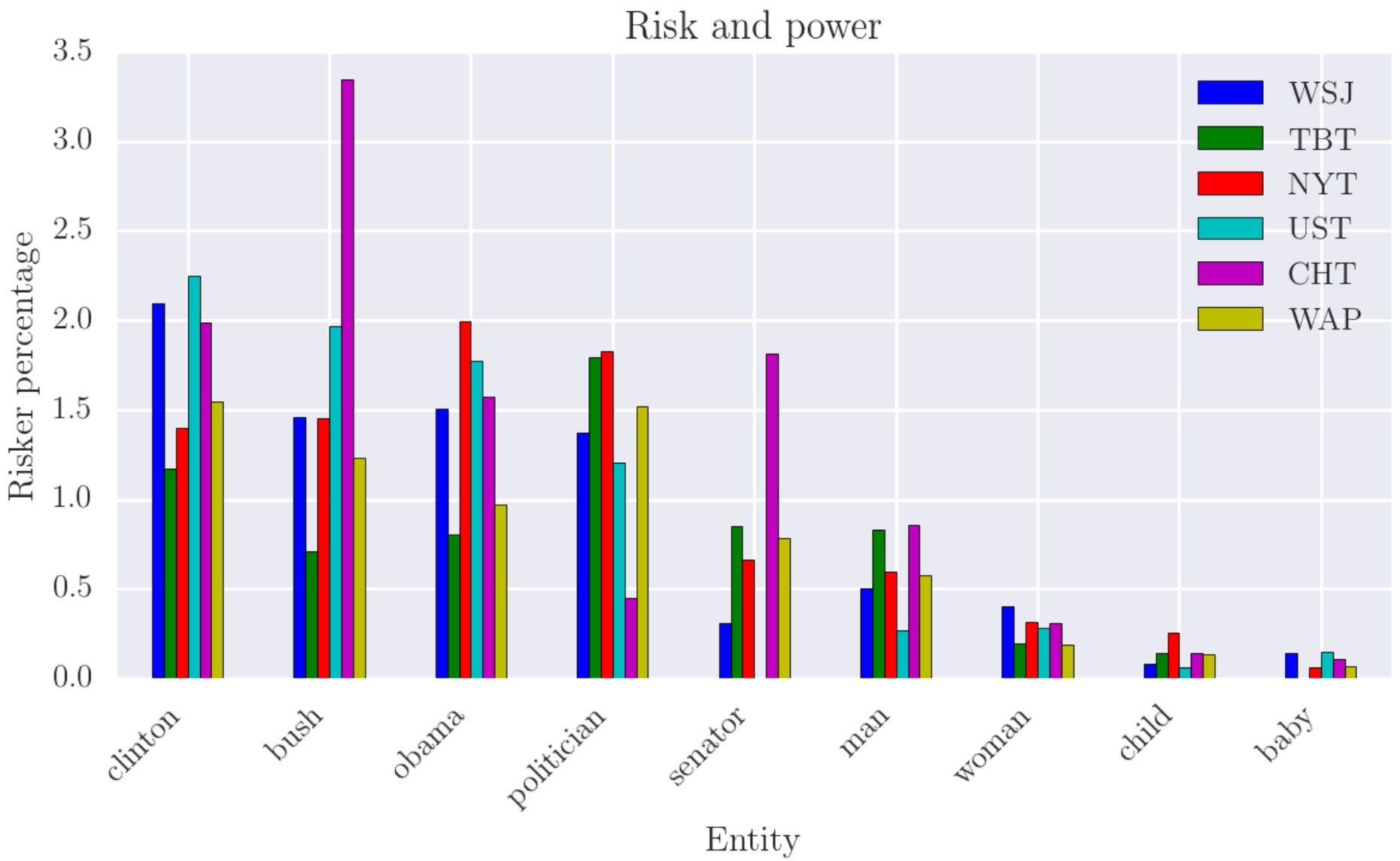
Risk processes



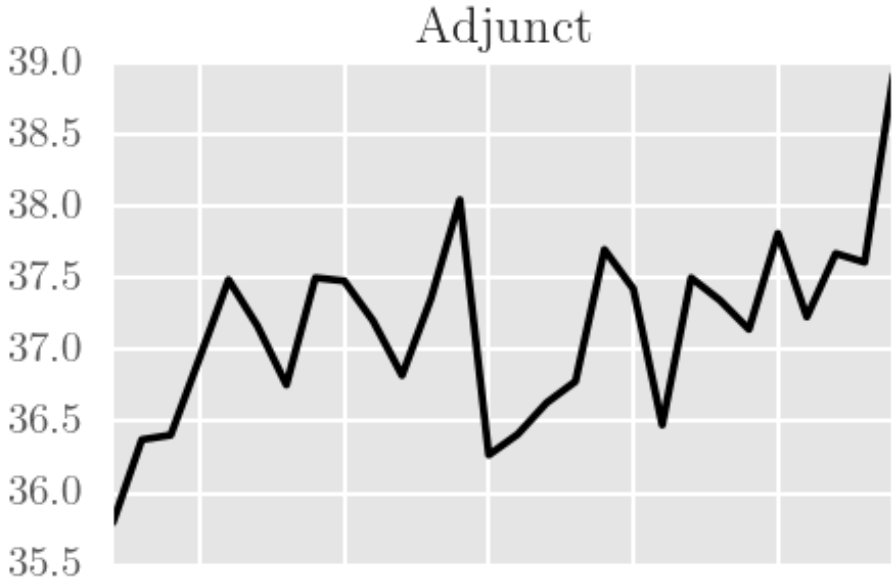
Less risking

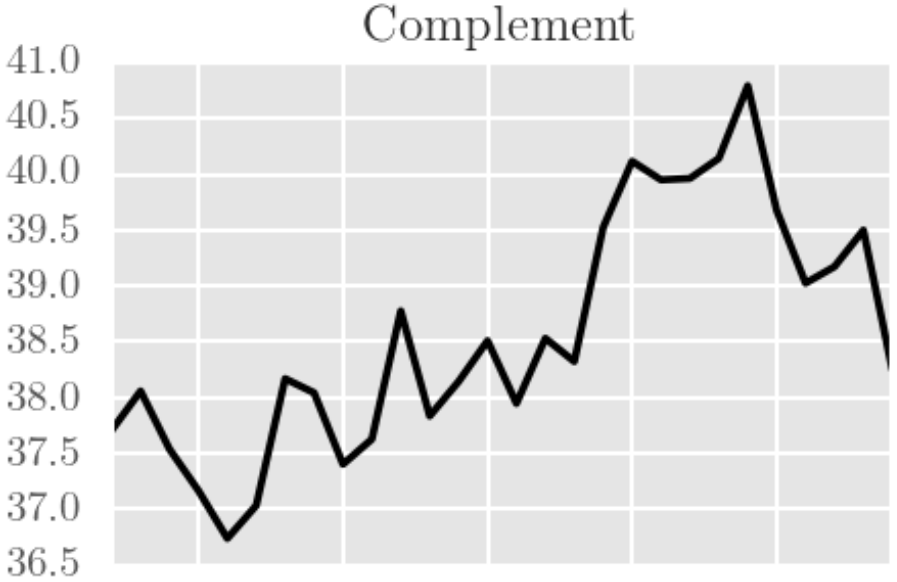


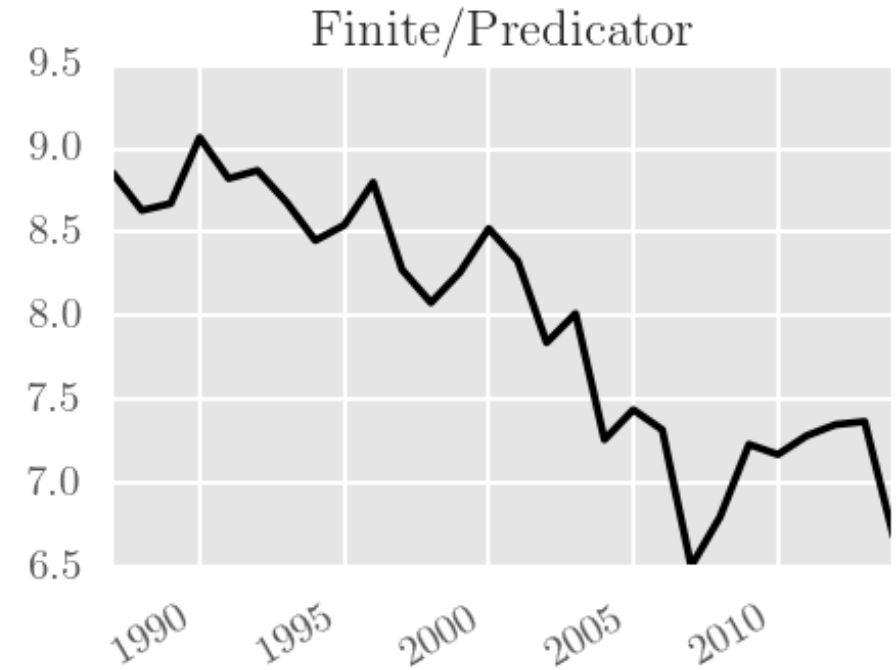
Risk and power across publications

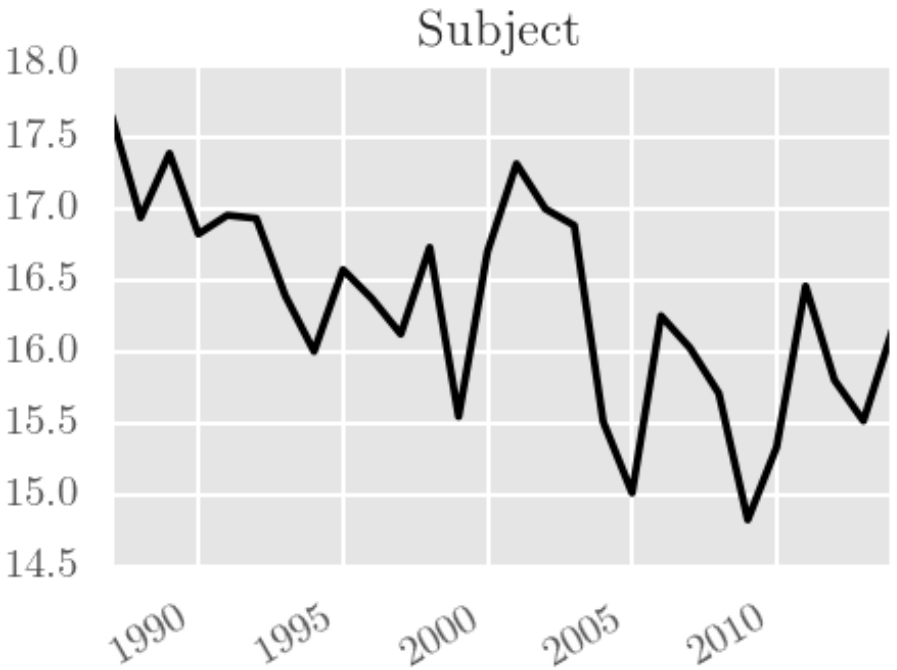


Mood role of risk words









Preliminary findings

- 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

Discussion: using SFL

THE UNIVERSITY OF MELBOURNE

- 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

Limitations

- 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

It's all open source

THE UNIVERSITY OF MELBOURNE

Data and tools are available for reuse:

- `https://www.github.com/interrogator/risk`
- `https://www.github.com/interrogator/corpkit`

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`

References I

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.

References II

Tulloch, J. C., & Zinn, J. O. (2011). Risk, health and the media.

Zinn, J. O. (2010). Risk as discourse: Interdisciplinary perspectives. *Critical Approaches to Discourse Analysis Across Disciplines*, *4*(2), 106–124.