Discourse-semantics of risk in the *New York Times*, 1963–2014: a corpus linguistic approach

Jens Zinn
jzinn@unimelb.edu.au

Daniel McDonald
mcdonaldd@unimelb.edu.au

University of Melbourne, Australia

February 21, 2015

Summary

In this report, we detail an investigation of risk words in the New York Times between 1963 and mid-2014. The investigation involves the creation of an annotated corpus of over 150,000 risk tokens and their co-text. Purpose-built functions for manipulating this dataset and visualising results were created and used to investigate the corpus according to a systemic-functional conceptualisation of the transitivity system.

Following the corpus interrogation, we attempt to use functional linguistics and sociological risk theory in tandem to analyse the findings. First, SFL is used to link lexicogrammatical phenomena to discourse-semantic meaning of the texts. Longitudinal changes in risk language are then mapped to key events, as well as broader social movements.

Contents

Chapter 1

Case study: the New York Times, 1963–2014

Our investigation centred on digitised texts from *New York Times* editions in 1963 and between 1987–2014. These texts (defined here as individual, complete chunks of content) are predominantly news articles, but depending on archiving practices, also included in our corpus is text-based advertising, box scores, lists, classifieds, letters to the editor, and so on. More specifically, we were interested in any containing at least one 'risk word'—any lexical item whose root is risk (*risking*, *risky*, *riskers*, etc.) or any adjective or adverb containing this root (e.g. *at-risk*, *risk-laden*, *no-risk*).[1]

We relied on two sources for our data. The *New York Times Annotated Corpus* (Sandhaus, 2008) was used as the source for all articles published between 1987–2006. ProQuest was used to search for and download articles containing a risk word from 2007–2014, alongside some metadata, in HTML format. We also created a subcorpus of articles from NYT 1963 editions through optimal character recognition (OCR) of PDF documents archived by ProQuest as containing a risk word in either metadata (i.e. title, lede) or content. Due to the time-intensive nature of manual correction of OCR, a random sample of one-third (1218 texts) was selected, with paragraphs of texts containing a risk word being manually corrected by hand.

Article text and any available metadata were extracted from this unstructured source content using *Python*'s *Beautiful Soup* module and *Shell* scripting, and added to uniquely named text files in annual subfolders. The kinds of metadata available varied according to the data source: The *New York Times Annotated Corpus* provides a number of potentially valuable metadata fields, such as author, newspaper section, and subject (manually added by trained archivists). We then value-added to this partially annotated corpus in three main ways. First, keywords and clusters for each article were calculated using *Spindle* (see Puerto, 2012) and added as metadata fields. Second, *MALLET* (see McCallum, 2002), a topic modelling tool, used LDA to algorithmically assign 'topics' to each article. The topics and their strengths were added as a metadata field. Finally, we used the *Stanford CoreNLP suite* (see Manning et al., 2014) to parse each risk token and its co-text for grammatical structure and dependencies.[2]

A key strength of the methodology is that subcorpora based on article or metadata attributes can can be easily created and compared. Our interest was in creating a small set of topic-specific corpora in order to look for changes in risk word behaviour within specific fields of discourse. As a case study, we decided to focus on three broadly defined topics: *economy*, *health* and *politics*. Librarian-added metadata concerning article topic/category (MC metadata field) was used to locate all articles tagged

Tag	Content
MA	Author(s)
MC	Librarian-added category tags
MD	Date of publication
MI	Unique identifier
MK	MALLET topic
MM	Manually annotated topic
MP	Section of newspaper
MS	Risk concordance line
MT	Article title
MU	URL for article
MZ	Annotator comment(s)

Table 1.1: Metadata tags and content

```
<MY>92 0.14 71 0.12</M>
<MV>13 0.26 96 0.21</M>
<MG>11 0.29 3 0.20</M>
<MO>28 0.33 21 0.24</M>
<MS>One family has lost a child and others may be at risk from a deadly brain
    inflammation, officials warned yesterday</M>
<MJ>center: 45.444118, officials: 28.536198</M>
<MT>New Jersey Daily Briefing;Meningitis Warning Issued</M>
<MC>MENINGITIS</M>
<MU>http://query.nytimes.com/gst/fullpage.html?res=9B06EFDA1239F933A05751C1A963958260</M
    >
<MF>0819209.xml</M>
<MA>KELLER, SUSAN JO</M>
<MD>1995-12-30</M>
One family has lost a child and others may be at risk from a deadly brain inflammation,
    officials warned yesterday. Bacterial meningitis recently killed a baby who attended
     the Center day-care program, officials say. They are urging parents and staff at
    the Center to contact their doctors or a hospital emergency room.
```

Figure 1.1: Example file: NYT-1995-12-30-10.txt

case-insensitive regular expressions `\beconom.*`, `\bhealth.*` or `\bpolitic.*`.[3]

We used some of the metadata fields to identify and remove listings (of best-selling books, plays, TV guides, etc.). Reasons for this were threefold. First, the jargon, abbreviations and non-clausal nature of listing language was not handled well by the parser. Second, list content was often repeated verbatim in multiple files, potentially skewing counts. Third, our two data sources archived listings in different ways.

Listings were located by querying metadata fields in a number of ways. Files with titles such as *Spare Times*, *Best Sellers*, articles with keywords such as 'theater', 'listing', or days of the week. If a file contained only a listing, the file was removed. If a risk word appeared only within the list portion of an article, the file was deleted. If a file contained both a body and listing, only the listing was removed.

After all data processing, we had a 150 million word corpus of nearly 150,000 articles containing a risk word published in the NYT or `NYT.com` in 1963, and between 1987 and mid 2014. The corpus had 29 annual subcorpora. The three subcorpora of economics, health and politics articles contained a subset of these articles. A breakdown of the size and composition of each annual subcorpus is provided in Table 1.2. Where necessary, frequency counts in the 1963 subcorpus were multiplied by four, to account for the smaller sample size. Frequency counts for 2014 were multiplied by 1.37 to fill in the uncaptured period between August 18–December 31.

Annual subcorpora	**Subcorpus**	**Words**	**Articles**	**Risk words**
	1963	83,188*	1218	1,584
	1987	4,885,883	4,878	7,690
	1988	4,834,791	4,703	7,430
	1989	5,059,517	4,997	7,810
	1990	5,416,187	5,250	8,244
	1991	4,748,975	4,774	7,493
	1992	4,923,509	4,818	7,329
	1993	4,686,181	4,615	7,330
	1994	4,857,729	4,762	7,384
	1995	5,130,206	5,150	7,834
	1996	4,969,911	4,773	7,257
	1997	5,121,088	4,759	7,318
	1998	6,085,810	5,437	8,351
	1999	6,053,731	5,392	8,248
	2000	6,472,727	5,717	8,434
	2001	6,603,456	5,902	8,722
	2002	6,865,631	6,423	10,288
	2003	6,795,591	6,481	10,066
	2004	6,776,200	6,215	9,989
	2005	6,722,240	6,191	10,031
	2006	6,722,592	6,278	9,965
	2007	4,757,290**	5,110	8,976
	2008	5,300,254	5,384	9,645
	2009	4,926,381	5,189	9,236
	2010	5,443,658	5,527	9,560
	2011	5,617,002	5,773	10,055
	2012	5,366,342	5,302	9,095
	2013	5,271,006	5,176	9,083
	2014	3,331,580	3,310	5,635
	Total	**153,828,656**	**149,504**	**240,082**
Topic subcorpora	**Subcorpus**	**Words**	**Articles**	**Risk words**
	Economics	10,489,137	8,286	32,448
	Health	8,524,023	6,944	36,547
	Politics	9,465,115	7,428	20,904
	Total	**28,478,275**	**22,658**	**89,899**

Table 1.2: Subcorpora, their wordcount, file count and number of risk words

* Only a small window of co-text—usually two sentences either side of the risk word—was preserved in this subcorpus, hence the smaller size of this sample.
* The drop in word-count here coincides with the switch from NYT Annotated Corpus to ProQuest as the datasource.

Chapter 2

Methodology

The challenge of making sense of enormous datasets is a formidable one, both at the practical level (the creation of scripts and search patterns, the transformation of search results into findings, etc), and at the more theoretical level of Big Data as both dataset and approach. *Big Data* approaches to social sciences and humanities research should be operationalised critically, with an acknowledgement that data size alone does not produce findings of higher truth or objectivity: automatic processing tools such as topic modellers and parsers do not provide perfect results, and their failures may often be buried within such large amounts of data.[4] Moreover, as boyd and Crawford (2012) note, even the imagination of phenomena as data itself constitutes an act of interpretation. There is also the potential for researchers to cherry-pick interesting or extreme examples from the set, rather than look for common patterns (Mautner, 2005). Finally, researchers must remain sensitive to the fact that the phenomenon under investigation (in this case, risk lexis) has been abstracted from its original multimodal context (as a component on a page in a daily paper).

To cope with these concerns in the context of natural language Big Data, we drew upon systemic functional linguistics (SFL) as a theory of language. SFL informed our study in two main respects: first, we relied on its conceptualisation of the stratal relationship between instantiated wordings in texts, their discourse-semantic functions, and the context they both respond to and construct; second, the systemic functional grammar (SFG) guided our attempt to locate specific sites of lexicogrammatical change in clauses containing one or more risk words.

1. A systemic-functional conceptualisation of language

SFL, as developed by Michael Halliday (see Halliday & Matthiessen, 2004) treats language as sign-system from which users select meanings for the purpose of achieving meaningful social functions. Inspired by the anthropological work of Malinowski, SFL divides the social functions of language into three realms of meaning: **interpersonal meanings**, which construct and negotiate role-relationships between speakers; **experiential meanings**, which communicate doings and happenings in the world; and **textual meanings**, which reflexively organise language into coherent, meaningful sequences.

One of the more radical dimensions of SFL is its inversion of the common discourse-analytic aim of analysing *texts in context*: in SFL, context is treated as being *contained within* instantiated texts—'context is in text' (Eggins, 2004). Based on the distribution of certain lexicogrammatical phenomena, we can accurately determine the overall genre/purpose of a text, even in highly decontextualised scenarios:

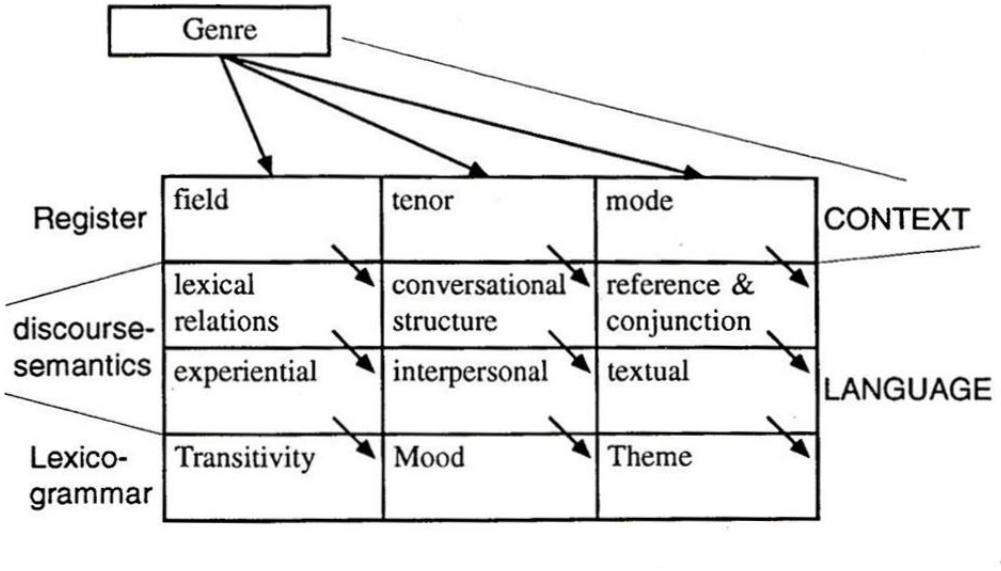

Figure 2.1: Strata and metafunctions of language (from Eggins, 2004)

'*Submissions must contain 3–5 references*' can be quickly identified as part of a set of instructions for an undergraduate assignment, based purely on its lexical (submissions, references) and grammatical (nominalisation, modalisation, etc.) properties. In the same way, Halliday conceptualises lexicogrammatical features of texts as probablistically determined by their context. That is to say, a given constellation of interpersonal, experiential and textual variables (e.g. the writing of a professor to undergraduates in a written course overview) will likely contain the kinds of lexicogrammatical features described in the example above (Halliday, 1991).

In SFL and its expansions (e.g. Martin, 1984; Christie & Martin, 2005), culturally recognised constellations of these three variables are treated as *genres*, within which other micro-genres may also be contained. In our case, the vast majority of texts under consideration are within the genre of newspaper article, with micro-genres such as sports-journalism, editorials, opinion articles and so on being differentiated by the appearance of different lexicogrammatical choices within both mood (i.e. use of interrogative mood, modalisation to connote subjectivity/objectivity) and transitivity systems (what is being spoken about).[5]

Three key factors informed our decision to adopt the SFL framework for our study. First, in contrast to most mainstream grammars, SFL conceptualises lexis and grammar as a different ends of the same system: lexis is the most delicate realisation of grammar (see Hasan, 1987). Such a conceptualisation, we believe, is vital to an investigation of the behaviour of a concept in a large text corpus, as much of this behaviour will indeed be grammatical. Accordingly, in this study, automated parsing of corpus texts is used to carry out (often simultaneous) searches of both grammar and lexis.

The second benefit of SFL to our research aims is that SFL is explicitly designed as a framework that to make it possible to say meaningful things about how real-world instances of language work to build meanings and perform social functions. It is thus an *appliable linguistics*, built to 'empower researchers to undertake projects of investigation and intervention in many contexts that are critical to the workings of communities and the quality of human life' (Matthiessen, 2013, p. 437).

Finally, SFL contains the best-articulated means of systematically connecting instantiated lexico-

grammatical units (i.e. wordings) to the more abstract stratum of discourse-semantics (i.e. meanings) (Eggins & Slade, 2004). On the strength of this link is the whole endeavour of corpus-discourse research predicated: absent a systematic connection of these two planes of abstraction, corpus-assisted discourse studies lose much of their explanatory power, and corpus-informed discourse research becomes a contradiction in terms.

2. Risk words and the systemic functional grammar

Perhaps the most laudable achievement of SFL is the ability of its grammar (admitted even by critics, e.g. Widdowson, 2008) to connect the three kinds of meanings to distinct components of lexicogrammar in consistent, stable ways. Interpersonal meanings are made through the **mood system**, including features such as *modality* and *modulation*. Textual meanings are made through the use of **systems of reference and conjunction** between and within clauses. Experiential meanings are made via the **transitivity system** (predicators, their subjects and object arguments, and adjuncts, in more mainstream grammars). This latter system is of most interest to us.[6] In SFL, transitivity analysis of a clause involves breaking it down into its *process*, *participants* and *circumstances*, realised by verbal groups, nominal groups and adverbials/prepositional phrases, respectively. Most central is the process, whose head (the rightmost verb in a verbal group), may be grouped into five types: **material processes** (doing and happening: *Risk declined*), **mental processes** (thinking: *She thought it risky*), **verbal processes** (saying: *We talked about the risks*), **existential processes** (*There are risks*) and **relational processes** (being and having: *It seemed risk-free*). Each type has different configurations of possible participants: mental processes have *Senser* and *Phenomenon* (the sensed); material processes generally have an *Actor*, in subject position, with optional participants such as *Goal*, *Range* and *Beneficiary*. Circumstances (e.g. '*this week*' in Figure 2.2) provide specifications such as the manner, extent or location of the process. Circumstances are more syntactically flexible, in that they are often able to be placed in a number of positions within the clause.

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

Figure 2.2: Transitivity analysis of a clause

An important caveat remains. SFL considers each kind of meaning as having a *congruent* realisation in the lexicogrammar—participants are congruently nominal; qualities as congruently adjectival. Aside from simply using native speaker intuition tests, SFL theorists argue that congruent forms often can be identified by their *typicality* and their *unmarkedness*: congruent realisations are expected to be more frequent in the language as a whole, and to involve fewer derivational morphemes (*nation* as a thing is less inflected than the quality, *national*) (Lassen, 2003). That said, as Halliday and Matthiessen (2004, p. ?) explain, 'it is by no means easy to decide what are metaphorical and what are congruent forms'. *Risk* is in itself a good example of a concept that straddles the terrain between participant, process and quality.

Incongruent choices, however, are also common in many kinds of texts, carrying a 'very considerable semantic load' (Halliday & Matthiessen, 2004, p. ?). First, through *grammatical metaphor*, semantic processes may be realised grammatically as participants ('I accepted *the invitation*') for the purpose of

Clause complex
Clause
Group/phrase
Word
Morpheme

Table 2.1: Rank Scale in SFL

packing more information into clauses—a key feature of written journalistic text (Simon-Vandenbergen, Ravelli, & Taverniers, 2003). Furthermore, similar meanings may be made at different ranks/strata of language: 'a good risk' and 'a risk is good' communicate the same positive appraisal of the same participant, but at different levels (group/phrase level via adjectival modification in the first example; clause level via relational ascription in the second). Incongruence poses serious challenges for corpus linguistic studies of discourse, as it limits our ability to locate, for example, all the ways in which risk is evaluated, graded or judged. This issue is exacerbated if, in line with SFL theory, we consider all lexicogrammatical choices to be meaningful and purposive, including the author's decision to invoke an incongruent form (as in Eggins, 2004). In some cases, rank-shifted meanings may be found using increasingly complicated lexicogrammatical search queries (see Figure 2.4 for an example). Automatic location of some other cases remain at this point beyond our capabilities: in appraisal at the level of clause-complex ('*I see a risk—it's a big one*') extremely complex grammatical searches would be needed to first recover the identity of *it* and *one* as *a risk*, before we could automatically determine that the risk is being semantically modified by *big*. Accordingly, our analysis is limited to group/phrase and clausal levels, with meanings made via the clause complex excluded.

2.1. Risk and the experiential metafunction

We situate our analysis of risk words predominantly within the experiential realm of meaning. At the most abstracted level of this dimension of language, we are interested in changes in the field of discourse in which risk as a concept is instantiated: *has risk shifted, as per key claims of sociological theory, from international relations toward population health?* Then, within these fields, we are interested in the constellations of happenings in which risk may play a role: *when risk is a process, what participants are involved? When risk is a participant, what is it a participant in, and with whom? And when risk is part of a modifier, what kind of participants and processes does it modify, and how?* Through categorisation of the kinds of fields in which risk appears, as well as the kind of participants who are positioned as riskers, risked things and potential harms, we can then empirically test the claims of influential sociological examination of risk discourse (See Table 2.2).

Either this needs to be expanded, or the mood description contracted...

2.2. Risk and the interpersonal function: arguability

Though our analysis is for the most part concerned with experiential meanings (via the Transitivity system), some aspects of interpersonal meanings (via the Mood system) are also relevant. Accordingly, a brief sketch of the mood system is required.

In SFL, the Mood system is used to give and request information (semiotic commodities) or goods and services (material commodities). Congruently, interrogatives request information, and imperatives request goods and services. Declaratives provide information. Being by far the most common mood

type in news discourse, our analysis is focussed on the structure of the declarative. A declarative clause contains a Mood Block, which contains a Subject and Finite (see Figure 2.3). Locating the constituents of the Mood Block is simple: if a tag question is added to this declarative (*the bang . . . can hold risks . . . , can't it?*), the tag picks up the Subject and the Finite (with polarity reversed).

Modality, also a component of the interpersonal metafunction, concerns modification of propositions with speaker judgements.[7] Prototypically, Modality is expressed through modal auxiliaries in the Finite position (*I can/should/might go*). Through Modality, speakers 'construe the region of uncertainty between yes and no' (Halliday & Matthiessen, 2004, p. 147). In Figure 2.3, for example, *hold* is modalised through *can* in order to express the author's judgement as to the possibility of the banging of the gavel holding risks.

But	*the bang of the gavel*	*can*	*hold*	*risk*	*for novices*
	Subject	Finite	Predicator	Complement	Adjunct
	MOOD		RESIDUE		

Figure 2.3: Mood analysis of a clause

At a greater level of abstraction, these Mood and Modality choices are responsible for the construction of role relationships between interactants: where interactants are of equal status (i.e. friends chatting at a cafe), similar overall frequencies in mood choices for each interactant may be observed. In a situation with interactants of less equal status, mood choice frequencies may vary more widely for the different participants: in a typical interaction between a professor and an undergraduate, only the professor is likely to use imperatives to issue commands. Importantly, as with experiential meanings, incongruence may occur, though the motivation for incongruence is an interpersonal one, such as politeness or face saving (*Shut the door!/Could you shut the door?*). For us, however, this kind of incongruence does not pose the same level of challenge as experiential incongruence, as print news journalism as a genre rarely commands or requests information from the reader, and as the faces of both writer and reader are rarely under threat.

We are interested in Mood mostly because Mood is the system through which *arguability* of propositions is mediated. In SFL, arguability is used to denote the relative ease of challenging or refuting a proposition, and thus, the level of implicitness of a meaning made about the world.

Chiefly, arguability rests in the two components in the Mood Block—the Finite and the Subject. To make a proposition arguable, it must be grounded in time and space, or to a speaker judgement of its validity. These are the two potential functions of the Finite. Locating a proposition within time and space is done through adding primary tense (*lives were risked*). Meanings are linked to speaker judgements through modality (*lives might be risked*) (Halliday & Matthiessen, 2004, p. 116). In either case, the Finite grounds the proposition with reference to the current exchange being undertaken by the interactants. Primary tense situates a proposition according to what is present at the time the utterance is made—it indicates 'the time relative to now' (Halliday & Matthiessen, 2004, p. 116). Modality either expresses an assessment of the validity (probability, certainty, obligation, etc.) of a proposition (*it might/will/must happen*) or, in an interrogative, invites the addressee to make this assessment (*might/will/must it happen?*).

The Subject is the second component of arguability. Semantically, SFL treats the Subject as 'something by reference to which the proposition can be affirmed or denied' (Halliday & Matthiessen, 2004, p. 117). In the contexts of proposals and commands, it is the one who is supposed to perform the action

Metafunctions		**Lexicogrammar**		
Experiential	*Interpersonal*	*Rank scale*	*Group structure*	*Arguability*
Process	Finite, Subject	Clause	Head	**Higher**
Participant(s)	Complement	Group/phrase	Modifier	**Medium**
Circumstances	Adjunct	Word	Submodifier	**Lower**

(*Shut the door, will you?/I'll speak to her, shall I?*). In the case of declarative information provision, the Subject is the thing upon propositional validity rests. In *the bang of the gavel can hold risk for novices*, for example, a refutation still requires a coherent Subject and Finite, while the Residue is only required if it is the challenged component:

1. No, *it should* hold risks (refuting modal finite/speaker judgement)
2. No, but *a handshake can* (refuting Subject)
3. No, but *it can* hold excitement (refuting Complement)
4. No, but *it can* for experts (refuting Complement)

Thus, the Mood Block is the most arguable part of a proposition—'it carries the burden of the clause as an interactive event' (Halliday & Matthiessen, 2004, p. 118). The steps an interlocutor needs to take to deny the validity of a meaning are fewest when the disagreement concerns the composition of the Mood Block. Meanings made within Complements and Adjuncts, or within groups or phrases, are more implicit: they support, rather than enact, meanings made within the Mood Block (Matthiessen, 2002).

With these descriptions, we can now operationalise arguability with regard to experiential and interpersonal metafunctions, as well as the internal structure of groups and phrases.

In the context of risk words, this conceptualisation of arguability can be used to empirically examine key sociological claims. Increasing prevalence of risk words generally would mean that risk words have an inbound trajectory in the NYT generally. Increasing risk words within the Mood Block would indicate that risk is discussed and argued about. A shift from Mood Block to Residue risk would indicate greater implicitness and inarguability of risk. At the same time, risk words as heads of groups/phrases would indicate greater discussion of risk, while risk words as modifiers would indicate implicitness.

3. SFL and corpus linguistics

Methodologically, our study may be characterised as an attempt to combine the systemic functional conceptualisation of language with practices from diachronic corpus linguistic (CL) research. As Hunston (2013) notes, SFL and CL share a number of underlying similarities, such as an emphasis on natural language a focus on register/genre as shaping the lexicogrammatical choices made in texts. More fundamentally, both CL and SFL posit that we can learn about these texts through quantification of their various lexical, grammatical and semantic properties.

We use SFL and CL in tandem to locate patterns in texts without manual interpretation or categorisation. Sociological insights into key events and movements are then mapped at later stages to observed lexicogrammatical and discourse-semantic change in the behaviour of risk words (challenges in balancing the systemic-functional notion of context-in-text with the use of sociological methods are discussed below). Such an approach is characteristic of the emerging field of *corpus-assisted discourse studies* (CADS). The oft-noted 'methodological synergy' of CL and discourse analysis allows researchers a greater degree of empirical and quantitative support for claims, as well as a larger body of examples that can easily be accessed and qualitatively analysed (Baker et al., 2008). In terms of risk, corpus-based methods allow an empirical testing of sociological literature that has tended to invent examples of clauses containing risk words, despite there being little evidence that these phrases are commonly instantiated

in general language use (Hamilton, Adolphs, & Nerlich, 2007). Research has also tended to conflate risk words with the concept of risk itself, even though the word may not be critical to the experiential meaning of a clause (the *risk management team went for coffee*) and even though the latter is often present without the linguistic instantiation of the former.

Work within CADS varies chiefly in the extent to which the corpus itself is the focus of the investigation. In *corpus-driven* work, researchers are attempting to demonstrate that the corpus itself contains particular patterns of discourse. Theories are developed inductively according to patterns located in the data. *Corpus-informed* studies, on the other hand, may use the corpus as a body of examples that can be drawn upon in discussion of broader trends in society (Baker et al., 2008). Theories to be tested are developed before the corpus interrogation

Our study is in the latter domain.[8] As a diachronic investigation, we can further situate our method within *Modern Diachronic CADS*. As Partington explains,

> [MD-CADS] employs relatively large corpora of a parallel structure and content from different moments of contemporary time ... in order to track changes in modern language usage but also social, cultural and political changes as reflected in language (2010, p. 83).

As newspapers are well-structured and archived in digital collections, they have formed a common data-source for CADS. Johnson and Suhr (2003) investigated shifts in the discursive construction of *political correctness* in German newspapers. Duguid (2010) performed thematic categorisation of the keywords from two collections of digitised newspapers from 1995 and 2005. Freake and Mary (2012) focussed on the ideological positioning of French and English in Canadian newspapers.

Ours is not the first corpus-based study of risk. Most well-known is Fillmore and Atkins (1992), who studied the behaviour of risk as both noun and verb in a 25 million word corpus of American English. Ultimately, the authors' aims were lexicographic, rather than discourse-analytic, limiting the usefulness of the study's methods for our purposes. A second key point of difference is the small size and lack of structure of their corpus (though their research was a certainly remarkable and groundbreaking effort at the time of publication). Finally, their study was neither longitudinal, nor designed to connect patterns to social/societal change.

More recently, Hamilton et al. (2007) used a frame semantics approach to understand the behaviour of risk in two corpora: the 56 million word *Collins WordbanksOnline Corpus* (N risk tokens) and the five million word *CANCODE* (235 risk tokens). We depart from their methods in five respects. First, they use general corpora, while we used a specialised corpus. Second, our study is diachronic, while theirs is largely monochronic. Third, we differ dramatically in the number of risk words analysed (n/n). Fourth, they relied on collocation (without lemmatisation[9]), while we performed specific queries of the lexicogrammar, using lemmatisation where needed. Sixth, they used frame semantics, while we use SFL (though informed by Filmore and Atkins' (1992) articulation of the components of the risk frame). Though these theories have a number of underlying similarities (both are semantically oriented grammars, for example), the two diverge in their treatment of the role of cognition and psychology. While frame semantics argues that lexicogrammatical instantiations are mapped by listeners to preexisting cognitive frames or schemata, SFL is largely silent on the subject of cognition, preferring to map lexicogrammar to external variables of field, tenor and mode.

Notably, our methodology also departs from typical methods of (MD-)CADS in a few key respects. First, CADS is often lexically-oriented, with techniques such as **keywording** used as a means of disinterring the 'aboutness of a text' (Baker, 2004) and **clustering** and **collocation** used to look for the

`__ >># (/(NP|VP|PP)/ > (VP <<# "RelationalProcess" $ (@NP <<# /(?i)man/)))`	In relational processes in which *man/men* is the Token/Carrier, what is the head of the Value/Attribute?

Figure 2.4: *Tregex*-based search query and gloss

co-occurrence of lexical items absent any consideration of grammar. Hunston (2013) contends that despite a number of areas of overlap, SFL and CL are at odds in the sense that SFL is grammatically oriented while CL is lexically oriented. Though the majority of CADS does indeed focus on lexis, this preoccupation stems more from the relative simplicity of searching for tokens in corpora, compared to grammatical features, than it does from any theoretical motivation.[10] Accordingly, our use of grammatically parsed data and equal consideration of lexical and grammatical features, though in line with SFL, is against the grain of much contemporary CADS literature.

The second key difference from mainstream CADS is that we did not rely on typical practices such as keywording, clustering, collocation and the use of stopword lists. Our reasons for avoiding these practices are varied. Keywording we found to be problematic due to its reliance on a reference corpus of general language. The usefulness of this reference corpus is predicated on the idea of corpus balance—that is, the notion that a corpus of texts, if comprised of a wide variety of genres, and if the relative proportion of these texts is akin to their prevalence in culture, may be taken to be representative of language generally (Chen, Huang, Chang, & Hsu, 1996). As corpus balance is well-acknowledged by CADS practitioners to be only a theoretical ideal (Gries, 2009), we took a different approach. Rather than keywording, we simply counted the base forms of the most common heads of participants, processes and circumstances in each subcorpus. This also liberated us from the arbitrary nature of stopword lists (lists of very common words that are automatically excluded from search results), as most stopwords are determiners, prepositions, conjunctions and so on, which rarely occupy key experiential roles.

Clustering and collocation, though mainstays of CADS, are also absent in our analysis, as they consider only the co-occurrence of lexical items within a specified (and arbitrary) number of words, and accordingly do not take grammatical relationships into account. As an example, *Men are from Mars, and women are from Venus* would contribute to an understanding of *Mars* and *women* as collocates, regardless of the fact that the experiential meaning of the clause has the opposite meaning. We instead created nuanced search queries capable of drawing on lemma lists and lists of process types (as in Figure 2.4). This luxury was afforded by grammatical (phrase structure and dependency) annotation of the corpus, as well as the development of scripts for quickly searching lexicogrammar.

4. Discourse-semantic areas of interest

Our interest is ultimately in discourse-semantic experiential and interpersonal meanings of risk words. The first point of interest is simply the relative frequency of risk words in the NYT generally, and by word class. These areas of interest are at the clausal level. Within experiential meaning, we are interested the relative frequency of risk as a Participant and as a Process, as well as the behaviour of risk when occupying these roles. At the same time, we are interested in meanings made below clause level, within groups and phrases. When risk is a participant or process, we are interested in the ways it is modified. Furthermore, risk itself can be a modifier of participants and processes. Accordingly, we are also interested in both understanding the ways in which this modification happen and finding

the participants and processes that risk commonly modifies. Finally, within the interpersonal realm of meaning, we are interested in the arguability of risk words—that is, the extent to which their meaning is symbolically available to negotiation by the writer/reader.

We can summarise our discourse-semantic interests with the following 10 questions. *In terms of longitudinal change in the NYT,*

1. *How frequently do risk words appear?*
2. *Which experiential roles do risk words occupy?*
3. *Is risk more commonly in the position of experiential subject or experiential object?*
4. *What processes are involved when risk is a participant?*
5. *How are participant risks modified?*
6. *What kinds of risk processes are there, and what are their relative frequencies?*
7. *When risk is a process, what participants are involved?*
8. *When risk is a modifier, what are the most common forms?*
9. *When risk is a modifier, what is being modified?*
10. *How arguable is risk?*

These questions are answered in this order in the Findings section. In the Discussion, these answers are synergised in order to perform a broader analysis of discouse-semantic change.

5. Lexicogrammatical realisations of discourse-semantic meanings

Discourse-semantic meanings are realised in texts by lexicogrammatical patterns. **Risk as participant** is congruently realised by a risk word at the head of a noun phrase that is an argument of a main verb. Other possible realisations of risk participants are adjectival risk words in participant positions (*The job is risky*) or risk words within prepositional phrases (*Votes were at risk*). SFL also treats prepositional phrases as partially realised relational processes, containing only object arguments. As this is perhaps a controversial analysis within linguistic theory generally, the treatment of risk within PPs is separated from risk as arguments of verbal groups. **Risk as a process** is congruently realised by a risk word as the main verb of a clause. When risk is instantiated here, we can extract the participants involved in the process. **Risk as a modifier** is realised by different word classes, depending on what is being modified. Risk can modify participants through pre-head or post-head modification. Analysed in this study[11] are adjectival pre-head modification (*a risky move*), nominal pre-head modification (*risk management*) and post-head modification via a prepositional phrase (*the electorate at risk*). **Arguability of risk words** can be determined by looking for the functional role of risk words within the Mood system: risk as Subject or Predicator is more arguable than risk as Complement and Adjunct.

The scope of our project necessitated some constraints on the kinds of patterns we analysed. Major constraints included our focussing on experiential meaning, perhaps at the expense of interpersonal meaning. Thus, the analysis contains little consideration of how risk may be operationalised in order to construct writer/reader or newspaper/readership relationships. Also largely unanalysed are the ways in which risk are appraised, judged, and graded in severity. This was mostly due to the lack of available automatic parsers for SFL's appraisal grammar (see Martin & White, 2005). Finally, queries returning less salient or ambiguous results are omitted from discussion here. Counting the kinds of determiners that occur before a nominal risk (this risk, a risk, the risk) uncovered no particularly interesting patterns, for example.

Author	**Claim**	**Discourse-semantic realisations(s)**	**Congruent lexicogrammatical realisation(s)**	**Example(s)**
Beck	Everyday life world characterised by risk	Common people as increasingly common participants in risk as a process; increasingly localised risked things/potential harms; risk appearing in articles about daily life	Women, children, non-celebrities appearing as heads of nominal groups that are actors in risk processes; Everyday processes and common things appearing as head of verbal and nominal groups that appear after risk processes	*The little girl risked tearing her coat*; *We risked getting rained on/one dollar*;
Giddens				
Zinn				
Author x	Increasing focus on risk in health discourse	Medical lexis (diseases, institutions, medications, etc) as increasingly common participants in risk as a process	Illnesses and risk in nominal compound words; health terminology in modifiers of risk as head of a participant	*The cancer-risk*; *the risk of heart disease*
Author y				

Table 2.2: Systemic-functional realisations of sociological claims concerning risk discourse

6. Operationalising sociological claims

The discourse-semantic and lexicogrammatical areas of interest can be summarised with regard to related sociological claims.

Chapter 3

Findings

Findings are organised according to the formulation of areas of interest as questions. These questions progress from general frequency counting (Q1), through experiential meanings (Qs 2–7), to risk as modifier (Qs 8 & 9) and finally to arguability (Q10). Discussion of the general significance of individual findings is also presented in this section, as the Discussion section synergises all findings to explain the discourse-semantics of risk.

An *IPython Notebook* interface for navigating the corpus (see McKinney, 2012), as well as the code used to interrogate it and the findings we produced, is available online: `https://github.com/interrogator/sfl_corpling`. A non-interactive version is available at `http://nbviewer.ipython.org/github/interrogator/sfl_corpling/blob/master/risk.ipynb`

1. How frequently do risk words appear?

The first point of interest was the overall frequency of risk words in the NYT and the distribution of risk words by word class (nominal, verbal, adjectival/adverbial), absent any consideration of surrounding grammar (see Figure 3.1). We found a trend toward nominal forms, with the other categories remaining more or less stable.

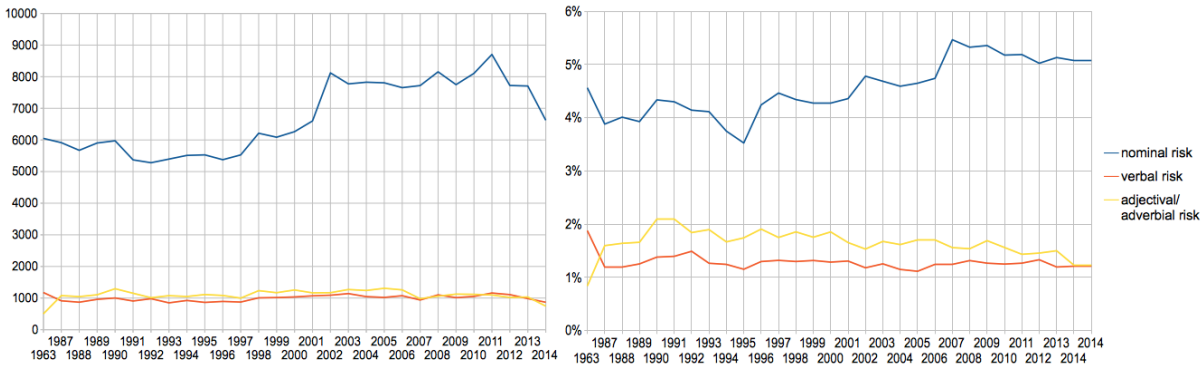

Figure 3.1: Occurrences of risk words by word class (absolute and relative frequencies)

We compared this against the relative frequencies of nominal, verbal and adjectival/adverbial lexical items in the corpus as a whole, in order to account for any trends toward nominalisation in the NYT more generally (Figure 3.1). This showed that even when compared to potential trends toward nominalisation

generally, nominal risks are still on an inbound trajectory. Verbal and adjectival/adverbial risks are both less frequent overall and more static in their trajectory.

These initial findings guided the rest of the investigation: particular attention was paid to nominal risks, as these were the site of the most longitudinal change. That said, these categories provide merely a categorisation of the formal features of risk words. Functionally, things are substantially more complicated: *running a risk*, for example, while featuring a nominal risk, is in reality a risk process; similarly, though risk is nominal in *risk management*, risk is nominal, it functions as a modifier, rather than a participant.

2. Which experiential roles do risk words occupy?

Within the Transitivity system, a risk word may take the form of a participant, process or a modifier. Using Stanford CoreNLP's dependency parsing, we counted the frequency of risk words within these four functional roles (Figure 3.2). Somewhat unexpectedly, the results were very similar to the word-class based results.

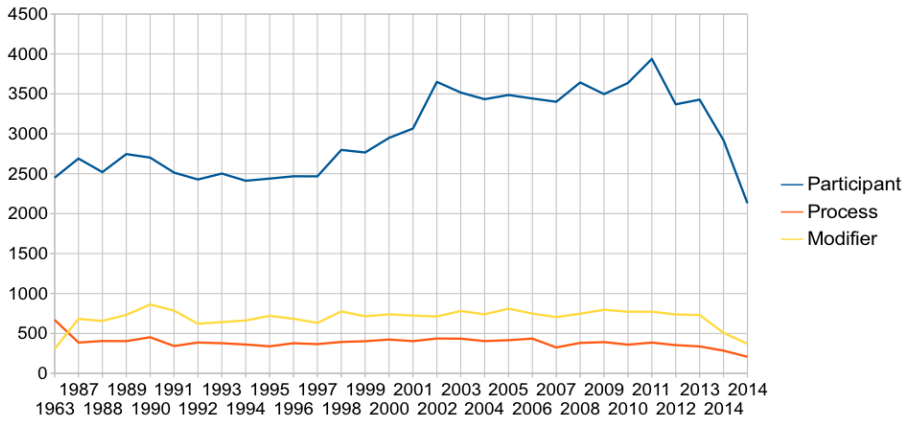

Figure 3.2: Functional roles of risk words

More discussion here, perhaps, as well as the above chart as relative frequencies. I may also have to account for risk within prepositional phrases here.

3. Is risk more commonly in the position of experiential subject or experiential object?

Risk as a participant may take the form of an experiential subject or an experiential object. Our first area of interest was the proportion of each, with respect to general trends in the NYT. As shown in Figure 3.3, risk is more commonly an object than a subject. It is also apparent that risk as experiential subject is on an static trajectory, while risk as experiential object is inbound. The significance of this is discussed in more depth in Section 10.

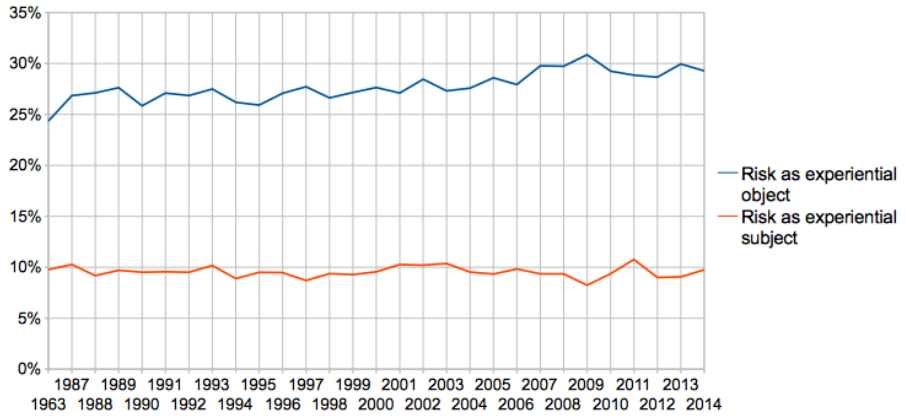

Figure 3.3: Risk as experiential subject and object as percentage of all risk roles

4. What processes are involved when risk is a participant?

We then wanted to determine the most common processes in which risk as a participant is involved. Tables 3.1 and 3.2 show the top twenty processes for risk as experiential subject and object, taking passivisation into account.[12]

Processes when risk is experiential subject	**Total**
be	8954
increase	460
outweigh	278
rise	269
say	222
come	201
remain	192
go	190
have	179
make	148
seem	148
involve	145
grow	133
exist	127
take	121
become	120
lose	120
include	113
appear	111
pay	100

Table 3.1: Processes when risk is experiential subject

Processes when risk is experiential object	**Total**
reduce	5609
pose	4179
increase	4063
have	2879
carry	2115
face	1477
raise	1115
minimize	1009
assess	841
create	731
outweigh	704
avoid	683
present	619
assume	593
consider	588
see	563
understand	493
accept	492
weigh	473
eliminate	450

Table 3.2: Processes when risk is experiential object

5. How are participant risks modified?

Most commonly, risk as a participant is modified through adjectival pre-head modification or post-head modification with a subordinate clause or prepositional phrase. Ignoring the distinction between subject and object risk, and collapsing pre-head and post-head kinds of modification, Tables 3.3 and 3.4 show the most common pre- and post-head modifiers of risk as a participant.

Pre-head modifier	**Total**
high	4753
great	3444
big	1672
political	1520
potential	1340
financial	1164
low	1056
more	1051
significant	1003
serious	935
real	869
little	761
own	713
substantial	547
less	541
such	514
calculated	469
considerable	463
possible	458
other	423

Table 3.3: Pre-head modification of participant risk

Post-head modifier	**Total**
cancer	2344
disease	1777
attack	1597
death	1025
injury	823
infection	811
loss	408
war	391
failure	383
inflation	368
problem	346
default	336
stroke	325
complication	288
damage	251
transmission	248
harm	244
aid	227
recession	217
accident	208

Table 3.4: Pre-head modification of participant risk

Some of these modifiers are undergoing longitudinal trajectory change. As can be seen in Figure 3.4, *calculated risk* has an outbound trajectory, decreasing steadily. The large number of occurrences projected for 1963, however, is largely the result of the 1962 Broadway play by the same name. Of course, the choice of name for the production may also serve as evidence for the salience of the construction in the earlier samples. *Potential risk*, on the other hand, is on an inbound trajectory. Also interesting is the spike in the *high risk* construction between 2002–2004.

Significance of this?

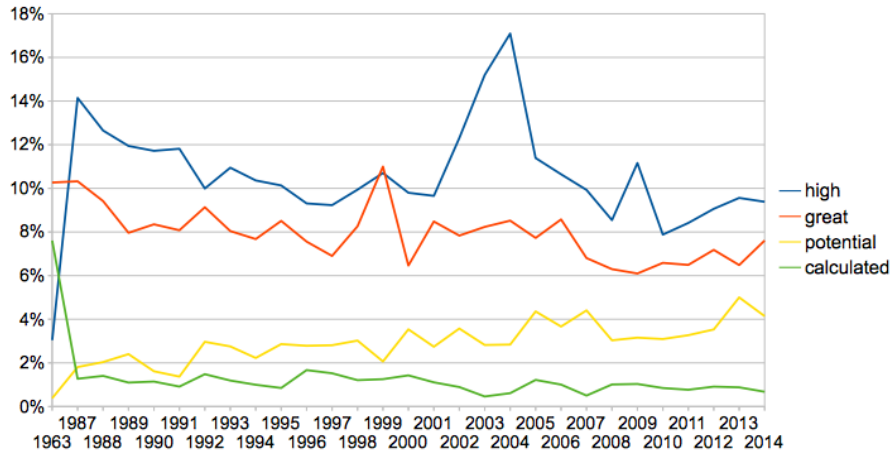

Figure 3.4: Selected modifiers of participant risk as percentage of all risk modifiers

6. What kinds of risk processes are there, and what are their relative frequencies?

Our second area of interest within the transitivity system is risk as a process. Within the corpus, we located four distinct risk processes. First, risk alone may be a process (*I won't risk it*). Second and third are *running risk* and *taking risk*—process–range configurations, where the verbal component is largely shorn of meaning, and with meaning conveyed primarily in the nominal in object position (Halliday & Matthiessen, 2004). The final process, *putting somebody/something at risk* involves an obligatory nominal object argument and a prepositional-phrase complement.

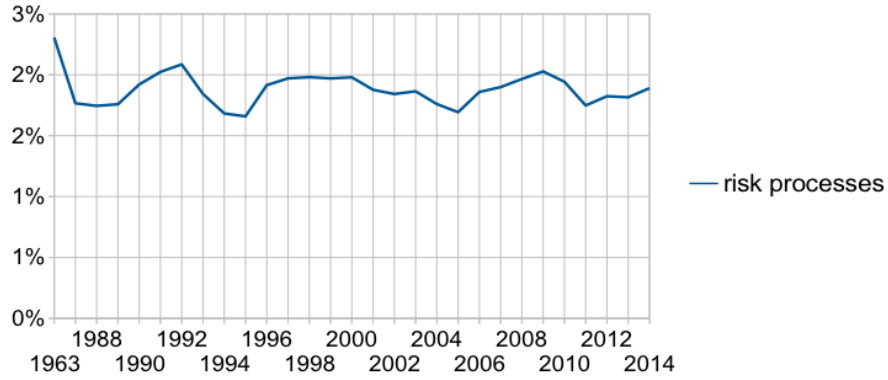

Figure 3.5: Risk processes as percentage of all parsed processes

Our first interest is the overall frequency of these four risk processes, when compared with the number of processes in the entire corpus. From Figure 3.5, we concluded that risk processes generally are on an static/slightly outbound trajectory, with a notable decrease in frequency between the 1963–1987 samples. Figure 3.6 charts the trajectory of the four identified risk processes. Most interesting here is that *putting at risk* has overtaken *running risk* in frequency.

Significance of this?

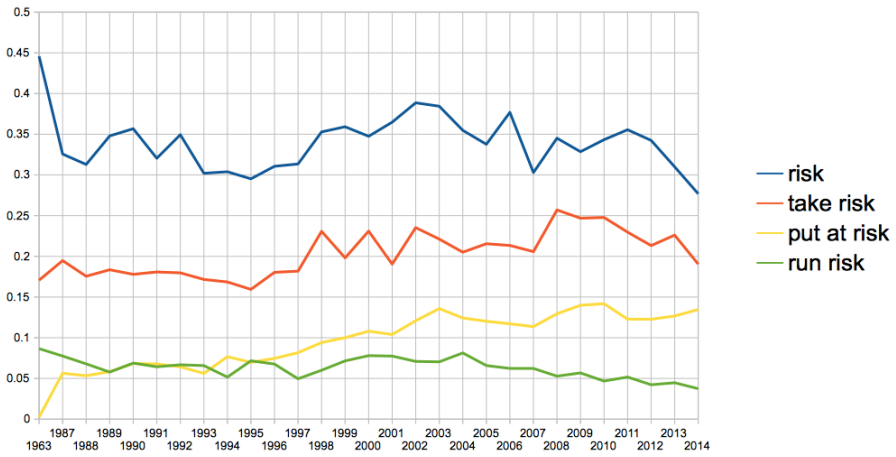

Figure 3.6: Four types of verbal risk as percentage of all verbal risks

7. When risk is a process, what participants are involved?

Clauses containing risk processes are a rich site for analysis, as the semantic roles of participants are determined by their placement with respect to the process. Experiential subjects of risk processes can be mapped to *riskers*. Experiential objects are either *risked things* or *potential harm* (*they risked their lives/death*). Table 3.5 lists the most common subject and object participants of risk processes. Also of interest are clauses embedded within risk processes (e.g. *she risks hurting herself/losing her life*). Table 3.6 lists the (lemmatised) top twenty subordinated processes in the corpus.

Risker	**Risked thing/ potential harm**
person	life
company	injury
state	loss
woman	everything
man	death
investor	money
bush	wound
player	war
government	career
worker	arrest
republican	health
clinton	damage
bank	reputation
democrat	fine
anyone	capital
obama	future
child	confrontation
move	job
firm	backlash
administration	failure

Table 3.5: Riskers and risked things and/or potential harms

Embedded process	**Total**
lose	1260
be	1095
alienate	379
have	347
become	285
get	184
make	166
turn	119
go	113
offend	110
take	86
look	85
undermine	82
anger	79
fall	78
create	76
put	74
miss	73
give	73
damage	62

Table 3.6: Most common embedded processes in risk processes

Riskers are most typically powerful institutions or individuals. Risked things and potential harms are generally serious and grave. A mismatch occurs here: *Bush* and *Obama* do not likely risk *wounds*, *arrest* or *death*. In terms of subordinated processes, notable is the appearance of processes that are fairly uncommon: *alienating*, *offending*, *undermining* and *angering* and are three key examples, ranking amongst expected processes like *being*, *having*, *getting*, *making* and *going*. Without considering longitudinal change, we can see from this that the embedded processes are often related to more powerful social actors: states, political parties and politicians risk alienating electorates; diplomats risk offending one another.

Longitudinal trajectories of a couple of constructions here?

8. When risk is a modifier, what are the most common forms?

Modifier risks are unique for their variety and diversity: through compounding, comprehensible new risk words and phrases can easily be created. The entire corpus contained 327 unique adjectival risk words, including *non-risk*, *de-risk*, *once-risky*, *take-no-risks*, *risk-swapping*, *risk-abhorrent*, *price-for-risk*, *post-risky*, *pooled-risk*, *personal-risk*, *optimum-risk*, *one-risk-factor*, *one-pitch-can-end-his-career-risk* and *low-risk-to-society*. That said, most of these occur no more than a handful of times. By far the most common were *risky/riskier/riskiest* (15588 occurrences), *high-risk* (5533), *low-risk* (1086), *at-risk* (902), *risk-free* (883) and *risk-taking* (789). Of these, four exhibited trajectory shifts (see Figure 3.7). The basic adjectival forms (*risky*, *riskier*, *riskiest*) are dominant in the 1963 sample, then decrease, and re-emerge in 2000. *High-risk* though very rare (two instances) in 1963, has become more common, and stabilised in trajectory. *Low-risk* and *at-risk* are on a consistent inbound trajectory.

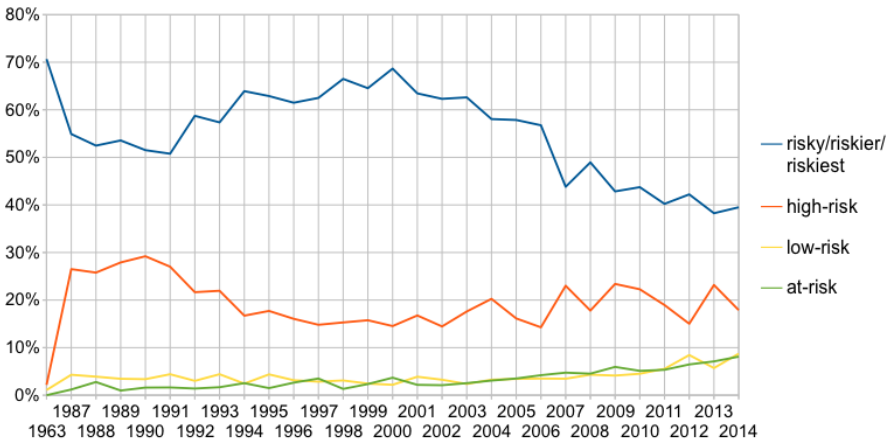

Figure 3.7: Common adjectival risk words as percentage of all adjectival risks

The prevalence of high-risk in the 1980s is largely due to the AIDS epidemic: concordancing reveals that certain populations (gays, African Americans, Haitians) are at high-risk of being inflected by HIV. *At-risk* is rare in earlier editions, but increases in prevalence steadily.

This shift in risk is modifier is an important one. Low, moderate and high risk comprises a gradient, or scale, while at-risk is a binary. As with the shift toward *potential risk*, this indicates both an increasing pervasiveness and a decreasing calcuability of risk.

9. When risk is a modifier, what is being modified?

Risk as a modifier can be placed either before or after the noun it modifies (*an at-risk person/a person at risk*). These two constructions are collapsed in Tables 3.7 and 3.8, which respectively list the participants most frequently modified by any risk modifier, and the participants most frequently modified by *at-risk/at risk*. Note that while risk-modified participants generally are financial and economic in nature (*investment, business, loan, asset*), the at-risk subset is comprised of vulnerable populations of people (*women*, *children*, *students*).

Risk-modified participant	**Total**
investment	696
business	515
behavior	508
group	466
loan	421
asset	388
strategy	377
bond	346
area	307
venture	301
security	287
patient	265
pool	239
bet	214
move	204
activity	201
proposition	199
child	170
woman	161
student	158

Table 3.7: Most common risk-modified participants in the corpus

At-risk participant	**Total**
person	439
child	368
woman	209
student	179
nation	135
patient	110
youngster	93
group	91
population	64
family	58
kid	50
youth	48
money	48
worker	45
life	41
job	41
man	40
area	35
teenager	32
other	32

Table 3.8: Most common at-risk participants in the corpus

10. How arguable is risk?

As noted earlier, our central concern with the Mood system is the degree of arguability associated with the concept of risk. Risk in Subject, Finite and Predicator positions is the most arguable. Risk words within Complements and Adjuncts are less arguable.

Based on the kinds of parsing provided by Stanford CoreNLP, it was possible to measure arguability in two ways. First, we can map dependency relationships to the systemic-functional notion of arguability. A dependency grammar locates the predicator of a clause and assigns it a position of zero. A '1' is then assigned to its most immediate dependent (other components in the verbal group, if present, or the head of the Subject, if not). This process continues until no lexical items are unattached, or 'ungoverned'. In effect, the higher the number attached to a word, the further it is semantically from being an important component in the meaning, and thus, in systemic functional terms, the less arguable the word.

Highlighting three sampling periods as in Figure 3.8 shows that risk is less and less often the predicator, or an argument of the predicator. That is, risk is less and less often in more arguable positions. In 1963, risk appears much more commonly as the main verb, or as one of its more immediate dependants.

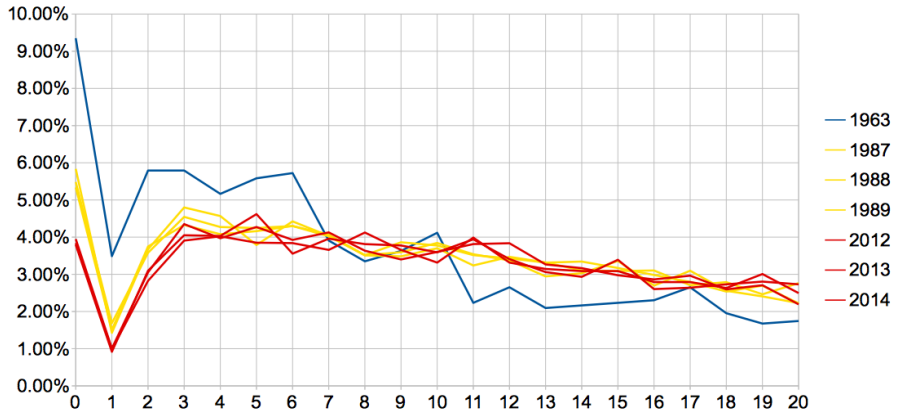

Figure 3.8: Risk words by dependency position in clause

By 2014, risk words are more commonly occupying roles within grammatical objects and adjuncts, and thus have greater numbers of governors.

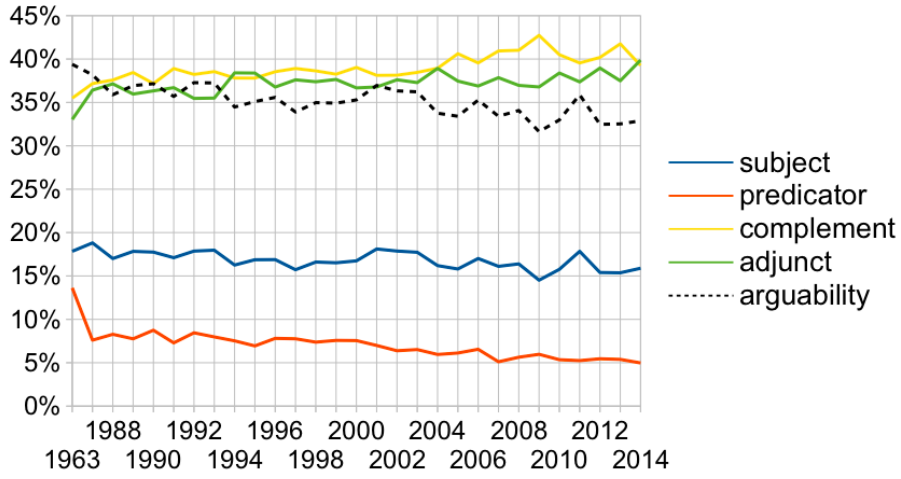

Figure 3.9: Trajectories of risk within each Mood component

The second thing we can use dependency output for is identifying the functional roles of risk words. This is more accurate than using the dependency ranking, but creates a very long list of functional roles. Of key interest, however, are risk words at the head of each major component of the Mood system—Subject, Predicator, Complement and Adjunct (risk cannot grammatically occur as a Finite, so it is excluded here). From Figure 3.9, we can see that risk is shifting from Subject and Predicator to Complement and Adjunct roles. By providing each role with a relative weight, we can plot arguability as a single decreasing trend line, showing the increased implicitness of risk within the language of the NYT.

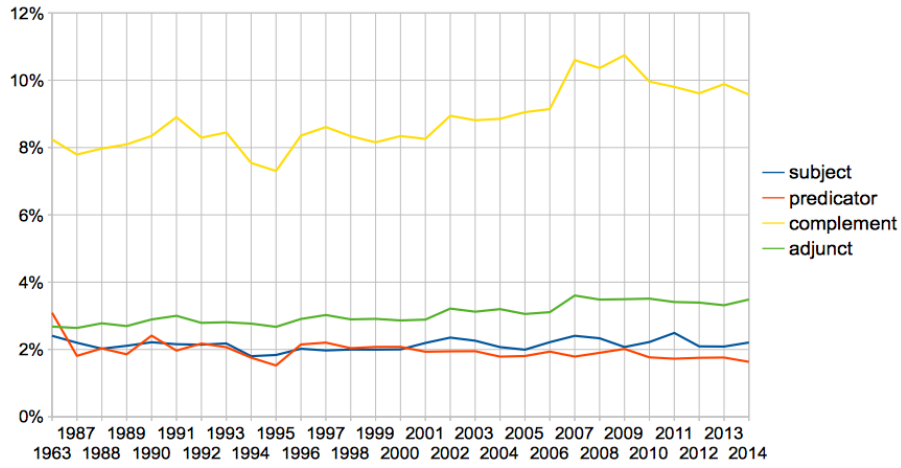

Figure 3.10: Frequency of risk words for each Mood component as percentage of all parsed data

11. Risk words and proper nouns

We searched for proper noun groups in parse trees containing a risk word. This is a departure from many of our earlier queries, as here we are looking only at which entities co-occur with risk language, rather than determining how risk words and non-risk words relate to other another lexicogrammatically.

The result of this query was n different proper noun groups. We took the 200 most common results, and merged any that denoted the same entity: *F.D.A./Food and Drug Administration*, or *Federal Reserve and Fed*.

We then grouped results into thematic categories:

1. People
2. US locations
3. Nations
4. Geopolitical entities
5. Companies
6. Organisations
7. Things

The results were then plotted.

A number of historical events were easily recognisable within the peaks and troughs of these charts.

Presidents and their rivals come and go

Though Bushes and Clintons are conflated, we can still reasonably infer which was being spoken about at which. Doubt can be eliminated by concordancing.

Chapter 4

A comparison of economics, health, and political risks

Key participants were first tallied (`/NN.?/ »# (NP !> PP)`) in order to broadly understand the key social actors in the three subcorpora (see Figure 4.1).

Figure 4.1: Key participants in the *Economics*, *Health* and *Politics* subcorpora

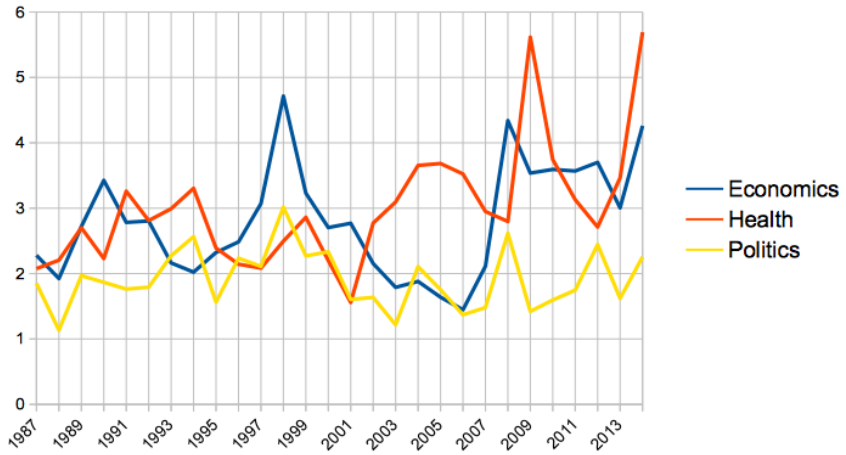

Figure 4.2: Risk words per total number of article topics per year

Due to time constraints, we restricted the topic comparison to domains that had yielded interesting

insights in the earlier interrogations. Further, we found that the smaller size of the subcorpora limited us to lexicogrammatical queries that outputted a large enough number of results for quantitative reliablity. Thus, we focussed on the following three areas:

Economics	**Health**	**Politics**
political	high	political
big	great	great
economic	low	big
financial	other	high
great	serious	own
high	financial	serious
more	potential	new
real	medical	real
systemic	more	considerable
significant	significant	more
new	cardiovascular	other
little	political	significant
global	possible	economic
serious	small	financial
other	real	potential
excessive	such	personal
potential	genetic	little
such	ovarian	such
much	same	public
own	bad	military

Table 4.1: Most common adjectives modifying nominal risks in the topic subcorpora

1. Summary

Unlike the general corpus of all articles, these subcorpora made it possible to observe the influence of key events:

1. Event 1 and 2 in Economics
2. Event 1 and 2 in Health
3. Event 1 and 2 in subcorpus 3

Chapter 5

Discourse-semantics of *risk* in the NYT

Accordingly to SFL, the sum total of lexicogrammar, abstracted, realises the discourse-semantics of texts. Accounting for discourse-semantic meaning involves sensitivity to realised lexicogrammatical forms, but also to the ways in which incongruence and grammatical metaphor can create similar meanings through differing grammatical constructions: as noted earlier, *potential harms* may be realised as a participant in a process of risk (*Bush risked losing the election*), or as a modifier of a risk participant (*the cancer risk/the risk of cancer*).[13] Given the diversity of roles in which risk words can appear, the delineation of risk by roles within mood and transitivity systems in the previous section was thus a methodological necessity, but one with heavy ramifications for analysis. At the level of discourse-semantics, it becomes necessary to discuss risk word behaviour more fluidly, with reference to both experiential and interpersonal meanings, and with distinctions between risk as participant, process and modifier largely collapsed. This is perhaps especially so in in our case, as risk is an example of a lexical item that may be congruently realised as either participant and process, straddling the semantic space between entity and event.

The first part of this discussion provides a description of risk in the NYT absent longitudinal considerations—something akin to the descriptions provided by Hamilton et al. (2007) and Fillmore and Atkins (1992), but from a systemic-functional, rather than frame-semantic purview. The second part is concerned with accounting for shifting discourse-semantics of risk, via the lexicogrammatical findings presented in the previous section. In the final section, longitudinal shifts are discussed in the context of specific events, broader social change, and sociological theory.

1. A monochronic description of risk

Before turning our attention to the behaviour of risk words over time, it is useful to provide a short description of the way risk words are generally used in the NYT.

Foremost, striking is the ability of risk to function within all open word classes (noun, adjective, verb, adverb), as well as the sheer diversity of risk words. 507 unique lexical items containing risk were found[14] , including many (albeit vary rare) words lacking existing lexicographical description: examples such as *risk-shy*, *risk-addicted*, *risk-elimination*, *species-at-risk* and *risk-happy* demonstrate the overall salience of risk and the nuance with which it is instantiated in news discourse. Further testament to this salience are the nuanced distinctions in riskers' awareness of potential harm in *risking*, *putting at risk, taking* and *running* risks.

In many respects, our findings agree with those of other monochronic descriptions of risk language.

First, we can see the usefulness of the frame-semantic categorisation of the kinds of participants/social actors that occur within the risk frame (i.e. Fillmore & Atkins, 1992): we often found it useful to divide corpus interrogation results into categories of *risker*, *potential harm*, *risked thing*, and the like. Promising is the fact that in many cases, we can use the grammatical structure of the clause to automatically return lists of each kind of participant. In cases where the grammar alone cannot tell us the participant role (*I risked my death*, *I risked my life*), manual sorting is not difficult, as there is little ambiguity. If we insert the *losing* participle (*I risk losing my life*, but **I risk losing my death*), we can quickly determine if a result is a *potential harm* or a *risked thing*. This is especially so when risk is the *process*, rather than a participant or modifier. With this in mind, focussing more exclusively on risk as process in very large parsed datasets may prove elucidating.

Our findings also agree with a key claim made by Hamilton et al. (2007): health and illness risks were surprisingly prominent within our data. As will be discussed below, however, this does not appear to be a purely static phenomenon: our longitudinal analysis points toward health risks as being far more common in contemporary language than in the language of our 1963 dataset.

A second point on which we agree is with their contention that risk words behave differently in different registers and genres, and that comparison of genres is worthy of further study (though here we rely on not on our dataset but on a long history of research in support of this contention within SFL):

> We find in these discourse environments that the focus of the semantic prosody and the semantic preference changes according to the context in which they occur. While this may be something that some (but not all) sociologists of risk may have intuitively sensed in the past, there are empirical data from corpus linguistics to suggest now that the semantic prosodies can and do change slightly from one context to another Hamilton et al. (2007, p. 177).

Their dataset included transcribed spoken conversations. This register is remarkably different to that of NYT articles, and examples of risk in these contexts demonstrate this quite clearly (e.g. *Don't don't risk it eh*; *Cos there isn't a risk of going of there*). The key characteristics of these examples (informal lexis, unrecoverable deictic references, low lexical density, etc.) contrast starkly with our examples.

Due to the composition of our dataset, we can have little to add to descriptions of risk in casual spoken language, aside from recognising that spoken risk talk is likely to point toward very different, and interesting, results. Though we believe our results may be generalisable to the behaviour of risk in relatively formal written contexts, extended investigation of risk in spoken corpora remains needed.

A key finding that received little attention in this earlier linguistic research of risk language is the notion of participants' *agency in risk*. Readily apparent when risk is process is that the kinds of people who risk are typically institutions or humans in positions of power and influence. Actors of risk processes are often states, politicians, or political parties. The *potential harm* being risked is often an abstract concern: *alienating* or *offending electorates* or *allies*. In these cases, risk is a process engaged in purposively by Actors who stand to gain something equally abstract. In contrast, when risk functions as a modifier of a participant, the participant is far less powerful: women and children are at-risk of sickness; workers are at risk of injury or death. Here, risk is a quality ascribed to the self. Risky behaviour is not often mentioned. For these people, the potential harm is often recoverable from context, but not outlined within the clause. This distribution was largely consistent throughout our dataset, and will be unpacked through sociological analysis in Chapter N.

2. Shifting discourse-semantics of risk in the NYT

Some lexicogrammatical and discourse-semantic phenomena have demonstrated consistent shifts over our sampling period. We turn our attention to them now.

First, though we noted above that risk as a process involves a different set of participants to risk as a modifier, there are still longitudinal changes within this area. When looking at the *risk of loss*, for example, we can see a general trend toward individual losers, rather than institutional losers. In 1963, the things at risk of loss were macro-level and abstract: athletic funding, market share, vital technology, sympathy in the west, and the like. Later, risked things are more individual assets—life and injury being the two most common. We link this conceptually to neoliberalism:

Sorry Jens, neoliberalism is not easy for me to write up.

2.1. Domains of risk disourse

In terms of the topics in which risk words are deployed, we saw that health risks are very prominent in the more contemporary data samples. Our comparison of *Risk of terror* attack* and *risk of heart attack* demonstrates this preference clearly. This change is indeed a longitudinal one: in 1963 editions, a number of constructions evidence that risk was commonly instantiated with regard to diplomacy, war, international relations, and the like. In their most prominent years, AIDS, Vioxx and Merck comprise over 1.6 per cent of all proper nouns that co-occur with a risk word. This is higher than Clinton, Bush or Obama at their peaks, as well as Soviet Union in 1963/1987 or Europe during the Eurozone crisis in 2011. Moreover, in the years following the AIDS crisis, health risk have increasingly related not to infectious diseases (which require institutional responses), but to kinds of illnesses where the responsibility for prevention falls upon everyday citizens through lifestyle choices, rather than politicians, hospitals, or the FDA. Even in the case of Vioxx, where the risk was created by the premature FDA approval, risk language surrounding Vioxx remained geared toward the risks faced by everyday people. Though Merck and the FDA may be blamed, risk remains a more appropriate frame for discussing the potential for heart attack than it does for discussing the potential harm caused by improper clinical trials or financial interests causing the FDA to approve the medication prematurely.

I'd like to unpack this above, and maybe see if there was a shift toward AIDS as something that needed to be prevented by individual action, rather than govt response ...

As Widdowson (2000) suggests, corpus linguistics often reveals things that are contrary to intuition, and this is certainly the case here. Our expectation of new risk meanings related to terrorism after 9/11 was for the most part not met. Rather than a limitation, this can be treated as an insight in itself: the events and topics that come to mind when we think of risk may not necessarily correspond to the reality of risk language generally. Such is the benefit of corpus linguistic investigation of risk, when compared with previous methodologies employed within the humanities and social sciences to better understand risk.

2.2. Implicitness and arguability

The most salient theme from the longitudinal mapping of risk is that of implicitness: increasingly common are grammatical constructions where potential harms and risked things are recoverable only from context.

Below are three examples in 2012:

1. `In 1999, we sold the company, and the next year, we moved to the United States with our two children - a third was born in 2003 - so I could pursue my idea of helping low-income, at-risk youth`
2. `Carolyn F. Blakely, then a new teacher at the school (who retired last year as the dean of the Honors College that now bears her name at the University of Arkansas at Pine Bluff), remembers Neal as an at-risk kid prone to challenge authority.`
3. `Mr. Lane is a sophomore at Lake Academy, an alternative high school for at-risk youths, some of whom take a bus from Chardon High School.`

In these cases, what the participant is at-risk *of* is not a specific negative outcome, but an interrelated set of negative outcomes that are more likely to happen to less powerful people in society. Evoked within this cluster is *poverty*, *drug use*, *disease*, *homelessness*, *abuse*, *fatherlessness*, *dropout*, *gang activity*, and the like. In many cases, *at-risk* takes on a euphemistic quality, most obviously as a substitute for *lower-class*, *non-white* or *poor*. Also interesting here is the muddying of the semantic frame: it is both difficult to determine the exact potential harm, and to classify the participant as a *risker*, which seems to imply some agency or comprehension of the risk. More accurately, these participants are *put at risk*—a risk process that itself is on an upward trajectory within our dataset.

This aligns with the decreasing arguability of risk. Risk in predicator or subject position is increasingly rare, as risk becomes less the nub of propositional meanings. Thus, less and less often is risk a fundamental component in meaning as exchange: in its role within complements and adjuncts, it now more typically plays a supporting role in the provision of information. A ramification of this is that risk becomes an inherent quality of participants in the field of discourse, rather than a process in which participants knowingly or by their own choice choose to engage. This shift is exemplified by the outbound trajectory of *calculated risk*, and its displacement by an uncalculated *potential risk*. In the former, the existence of the risk itself has been acknowledged, and the potential harm/reward have been weighed. In the latter, though the situation can be identified as having potentially negative outcomes, these are formless and immeasurable. *Potential risk* is in fact *a risk of risk*. This aligns with the idea that risk (sociological reference) has come to be simply *threat*.

2.3. Low-risk, moderate-risk, high-risk

During the first years of AIDS, people could be classed according to low, moderate and high-risk groups. Here we have basic quantification of levels of risk. This stands in contrast to the *at-risk* construction discussed above. Of these modifiers, only *low-risk* emerges as an increasingly frequent form. This is also interesting, as it points to a broadening of the semantic scope of risk to include situations where risk remains present: *low-resolution image* does not point toward the increased prominence of low resolution images, but more to the prominence of resolution as thing that meanings are made about. In the same way, the inward trajectory of *low-risk things* does not point toward a culture of less risk, but toward a culture where even things that do not have risk are characterised by their nature to it. We could not locate existing literature supporting a claim that the salience of a concept may be evidenced not only through *extreme case formulations the riskiest, high-risk, very risky*, but through minimisation. Nonetheless, our analysis points to the idea that the increased salience of risk as a concept is in part demonstrated through its instantiation in situations where its significance is claimed to be negligible or banal.

2.4. Risk as modifier

Risk occurs within many different modifier positions:

Of these, pre-head nominal types are rising, and adjectival pre-head types are falling. From these shifts, we can surmise some sociological insight related to arguability (as conceptualised by SFL). In the increasing frequency of pre-head nominal modifiers (*risk management, risk arbitage, risk factor, risk insurance*), we can see increased social significance of risk as a concept through the evolution of specific jobs whose central concern is risk. Pre-head nominal modification is an indication of codification of a concept: such constructions must be culturally recognised constellations of meaning. In comparison, adjectives attach to head nouns relatively freely in English. Cultural recognition of the adjective-noun combination is not a prerequisite for meaning to be understood.

2.5. Arguability

Longitudinal change in the arguability of risk words is consistent. In earlier editions, risk words more commonly occupy more arguable roles, according to systemic functional grammar. In later editions, risk more commonly occurs in heavily dependent positions. Less often does a risk word form the central component being discussed; more often, it exists as a modifier of one of these components, or as a part of a supporting, subordinate clause.

We are limited in our ability to interpret this result. Little has been written about the relationship between dependency grammars and SFL. As dependencies are inherently functional-semantic, rather than generative-grammatical, dependency is perhaps the most useful [15] mainstream grammar for learning about the semantic behaviour of a given word. That said, though functional categories provided by Stanford CoreNLP's dependency parser overlap in many respects with categories in the Mood system of SFL, there are still mismatches, or shortcomings. Most critically, dependency grammar conflates interpersonal, experiential and textual systems, while SFL demands three separate parses. As discussed earlier, the systemic-functional conceptualisation of subjecthood is threefold, whereas CoreNLP simply nominates the interpersonal subject.

Due to the availability of nuanced querying languages for phrase structure grammar annotation, our investigation leaned toward grammatical structure annotation over dependency grammar. This is despite a problematic relationship between functional and phrase structure grammars. Given that interesting preliminary findings were unearthed by querying dependency information, we conclude that further exploitation of dependency annotation for the purpose of risk language analysis appears to be a promising area for further analysis.

3. Sociological perspectives

Below this point is not particularly readable, sorry.

The task that remains is to connect observed shifts to their temporal context. In terms of the annual subcorpora, this was by no means a clear-cut task.

Take it from here, Jens.

When focussing on the subcorpora of economic, health and political risks, linguistic reactions to real-world events were much easier to locate. We concluded that further investigation of risk would do well to focus on risk as instantiated within texts sharing a semantic field.

Our investigation of topic subcorpora was limited by scope. That said, the open-source tools we have developed for interrogating corpora for discourse analysis could easily be put to use in an investigation of a topic subcorpus.

Mapping events to risk instantiation in the topic subcorpora here

We found little evidence that health crises resulted in increased frequency of risk in articles centred on economics or politics. This seems to suggest that while real-world events influence the instantiation of the risk semantic, this instantiation remains more or less limited to the field(s) of discourse to which the real-world event is most related.

A final point of interest is that adjectival risk words behaved largely contrary to expectations. Adjectival risks as modifiers of participants appear to be decreasing in frequency. Furthermore, though there is a very large variety of adjectival risks, this variety does not seem to be expanding.

Perhaps in this finding there is some evidence for the Risk Society thesis, in that the ways in which risk can characterise a situation were more or fully articulated during high modernity. Though these characterisations continued to be applied today, saturation point may have been reached.

What can be concluded from the finding that real world events do not appear to have long-lasting effects on the behaviour of risk words?

Ultimately, perhaps we should not be surprised by this finding. Language is a system that must be resilient against such influences: if single events caused meaningful changes in the lexicogrammatical behaviour of a single word, communication between those aware of and unaware of events would be made more difficult. Accordingly, our suggestion would be that temporary change in the behaviour of a word (as can be seen in spikes in the number of risk words surrounding certain events) are interesting in and of themselves. Moreover, these changes can potentially be measured in pseudo-real-time by mining RSS feeds, using the Twitter API, and so on. Lexicographers could take note of which kinds of events bring about instantiation of a certain word or concept, and create definitions accordingly. Discourse analysts and sociologists could hypothesise the co-occurrence of certain kinds of language with certain kinds of events, and use real-time data to confirm or refute these hypotheses. Cooperative efforts between functional linguistics and sociology, however, are dependent upon a reconciliation of divergent conceptualisations of the relationship between text and context. This issue is elaborated below.

Predictive applications of Big Data/corpus linguistic methods have already been discussed: Michel et al. (2011) and Leetaru (2011) argue that nuanced mining of large quantities of language can potentially predict civil uprisings such as those seen in the Arab Spring, for example. It must be remembered that these studies have been criticised for their far-reaching conclusions (e.g. Zimmer, n.d.), and of course that predictive applications of corpus linguistics to date have had the benefit of hindsight. Interpreting peaks and troughs in particular kinds of language is also far from straightforward: increasing numbers of risk words before 9/11 could be interpreted as either a possible predictor of the event or as evidence that the event did not itself cause an increase in the use of risk words. As such, we remain cautiously optimistic about predictive applications. More practically, it seems that such applications are feasible only when there is little delay between text production and text analysis: automated analysis of language circulated via the Web seems a much more sensible starting point for predictive work than static corpora of digitised newspapers and books.

4. Reconciling sociological and systemic-functional conceptions of text and context

Functional linguistic theories such as SFL not only provide a grammar, but also a conceptualisation of the relationship between text and context. In the case of SFL more specifically, the argument is that context is contained within text. Compelling evidence of this is that understanding of the more abstract genre/context from which a text is taken can be gained through exposure to lexicogrammar only.

At issue for sociologists is that this argument rests on a particular operationalisation of the idea of context. The definition according to SFL (though indeed inspired by scholars with significant in sociology, e.g. Malinowski) remains deeply concerned with language. In reality, it is a projection of grammatical phenomena (mood, transitivity, theme) onto the situations and cultures in which texts are produced. Though this has proven a useful heuristic within (critical) discourse analysis, it is in many ways alien to sociological theory, where texts tend to be considered with respect to political, historical and social events and movements, rather than with respect to communicative systems.

The systemic-functional description of the context of culture contains no references to the effects of current events on language production. While SFL has demonstrated its usefulness without such considerations, this usefulness has been for linguistics. Corpus linguistic applications of SFL have seldom traced the influence of real-world events.

It is not hard to imagine scenarios where important meanings can be made through absence of references to certain things. In seldom considering both cognitive elements of language production and the influence of specific events, SFL has remained largely unable to conceptualise the notion of meaning made through omission. ...

Naturally, a current event can influence the likelihood of certain parts of lexicogrammar (the most banal example is in proper nouns, where the appearance of a politician is certain to influence the likelihood of his or her mention in texts). SFL is predictive in the sense that it can predict that different genres will be more or less likely to involve certain speaker choices. So far, it has not been able to ...

Like SFL, sociology aims in part to provide a link between text and context. Context, however, is generally not treated as simply a Malinowskian constellation of field, tenor and mode, but also a backdrop of current and past events that inform and shape discourse at the time of its production.

With these issues in mind, we propose that SFL and general sociological theory are useful partners. SFL provides a means of relating lexicogrammar of texts to discourse-semantic meanings. It can then

Key sociological ideas such as reflexive modernity or neoliberalism can be expected to exert influence over texts produced during these movements. Though earlier SFL treats ideology as the most abstract stratum affecting the production of texts, this conceptualisation has been abandoned by a number of current SF linguists.

Earlier SFL indeed devoted significant energy to exploring the ways in which ideologies such as capitalism are manifested in the content strata of language.

What often goes unsaid in SF theory is that an additional usefulness of SFL is in its ability to draw a line between the kinds of context (field, tenor and mode) that are embedded within the lexicogrammar of a text and the kinds of context that leave no immediately identifiable trace.

At this point, sociological theory can fill in the missing parts of the picture.

At a level of greater abstraction, functional linguistics and sociological theory can be combined to flesh out the text/context relationship. Functional linguistics is concerned with language as a tool to make things happen in the world; sociology can add to the understanding of how culture informs our

motivations for making these things happen, for presenting ideas in certain ways, etc.

Chapter 6

Limitations of the study

Our methodology was innovative, and involved fitting theories, practices and tools together in novel ways. Through the course of our investigation we noted two major clusters of limitations. The first were issues relating to the performance and epistemological consequences of digital tools used during the investigation. In short, available digital tools may not perform as desired. In the case of parsers, this is generally an incorrect parse.

The second major issue unearthed during the investigation concerned the size of the dataset, which, aside from being simply computationally intensive, was also so large that it constrained the kinds of analytical methods available to us. With 29 annual subcorpora, as well as three topic subcorpora, we struggled to simultaneously maintain a focus on minute changes in lexicogrammar and to connect change generally to events of interest to sociologists. Indeed, though instantiations of risk words may react to current events, further subdivision of the corpus into weekly/monthly subcorpora proved too unwieldy. A similar investigation could be carried out on one subcorpus alone, divided into weeks or months, in order to better assess the influence of individual events. The richness of the data also prevented direct comparison of more risk fields, with only a cursory treatment of government and health risks given here. A final issue caused by issues of data size was that we were unable to manually check each

Search query output was manually read to determine that the correct features were being located. What was missing as a result of parsing problems or query design likely went unnoticed amongst the streams of text. By the conclusion of the interrogation, millions of clauses had been manipulated, millions of features extracted and counted—mistakes are unfortunately bound to remain.

1. The limits of lexicogrammatical querying

Discuss how our investigation focusses more or less on congruent realisations, and is monomodal...

2. Conclusions

Instantiation of risk words is linked to real-world events: the beginnings of the AIDS epidemic are accompanied by a spike in health risk discourse; 9/11 appears to be a catalyst for increasing discussion of risks and threats of terror and war. It remains very difficult, however, to fully disentangle the constructive-

responsive relationship between real-world events and instantiation of particular concepts in language. Broader ideologies and social movements may indeed be more reliable predictors of linguistic change.

- We found that it is very difficult to pinpoint the effect of individual events on risk word usage in the NYT as a whole.
- Our interrogation of economics, health and politics articles turned up clearer evidence of the effect of real events.
- Even so, our approach to the creation of the subcorpora was simple, and the topics were very broad. Manually selection of automatically located articles for more specific topics would likely result in clearer indications of the effects of events on risk word usage.
- The main thrust of our approach was to investigate the sum total of NYT articles during the sampling years.
- While political risks peaked during US election times, this could not be observed when analysing the corpus as a whole.
- Accordingly, our findings pointed more toward the influence of broader social movements than to specific events.
- We interpreted many of the changes in the behaviour of risk words as evidence for **neoliberalism** and **reflexive modernity**.
- Indeed, the corpus developed for this study could be reused in a multitude of ways to
- Finally, it must be borne in mind that the NYT is merely one newspaper, and newspapers are merely one genre
- We selected NYT for its size, consistency, the availability of digitised content, and its influence in global discourse.
- Most obviously, the emergence of the Web challenges this set of criteria in a number of ways. Popular social networks, as well as the public web, produce exponentially more content than a single newspaper.
- Mining global news or blog posts via RSS feeds, or Tweets via the Twitter API, allows the quantitative analysis of voices typically marginalised or absent within mainstream media.

Notes

1. This was operationalised through the case-insensitive regular expression **brisk.*?****b**, where **b** acts as a word-bounrdary marker.

2. As a part of an ongoing Australian Digital Humanities initiative, since the beginning of our analysis, we have been allocated resources for creating and cloud-hosting a much larger corpus. All 1.8 million articles from the *New York Times Annotated Corpus* were turned into an identically structured, though dramatically larger, cloud-hosted corpus. Interrogation of this corpus was used to determine whether trends in the behaviour of risk words were localised to the word itself, or to general stylistic/language change in the *New York Times*. More details on this project are to be presented in Zinn and McDonald, forthcoming.

3. We tried a number of strategies for collecting topic subcorpora, such as exploiting topic modeller and keyword metadata. Ultimately, however, we relied on the hand-classification. A limitation of the selected approach is that—an article collected in the health subcorpus was tagged with 'Livestock health', for example. Similarly, article categories may be lacking: Figure 1.1 is tagged only with MENINGITIS, and thus is not included in our health subcorpus. More obviously, 1963 articles had not been classified this way, and thus do not feature in the three topic subcorpora.

4. A key cause of incorrect parsing is non-standard language (perhaps regional, colloquial, etc.). Examples of this kind of language in news publications are interesting in their own right, but due to misannotation, are likely to go unfound during corpus interrogation, and thus unanalysed.In our case, this problem was exacerbated by the fact that time contraints precluded a manual scoring of parser accuracy.

5. The mode dimension, responsible for reflexively organising language into comprehensible sequences, remains largely static between print news micro-genres, though mode features are likely to be at risk when news is transmitted via different media

6. Though role relationships between journalists and their readership have undergone significant shifts (especially since the popularisation of online news), charting these changes falls largely outside the scope of this project.

7. The corpus contained too few modalised risk predicators for analysis of longitudinal change in modalisation.

8. Though we are focussed on corpus-assisted investigation at present, indeed the dataset under investigation is of size and scope as to be of interest to corpus-driven researchers, language and media specialists, etc., and indeed, such projects are forthcoming.

9. Lemmatisation is the process of counting the base forms of tokens, rather than the token itself. *Taken* would be classified under *take*, for example. While lemmatisation is not *always* the best option, as it can collapse different parts of speech, tense information, etc., it is certainly appropriate when determining the most common predicators, etc.

10. We need only to look at the number of lines of code needed to develop an accurate tokeniser and an accurate grammatical structure parser to understand the reasons why lexis appears as the de-facto centre of CL/CADS today

11. Modification through embedded clauses (*the children who were at risk*) has been left out for reasons of scope.

12. *Take* and *run* are removed from the object column here, as *take risk* and *run risk* are considered risk processes.

13. A key issue in CADS is the ability to systematically account for rank-shifted meanings (See McDonald, Forthcoming).

14. This naturally depends on your definition of a word/token. If we removed hyphenates or tokens containing a slash (*risk/ reward*), the list would be dramatically reduced in size. Lemmatisation would compress this list even more.

15. Current systems for automatic systemic functional annotation tend to rely on dependencies generated with Stanford CoreNLP

References

Baker, P. (2004). Querying Keywords Questions of Difference, Frequency, and Sense in Keywords Analysis. *Journal of English Linguistics*, *32*(4), 346–359.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, *19*(3), 273–306.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Chen, K.-J., Huang, C.-R., Chang, L.-P., & Hsu, H.-L. (1996). Sinica corpus: Design methodology for balanced corpora. *Language*, *167*, 176.

Christie, F., & Martin, J. R. (2005). *Genre and Institutions: Social Processes in the Workplace and School*. Continuum.

Duguid, A. (2010). Newspaper discourse informalisation: a diachronic comparison from keywords. *Corpora*, *5*(2), 109–138.

Eggins, S. (2004). *Introduction to systemic functional linguistics*. Continuum International Publishing Group.

Eggins, S., & Slade, D. (2004). *Analysing: Casual Conversation*. Equinox Publishing Ltd.

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*.

Freake, R., & Mary, Q. (2012). A cross-linguistic corpus-assisted discourse study of language ideologies in Canadian newspapers. In *Proceedings of the 2011 Corpus Linguistics Conference, Birmingham University. Available at http://www. birmingham. ac. uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-17. pdf.*

Halliday, M., & Matthiessen, C. (2004). *An Introduction to Functional Grammar*. Routledge.

Halliday, M. A. K. (1991). Corpus studies and probabilistic grammar. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics: studies in honour of Jan Svartvik* (pp. 30–43). New York: Longman.

Hamilton, C., Adolphs, S., & Nerlich, B. (2007). The meanings of 'risk': a view from corpus linguistics. *Discourse & Society*, *18*(2), 163–181. doi: 10.1177/0957926507073374

Hasan, R. (1987). The grammarian's dream: Lexis as most delicate grammar. In M. A. K. Halliday & R. P. Fawcett (Eds.), *New Developments in Systemic Linguistics* (pp. 184–211). New York: Pinter Publishers.

Hunston, S. (2013). Systemic functional linguistics, corpus linguistics, and the ideology of science. *Text & Talk*, *33*, 617. doi: 10.1515/text-2013-0028

Johnson, S., & Suhr, S. (2003). From 'Political Correctness' to 'Politische Korrektheit': Discourses of 'PC' in the German Newspaper, Die Welt. *Discourse and Society*, *14*(1), 49–68.

Leetaru, K. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, *16*(9).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60).

Martin, J. R. (1984). Language, register and genre. In F. Christie (Ed.), *Children writing: reader* (Vol. 1, pp. 21–29). Geelong, Victoria, Australia: Deakin University Press.

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: appraisal in English.* New York: Palgrave Macmillan.

Matthiessen, C. (2002). Combining clauses into clause complexes: A multi-faceted view. In J. Bybee & M. Noonan (Eds.), *Complex Sentences in Grammar and Discourse. Essays in honor of Sandra A. Thompson* (pp. 235–319). Amsterdam: Benjamins.

Matthiessen, C. M. (2013). Applying systemic functional linguistics in healthcare contexts. *Text & Talk*, *33*(4-5), 437–466.

Mautner, G. (2005). Time to get wired: Using web-based corpora in critical discourse analysis. *Discourse & Society*, *16*(6), 809–828.

McCallum, A. K. (2002). *: A Machine Learning for Language Toolkit.*

McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython.* " O'Reilly Media, Inc.".

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... Orwant, J. (2011). Quantitative analysis of culture using millions of digitized books. *science*, *331*(6014), 176–182.

Partington, A. (2010). Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: an overview of the project. *Corpora*, *5*(2), 83–108.

Puerto, S. G. (2012). *Automatic Keyword Generation: Step by Step.*

Sandhaus, E. (2008). *The New York Times Annotated Corpus LDC2008T19.* Linguistic Data Consortium.

Simon-Vandenbergen, A. M., Ravelli, L., & Taverniers, M. (2003). *Grammatical metaphor : views from systemic functional linguistics / edited by Anne-Marie Simon-Vandenbergen, Miriam Taverniers, Louise Ravelli.* Amsterdam ; Philadelphia : Benjamins Pub. Co., c2003.

Widdowson, H. G. (2008). *Text, Context, Pretext: Critical Isssues in Discourse Analysis* (Vol. 12). Wiley. com.

Zimmer, B. (n.d.). *When physicists do linguistics - Ideas - The Boston Globe.*