Discourse-semantics of risk in the *New York Times*, 1963, 1987–2014:

a corpus linguistic approach

Jens O. Zinn
jzinn@unimelb.edu.au

Daniel McDonald
mcdonaldd@unimelb.edu.au

University of Melbourne, Australia

August 3, 2015

Abstract

Since the 1980s and 1990s the notion of risk has become increasingly influential in societal discourses and scholarly debate (Skolbekken, 1995). From early work on risk and culture (Douglas, 1986, 2013) to the *risk society* thesis (Beck, 1992, 2009; Giddens, 2002), from governmentality theorists working in the tradition of Foucault (Dean, 2010; O'Malley, 2012; Rose, 1999) to modern systems theory (Luhmann, 1993) all have built their work around the notion of risk and implicitly or explicitly refer to linguistic changes. Though this body of literature offers different explanations for the shift towards *risk* and its connection to social change, to date there has been no attempt to empirically examine their relative ability to explain this change in the communication of possible harm.

To address this defecit, we conduct a corpus-based 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 unqique 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 and mood systems of language. This toolkit is freely available at `https://github.com/interrogator/corpkit`. Following the corpus interrogation, we use functional linguistics and sociological risk theory in tandem to analyse the findings. First, systemic-functional linguistics 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.

This report is accompanied by an interactive *IPython Notebook* interface to our corpus and developed computational tools. Key findings from this report are stored there, as well as additional information (e.g. concordance lines, keywords, collocations), that could not be included in this report due to spatial considerations. It is available for both interactive and static viewing at `https://github.com/interrogator/risk`.

Contents

Executive Summary

Since the middle of the 1980s sociologists have started to argue that after WW2 significant social transformations took place in most Western industrialised societies which have manifested in a shift towards the risk semantic. Risk has become pervasive in public and scholarly debates and practices in Europe and elsewhere in the world. Following the common assumption that societal changes and language changes are closely linked, this research report aims to contribute to advancing understanding of this social shift towards risk.

There is a wealth of literature and several sociological theories (e.g. risk society, socio-cultural, governmentality, systems theory, edgework) which offer different explanations for the shift towards risk and its connection to social change. Yet, to date there has been little attempts to empirically examine their relative ability to explain this change in the communication of possible harm. This research report addresses this deficit by:

1. Examining a number of claims made by different sociological theories and exploring linguistic changes which might not have been theoretically addressed yet.

2. Developing a corpus based research strategy and computational research tools to examine in detail how the institutional and sociocultural shift towards risk has manifested linguistically.

The study utilises The New York Times (NYT) as a case study for the US to reconstruct the growing usage of the term risk from 1987 to 2014 and examines how discourse-semantic shifts are linked to institutional and socio-cultural changes as well as socially relevant events (e.g. crises and disasters). In addition the study uses a sample of volume 1963 articles of the NYT to contrast the results with much earlier ways of risk reporting.

The investigation involves the creation of an annotated corpus of over 150,000 unique 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 and mood systems of language. This toolkit is freely available at `https://github.com/interrogator/corpkit`.

Following the corpus interrogation, we use functional linguistics and sociological risk theory in tandem to analyse the findings. First, systemic functional linguistics 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 processes and changes.

This report is accompanied by an interactive IPython Notebook interface to our corpus and developed computational tools. Key findings from this report are stored there, as well as additional information (e.g. concordance lines, keywords, collocations), that could not be included in this report due to spatial considerations. It is available for both interactive and static viewing at `https://github.com/interrogator/risk`.

Research Question, hypotheses

Most generally, the study examines *how the institutionalisation of new societal practices manifest linguistically in the change of risk discourses and the use of risk language.* We also examine a number of hypotheses derived from mainstream sociological risk theorising:

1. Is there a shift in discourses from the positive notion of risk-taking to the negative meaning of risk?

2. Can we observe an increasing salience of the at-risk status of social groups?

3. Is there evidence for the assumption of individualised experience of risk?

4. Is there any evidence for the decreasing calculability/controllability of risk?

5. Are there any differences between different social groups in the exposure to risk?

Outcomes

1. The linguistic analyses on the lexicogrammatical level and in discourse semantics give clear indications for a growing routinisation and institutionalisation of risk. We found a growing number of forms such as risk assessment or risk regulator which indicate the institutionalisation of risk practices. Decreasing arguablity and growing implicitness of risk in the clause indicated a shift from the centre to the ancillary parts of the sentence.

2. There are also clear hints that the negative notion of risk is becoming more common, indicated by a clear tendency to nominalisation but also to processes where agency is minimal (e.g. put at-risk).

3. There is a clear trend towards decreasing agency in risk processes from risking and taking risk to running risk and putting at-risk. Thus the strong norm of individual responsibility and risk-taking is accompanied by media-coverage which emphasises lacking agency in risk.

4. There is a clear increase of reporting on issues where people and particular social groups are presented as lacking control especially regarding health issues.

5. Risk reporting, in particular in the health sector, is driven by reference to scientific research supporting a rationalised approach to risk with formalised concepts such as risk factors and reference to research terms on the rise. However, expressions of control such as calculated risk are decreasing while expressions indicating the possibility of negative outcomes are increasing indicating a rise of a possibilistic notion of risk.

6. New coverage in the NYT shows a clear difference between powerful risk-takers and relatively powerless at-risk groups. The difference between the groups is sharpening over time. The powerful take more social risks while the powerless take much more substantial risks often related to illness, injury and death.

Key methodological issues

The study demonstrates the potential for large annotated corpora to examine key claims of sociological theory about social change. It is innovative in creating a corpus comprised by annual subcorpora which allows detailed longitudinal interrogations. It takes advantage of major developments in computational linguistics/natural language processing such as constituency and dependency parsing. Both allow much more detailed and nuanced investigations than generally seen in CADS.

The advantages of the longitudinal analysis of one newspaper had a price, however. It is limited to a case of one genre of newscoverage. Other forms of text are not considered and might show different patterns.

The Language processing tools developed and/or used in this study also have limitations. We have used the *Stanford CoreNLP* NLP suite, which outputs constituency and dependency parses, rather than systemic-functional parsers. Translating the former into the latter poses a number of serious challenges. In some cases, systemic features cannot be automatically recovered from the parses, limiting the kinds of phenomena that can be automatically located and/or counted.

Perspectives

At the time of writing, our investigation is ongoing. In forthcoming work, we seek to:

1. Broaden the empirical basis integrating a growing number of US newspapers, in order to determine whether findings generated here may be generalised across mainstream U.S. print media.
2. Integrate data from newspapers of other countries (e.g. UK, Australia) into our corpus, in order to identify differences in institutional and linguistic practices across Anglophone countries.
3. Perform qualitative analysis of indivudal texts, in order to explore particular observations made in this report and their institutional contexts in more detail.

Chapter 1

Introduction

There is little doubt that risk has become a common experience of our times. This is most apparent in technical catastrophes (e.g. Chernobyl), environmental changes (e.g. climate change), international terrorism (Al-Qaida, IS) and global epidemics (BSE, bird flu), but it is in everyday life as well. We are concerned about whether and who to marry, what to study, which occupation to learn, how to be financially secure in retirement, and even what to eat or drink (Beck 1993, 2009, Giddens 2002). We can increasingly less build on established traditions or generally accepted models of a normal life or intimate relationship. Instead, we have to make risky decisions ourselves and negotiate them with others (Beck & Beck-Gernsheim 2002). There is a vast amount of advisory literature, trainings, and advisors available which all offer their help. Risk and how it is assessed, evaluated, audited and managed have also become institutionalised in both public administration and private companies (Power 1997, 2004). Risk based procedures are in the core of the New Public (Risk) Management (Hood 2001; Black 2005; Kemshall 2002). This focus on risk avoidance is accompanied by the marketing campaigns of the insurance industry to take care for all the undesirable possible events which could happen to us from accidents, to burglary or even death. But is it possible or really desirable to live without risk? Wouldn't life 'be pretty dull without risk' (Lupton & Tulloch 2002)? Some might seek risks in activities such as base jumping (Lyng 1990), aidwork (Roth 2011) or mountain rescue work (Lois 2003) to experience their real self while others might prefer risk free excitement.

The diversity of risk issues makes it even more difficult to understand what drives current debates about risk (Garland 2003). Has risk just become a buzz word in news media that will go away sooner or later as quick as it appeared? Is it too complex to find a shared core of all these ideas of the term risk referring to issues from harm to risk calculation and risk-taking? Is there any good reason that at times where we live in average longer, healthier and wealthier as ever before, risk has become such a common semantic in public discourses and news media? Haven't humans always been exposed to dangers, threats and harm and often their own behaviour was connected to it, whether they exposed themselves voluntary or involuntary to risk, they were aware of the risks or ignorant, whether they calculated them or just hoped that nothing goes wrong? What are the roots of our increased usage of the term risk?

The term 'risk'—as etymological research has shown—became more common in the 14th and 15th century—in particular in marine trading (Luhmann 1993, 9f.; Bonss 1995, 49f.). Luhmann (1993, 10) suggested that the invention of a new term —'risk' —became necessary to express a socially new experience: 'Since the existing language has words for danger, venture, chance, luck, courage, fear, adventure (aventuyre) etc. at its disposal, we may assume that a new term comes into use to indicate a problem

situation that cannot be expressed precisely enough with the vocabulary available'. This new experience was '... that certain advances are to be gained only if something is at stake. It is not a matter of the costs, which can be calculated beforehand and traded off against the advantages. It is rather a matter of a decision that, as can be foreseen, will be subsequently regretted if a loss that one had hoped to avert occurs' (Luhmann 1993, 11).

This was a typical experience for the merchants who sent out their ships to gain huge profit but were always in danger to lose their vessels and face bankruptcy. In response to these challenges they developed early support schemes and later marine insurance. While this notion of risk-taking developed and became common in the transition to modernity, it is only after WW2 that in scholarly work and public discourse the term risk experienced a triumphal procession in contrast to other terms such as danger, threat or harm (Zinn 2010). Why is it that the term risk shows such an outstanding development? How did the term behave and does risk show different behaviour in different contexts? Did it change its meaning as some scholars claimed? What does the term's behaviour tell us about our life and how it is changing?

This project combines the sociology of risk and uncertainty with linguistics to shed light on the historical social and linguistic changes linked to the term risk.

Since the 1980s and 1990s the notion of risk has become increasingly influential in societal discourses and scholarly debate (Skolbekken, 1995). From early work on risk and culture (Douglas, 1986, 1992) to the risk society thesis (Beck, 1992, 2009; Giddens, 2002), from governmentality theorists working in the tradition of Foucault (Dean, 2010; O'Malley, 2012; Rose, 1999) to modern systems theory (Luhmann, 1989, 1993) all have built their work around the notion of risk and implicitly or explicitly refer to linguistic changes. Though this body of literature offers different explanations for the shift towards risk and its connection to social change, to date there has been no attempt to empirically examine their relative ability to explain this change in the communication of possible harm.

This project conducts such an analysis to advance sociological theorising. It takes advantage of two developments: Firstly, in computational linguistics (Baker 2006, CASS 2015) more complex research tools have been developed which allow the analysis of large text data sets ('text corpora'). Secondly, news publishers have advanced in digitising and making publicly available their archives so that they can be used more easily for research purposes.

The project is the first that combines corpus/computational linguistic approaches with risk studies for a detailed analysis of the discursive social change towards risk after WW2 in the US. It breaks new ground by building the first grammatically parsed text corpus of the New York Times (NYT) including all articles containing risk tokens from 1963, 1987 to 2014. The study proves the value of the used methodology, the applicability of a corpus approach for the analysis of long term social change and the fruitfulness of combining linguistic and sociological approaches in a research design to advance understanding of social change.

In the following we first introduce key elements of sociological risk approaches, review empirical shortcomings and present key hypotheses derived from risk theories. We then outline our linguistic approach starting from insights and shortcomings from frame semantics and utilising systemic functional linguistics for a more elaborated analysis of the corpus and the key hypotheses.

In the next step we outline our research design, including the selection of the case study of the New York Times, the building of the text corpus, and the tools to interrogate the text corpus. We follow two major lines of analysis: First, we are looking for different ways in which risk is instantiated, and how these have changed longitudinally. Second, we are also looking for specific claims made by sociological theories to find out whether is any indication that they are correct or require specification.

Our investigation begins with a linguistic analysis of risk language in the NYT, exploring lexical

and grammatical phenomena, and moving freely between different levels of abstraction (from frequency counting to concordancing of linguistic phenomena, for example). Findings from this lexicogrammatical exploration are then abstracted, according to SFL theory, to form a description of the changing discourse-semantics of risk in the NYT. This description is linked to key sociological questions, as well as discussions concerning the extent the linguistic observations can help and inform these social changes.

Given the vast array of changes in the behaviour of risk words uncovered, as well as limitations of time and scope, our analysis is at this stage oriented more toward a longitudinal account of language, rather than sociological theory. We outline a number of promising leads for sociological analysis; developing links between linguistic and sociological reasoning that create pathways for further research and research strategies to answer key sociological questions about social change. We finish with discussing some technical issues and perspectives for digital humanities perspectives.

Chapter 2

Conceptual foundations of our project

1. Theories explaining the shift towards *risk*

Interdisciplinary risk research had traditionally been dominated by technical and psychological approaches examining public understanding and acceptance of risk. Since the 1980s social science have become more influential focusing on the social shaping and construction of risk (Zinn & Taylor-Gooby 2006) and thereby open debates for social explanations for the historical shift towards risk.

Seminal work of Mary Douglas introduced a sociocultural approach focussing on the social values which would determine what risks are selected and which responses are considered appropriate (e.g. Douglas & Wildavsky 1982). Douglas and Wildavsky argue that with the social institutions we also select the risks we are concerned about. Real dangers would be transformed into risks for the institutions and value of a social unit such as particular social group, organisations or society. They developed the grid-group schemes that provide a number of ideal types to characterise the different empirically observable cultures. They distinguished their types on two dimensions: the extent to which individual identities and cultural outlooks are fixed and predetermined and the amount of control members accept (grid-dimension), and second by the degree of commitment or solidary individuals exhibit or feel towards a social group (group-dimension). The combination of these two dimensions resulted in four ideal types. Hierarchy, is characterised by a high predetermination of identities and control and a high degree of commitment to the group, which is typical for organisations such as the military, the Catholic Church and traditional bureaucracies. Conversely, in markets we find low predetermination and control and relatively little commitment to the group. This second ideal-type is termed individualism. Low predetermination and control of individual identity but high degree of commitment and solidarity characterises the third ideal-type, egalitarianism, which is typical for grassroots movements and communal groups. The fourth type, fatalism, occurs when predetermination of identity and control is high but solidarity and commitment are low, and when people are rather isolated and lack influence and commitment to a social group. On this basis, Douglas (1990, 15f.) argued that in an individualist culture, social concerns focus on the risks associated with the competitive culture of markets where the weak and the losers have to carry the blame for their failure. In a hierarchical culture, the focus is on social risks and those who deviate from the dominant social norms tend to be blamed for risks. In an egalitarian culture there is the tendency to focus on natural risks and to blame the system and faction leaders. In the fatalist perspective only fate, itself, is blamed.

Douglas' argues on the basis of her earlier anthropological work (1963, 1966) that concerns regarding

dirt and pollution are less about bacteria, viruses, or pollutants and more about socio-symbolic disorder than the lack of control of a group's boundaries. The control of the body and its margins serves as a symbol for controlling the rules that define a social group. The reality of danger becomes important for a social group as a threat to its boundaries, orders, and values. Similarly the modern notion of risk in secularized societies serves to protect the boundaries of social groups.

However, Douglas (1990) suggested, the term risk has changed its meaning in contemporary western societies. It is no longer a neutral but a negative term. Risk has come to mean danger and 'high risk means a lot of danger' (Douglas 1990, p. 3).

Ulrich Beck introduced the most influential risk society theory with a focus on the impact of new risks which accompany successful modernisation processes. Economic, scientific and medical advancements manifest not only in increases in average wealth and health but also new risks. Beck and also Giddens emphasise that the modern world is characterised by risks which are increasingly produced by humanity itself rather than exposed to us by the environment (Beck 1992; Giddens 2002). As a result it was mainly up to us to deal with the risks and uncertainties of our world and the reality of increasingly self-produced risks and uncertainties ('manufactured risks' and 'manufactured uncertainties'). Nature and our environment would increasingly be experienced as shaped by humanity (e.g. climate change, genetic engineering). Beck also argued that individualisation processes would transform a society stabilised by traditions into social forms characterised by individual decisions (Beck 1992, 2009). But this is an ambivalent process. The social expectation that people should take on the responsibility to individually plane and shape their life were emphasised at a time of growing social complexity, instability and volatility. Individual control of outcomes becomes even more unlikely. As a result of new risks and individualisation risk and uncertainty have become a common experience of our time. It is characterised by increasing social and individual responsibility at a time when growing complexity and uncertainty makes it even more difficult to plan and shape our future purposefully.

Following Michel Foucault's work, a number of scholars understand risk as characterising a new way of governing societies on the basis of normative discourses of individual responsibility and improvement on the one hand and calculative technologies such as statistics and probability theory on the other (e.g. Dean 1999; Rose 1999; O'Malley 2004). In the governmentality perspective, risk is not so much about the reality of risk but a specific form to govern societies utilising calculative technologies and normative discourses of individual self-improvement and responsibility (e.g. Dean 1999). Framing the world in terms of risk was an expression of a new form of discursive power in late modern societies. Statistic probabilistic technologies are only part of these discourses among others though an important one. It produces an own rationality. Individuals are no longer approached as a whole but defined by particular factors which determine their status regarding particular regulative action (e.g. offenders risk to reoffend defined by a number of risk factors). How social groups are defined by such factors in tandem with normative discourses of individual self-improvement is a key theme of research in the governmentality tradition.

In addition to these mainstream approaches, Niklas Luhmann's systems theory emphasising the new ways how the shift from stratificatory differentiation to functional differentiation contributes to increased debates about risk (Luhmann 1993). Social systems would try to get shift risk to other areas which would result in risk conflicts and negotiation of risk.

Finally, Steven Lyng's work on edgework engages with forms of voluntary risk taking (Lyng 1990, 2005). He emphasises that people take risks out of desire to experience their real self, unmediated through social norms. He also considers the possibility that in advanced modernisation the ability to take risks and to deal with highly complex and volatile situations becomes a socially desirable skill. People's

increasing voluntary risk taking would therefore be an expression of common normative expectations.

It is likely that all these social changes proposed by different social theories manifest themselves somehow linguistically in everyday life usage of language as much as in the news media. However, these approaches have usually not explicitly developed a theory or explicit hypothesis about the connection between social change and linguistic change or news reporting. Usually the media get involved in risk studies since we do mainly only know about many risks because they are reported in the media and the media are often accused of contributing to public's distorted knowledge about risk (e.g. Jensen et al. 2014). It has also become common to emphasise the media's own rational in news production (Kitzinger & Reilly 1997; Kitzinger 1999) and the need to examine the missing link between risk studies and media studies to understand social perception and responses to risk (Cottle 1998; Tulloch & Zinn 2011). For example, at times of restricted financial resources, using scientific press releases, which typically contain some risk language, might the most efficient strategy to produce news.

There is also a lot of research how particular risks have been communicated and presented in the media such as climate change. However, there has not yet been an analysis that—similarly to Luhmann's claim of the condition of a new semantic—examines how risk is utilised and behaves relatively independent from trends in the media (e.g. increasing nominalisation) and how broader social changes and significant events might influence the way how risk is understood and used in discursive practice—not only but also in the media. However, when sociology scholars referred to issues such as the common usage of risk language they often rely on more or less anecdotal analysis of historical and semantic change (e.g. Luhmann 1993, Giddens 2002; Beck 1992) or invent examples themselves (Lupton 1999).

2. Empirical evidence and shortcomings

Claims about historical social change made by Ulrich Beck (Beck 1992, 2009) with the most famous risk society thesis are based on general observations and exemplary evidence. It is difficult to provide comprehensive empirical evidence for the kind of structural and institutional changes Beck addresses with his theory. Tracing such changes through detailed analysis of linguistic changes could be a valuable strategy to show systematically in which societal areas social changes have manifested.

Detailed historical studies on risk are mainly provided by researchers from the governmentality perspective (Ewald 1986; Hacking 1991; Valverde 1998; but compare: Strydom 2002; Gamson 1989). They produce valuable knowledge on the prerequisites for, and impact of, the introduction of statistics and probability calculation, and how they contribute to the governing of societies. These studies are convincing in the reconstruction of changes in institutional risk practices by specific area- or case-studies. They contribute less, however, to our understanding of how these developments compete with or complement others, and how they combine to influence a general shift in the communication, comprehension and semantics of risk in the media.

Many theorists claim that the media are particularly influential in social risk discourse (Beck 1992), though conceptualisations like risk society are criticized for being undifferentiated and ignoring current trends in media research (Kitzinger & Reilly 1997; Cottle 1998; Kitzinger 1999). Media-oriented risk research mainly examines specific events or debates, such as Mad Cow Disease (BSE), genetically modified food, or international terrorism, and how news and risks are produced by the media (e.g. Allan, Adam & Carter 2000). It does not reconstruct how risk enters the media and how the understanding and usage of the term may have changed over time. Even the most recent special issue 'Media and Risk' in the Journal of Risk Research (vol.13, no.1) ignores this important aspect. One exception is Mairal (2008). He has

reconstructed how risk discourses developed over time in Spain and showed how earlier experiences and symbolical representation of risk influenced later discourses; but he did not examine semantic changes of the term risk.

There are strong streams of risk research on technological risk and risk assessment, health, social work and insurance. Authors such as Strydom (2002) claim that the nuclear power debates and technological risk analysis has been the major drivers for increasing concerns about risk. Similarly Beck (2009) focuses on new technologies as the driver for the growing anxiety about our future. This might underpin the different conceptualisation of risk as unexpected harm, part of statistic probabilistic calculation, or a conscious decision. Differing from Beck, first analyses have shown that in media discourses the risk semantic is less used in articles describing new risks. Instead, a majority of articles are on health and illness and economics (Zinn 2010, p.111f.).

Many linguists are interested in overcoming the strong focus on language in discourse analysis and in incorporating social dimensions (e.g. Van Dijk 1997, Wodak & Meyer 2001). In general, this stream of research has contributed little to the reconstruction of the historical development of discourses (Brinton 2001; Harding 2006; Carabine 2001) although many cognitive linguists examine long term semantic changes (e.g. Nerlich & Clarke 1988, 1992, 2000; Traugott & Dasher 2002). Regarding risk, corpus linguists have shown that sociologists' assumptions about the usage of risk are often informed by everyday life knowledge rather than systematic empirical analysis of how the term risk is actually used (Hamilton et al. 2007). Frame Semantics has provided a detailed analysis of the available risk-frames (Fillmore & Atkins 1992); but neither approach examines historical changes of the usage and notion of risk.

In interdisciplinary risk research there is a long-standing body of research focusing on risk communication between decision-makers and the public (e.g. Kasperson & Stallen 1991). This research has produced valuable knowledge about how to improve the communication of risk, while media coverage is discussed from the point of view of the public's risk perception (Bennett & Calman 1999; Slovic 2000). Some typical patterns of risk reporting are identified as well as factors which amplify and attenuate the communication of risk (Kasperson et al. 1988; Pidgeon et al. 2003; Flynn et al. 2001). However, this research contributes little to a historical perspective of how the risk semantic became pervasive in daily newspapers.

For risk research, the phase after WW2 has been identified as particularly crucial in the establishment of increasing debates about risk and the success of the risk semantic. Other semantics such as 'threat' had its establishing phase between WW1 and WW2 during which it established as a most common term in New York Times' newspaper coverage which remains relatively stable after WW2 (Zinn 2010, p.117).

The triumphal procession of risk took off before iconic events such as the Chernobyl disaster or the 9/11 terrorist attacks took place. A more systematic analysis of the dynamics of the usage of the risk semantic would allow a detailed understanding of how our framing of the future in terms of risk was influenced by different forces and events.

Originally, social science debates had been dominated by the introduction of nuclear power and the social controversies accompanying them (Douglas & Wildavsky 1982; Perrow 1984; Beck 1992; Luhmann 1993). However, the debates about DDT-based insecticides had driven public conflicts much earlier. The publication 'The silent spring' (Carson 1962) did not trigger social science risk debates and did not stand out in the early debates of Douglas (1985) and later of Luhmann (1993) and Beck (1992) on risk. One reason might be that the semantic grounding of a risk framework had not been established at the time.

However, there are clear indications that the risk semantic and related discourses using a risk frame became increasingly dominant during the 1980s. With our study we wanted to examine in more detail how institutional social change manifests in language. We assumed that fundamental changes such as

towards a society increasingly concerned with self-produced risk would manifest in linguistic patterns even in a single genre such as print news media (similar to the Books of Manners in Norbert Elias' study). Building on an exploratory study that only counted the numbers of articles where a risk token was used at least once, Zinn (2011) provided evidence that even during a relatively short period of 1987 to 2014 we should be able to identify relatively short-term social changes within language.

3. Central hypotheses in risk studies

In order to test the applicability of our research strategy and computational tools for historical research we derived some central hypotheses from mainstream sociological risk theories.

First, a number of approaches frame risk not only as possible harm but a calculative technology to estimate possible harm and through this knowledge manage it. For example, in the governmentality perspective risk is considered a calculative technology (e.g. Ewald) which is used to manage harm. Similarly, in the risk society perspective insurance and science are characterised by risk calculation to minimise risk. In the risk society perspective the calculability of risk characterises first modern experience. If risks are not directly controllable by science/knowledge we still have the opportunity to manage them by insurance, for example. However, both governmentality theorists and risk society researchers have emphasised that uncertainties and non-knowledge would increase and we would observe a shift from the calculability of risk to the potentiality of harm. If this is correct it is more likely to find phrases which indicate the pure potentiality of risk rather than the calculability of risk.

Second, the risk literature about societal changes has also emphasised that the experience of risk has started to change during modernisation on another dimension. The positive side of risk as risk-taking would lose influence (Douglas 1990, Lupton 1999). Risk would mainly mean harm or danger and verbal forms involving an active decision to take a risk for something positive would decrease. We might even observe within the verbal forms a shift from positive risk taking to a pure exposure to risk where the possible gain disappears. For example, the notion of taking a risk or running a risk might increasingly be supplanted by notions of exposure to risk.

Third, governmentality theorists have claimed in recent decades that a neo-liberal agenda has become more dominant that shifts responsibility to individuals and the expectation that individuals actively make decisions and take risks. If this is correct we would expect more individualised phrases which express more active risk-taking.

However, Beck (1992) claimed that in recent decades one has to understand and act as an individualised planning office exactly at times where knowledge and control of the future is limited. That means an active risk taking citizen is expected at a time where it is even more unlikely that an individual can control outcomes. For such a contradictory situation we would expect less the communication of self-confident decision making and risk taking but individualised suffering and exposure to of all kinds of risk.

This would support the suggestion of social policy researchers that risk is increasingly shifted from organisations and institutions to individuals (Hacker 2006: risk shift). It is important to see that this happens as a legitimate shift not something what happens against public resistance. At least in a country such as the US where individual action is highly valued we would expect that this shift takes place legitimately and deeply rooted in the societal institutions. Rather than as a surprise it would be a consequent development following an already prepared path. As a result we would expect not only as a rational of consequent media reporting that individual stories are but to sell to the public but that more

generally the individual exposure to risk rather than individual agency would be emphasised.

Fourth, Ulrich Beck claimed in the chapter Beyond Class and Status in his famous book Risk Society that social inequalities and disadvantage would increasingly be framed in individualised terms. That means that risk is no longer attributed to social class or status but to social groups which are at-risk because of their particular behaviour rather than class affiliation. Researchers examining shifts in public/social policy and social work support Beck's suggestion and claim that social institutions would increasingly use practices that identify social groups at-risk on the basis of particular indicators which then characterise particular groups such as drug users, homeless, fatherlessness etc. as at-risk groups which require regulation, support, encouragement or protection. If Kemshall and others are correct that risk thinking has become a common societal practice this should be reflected in media coverage.

We would expect that groups reported on in the media are identified and reported about using their at-risk status rather than social class affiliation or general socio-structural conditions which influence their behaviour or shape their living conditions. Such generalised factors would be rather silenced or made invisible. We would expect that it is increasingly likely that we find groups characterised by attributed risk status.

Fifth, there is a tension in the debates about risk in the literature. Relative powerful middle class people are assumed to be individualisation winners, that means they have agency and can make decisions while more disadvantaged people lack agency and are approached by the state, encouraged or more broadly managed. They have a more intrinsic quality of being at-risk. For example, drug users might be a population at-risk by social definition. We would expect finding a clear distinction between powerful risk takers and powerless at-risk or vulnerable groups identified and characterised by a specific variable or characteristic.

Finally, we not only examine whether we can say something about these hypotheses. We will also use the data to explore and generate new insights with the help of the corpus linguistics research tools.

4. Linguistic concepts for researching risk

Recently, with rapid technological developments in the digitisation of historical newspaper archives and the computational analysis of text data, it has become possible to examine long term changes in media reporting and using the media as a source for analysis of long term societal changes. Accordingly, our research takes advantage of sophisticated linguistic tools for the analysis of long-term social change—a research agenda with roots not only within the media but in the larger (social) world which effects and shifts the lexis and grammar used when reporting risk.

Central to any well-considered study of language use is a theory of language, which may either implicitly or explicitly inform the kinds of analyses being done. A number of frameworks exist for connecting lexis and grammar to functional meanings. Notable within risk research has been frame semantics, which has been used to characterise risk as one or more cognitive frames/schemata involving a number of possible components, such as risker, risked thing, chance, and positive/negative outcomes. This theory has then been put to use within corpus linguistic approaches to risk, which have used large digitised datasets to understand how the risk frame(s) are typically constructed. Despite successes within this approach, it remains limited by the fact that corpora seldom provide researchers with opportunities to confirm cognitive hypotheses regarding the intentions of the writer, or the comprehension of the reader
.

Another popular functional linguistic framework is systemic functional linguistics <see>hallidayintroduction2004,

which conceptualises language as a sign system that is employed by users in order to achieve social functions. While sharing a functional view of language (as opposed to formalist views proposed by (e.g.)), SFL is a functional-semantic theory, rather than a cognitive-semantic one. While the remarkable achievement of frame semantics is its mapping out of cognitive frames, we are largely unable to operationalise these with our dataset, as we have little information regarding the specific interactants (writers and readers) of the original texts. Moreover, cognitive understandings of text are complicated in situations where the text's author is producing the text within an institutional context, for a readership. Without downplaying the potential importance of cognitivist accounts of risk, we have instead opted here to focus on risk words as instantiations of parts of the linguistic system for the purposes of meaning-making, rather than as a representation of the cognitive schemata that underlie our behaviour.

A second benefit of SFL for our purposes is that it provides the most detailed functional grammar of English (Eggins, 2004): when compared with frame semantics, it provides a more rigorous description of how risk can behave lexicogrammatically—that is, in relation to both other words and grammatical features—within a clause. This makes it possible to search parsed texts in nuanced ways.

The third benefit of SFL is that it provides not only a grammar, but a conceptualisation of the relationship between text and context. A foundational tenet of SFL, and a point of departure from other linguistic theories, is the notion that we can create a description of context based solely on the lexicogrammatical content of the text. This is particularly suitable for us, given that our texts arrived to us abstracted from their original contexts. This context was then further obscured through the parsing process. As such, SFL provides an ability to account for discourse-semantics using corpora that other theories cannot.

In many respects, the major challenge of this project has been to find ways how to combine a linguistic analysis that goes beyond tallying the co-occurrence of lexical and grammatical features with the sociological understanding and analysis of long-term social change. As a linguistic theory that provides a taxonomy of both language and context, SFL practitioners have to date been reluctant to engage with conceptualisations of context from other traditions within the Humanities and Social Sciences. This is disappointing, especially when considering that the most common criticism of SFL is that its theory of context is heavily influenced by its theory of grammar: in SFL, context is divided into three major dimensions (*Tenor*, *Field* and *Mode*), which are essentially projections of a language's major grammatical systems (*Mood*, *Transitivity* and *Theme*).

5. Our research approach

The social sense-making processes of risk depends on risks being communicated. Though people experience risk when they manifest personally, since risks are usually expectations towards the future, the social process that shape these expectations are crucial. Even when we make personal experiences it depends on broader social processes whether we interpret a hot summer as an indication for climate warming or just a normal variation.

Communication is mediated through language, and language is by no means restricted to a neutral communication of knowledge or information about events and happenings in the world. Language may shape what seems possible as much as what seems appropriate or inappropriate. It both constructs and responds to all kinds of information about the context in which it has been generated, the values underpinning it, the power structures it reproduces or is structured by (sociolects; gendered language, etc.). Since language is such a rich resource for communicating information about social reality, it is also

data that can be used to examine social change (e.g. Norbert Elias' historical analysis of the books of manners to examine the civilisation process).

The media plays an important role in communicating social life. It not only influences but also reflects what is considered important at a historical point in time not only in the form of selecting particular content but how it is presented. A careful analysis of linguistic change therefore requires not only investigation of what has been communicated through language, but how it has been communicated and how both lexis and grammar have changed over time in the communication of issues such as risk.

Sociology, linguistics and media studies provide slightly different concepts of both 'context' and of the forces that influence the selection and communication of social issues such as risk. Sociology is interested in wider and long term social changes. In a historical perspective, the focus is on how institutional and sociocultural social changes are reflected in the use of language. Sociologists are well aware of that the use of language is, for example, influenced of the social milieu a person is part of (e.g. working class, middle class) and such a context manifests not only in the content but also the form and the use of grammar of language. For sociologists, contexts and events within contexts are not necessarily socially triggered or caused. But how they are dealt with is mediated through language. The suppression of women in a society might be openly debated or not talked about. It might even be engraved in a language, where masculine nouns and/or pronouns have historically also been used to refer to general populations (e.g. A giant leap for mankind) or singular entities whose gender is unknown.

In many branches of functional linguistics, the understanding of context focusses on text. A particular text can be analysed regarding its form and structure and its origin. Through linguistic features of texts alone, genre can often be clearly determined. Whether a text is a newspaper article or a university lecture, a talk of a party leader to party members or a general public, can often be determined simply through an analysis of the lexis and grammar in a text, as well as the way in which stages of the text are ordered. The larger social conditions and how these might have influenced the content and use of language are less commonly examined. Despite increasing awareness and sensitivity to context in functional linguistics, context is more commonly operationalised as observable constellations of variables of a given interaction (speaker demographics, spoken/written, formality, etc), rather than as a set of broader social movements, ideas and values. Even researchers within systemic functional linguistics (SFL), which at one time explicitly attempted to delineate the relationship between realised language and social class and ideology, have revised the conceptualisation of context to exclude ideology as the greatest level of observable abstraction. Long-term historical analyses remain centred on language, and empirically driven attempts to connect language change to broader social change are exceptionally rare.

This is not to say that there is no value of linguistic theory and methods for the purposes of understanding the changing status of risk in society. In fact, the opposite is the case: linguistics (in our case, SFL) provides a framework for delineating the kinds of changes that risk language undergoes. For example, in order to understand how risk language has changed, we must first distinguish between risk as a participant within a communication about the world (The risk was serious) and risk as a process (Lives were risked). Our addition to more standard linguistic methods is not that we abstract the significance of linguistic changes—as this is a common task within linguistic discourse analysis—but rather that following from an abstracted discourse-semantic analysis of risk, we abstract again, to consider the influence of factors beyond what is captured within linguistic taxonomies of context.

Media studies are positioned in between sociological and linguistic approaches. Discourse analyses using media or print media often focus on content and the positive or negative representation of issues. These studies do often not go into further detail regarding long term linguistic changes. They tend to focus on short term ways of representation of issues such as climate change. However, media studies have

also raised awareness of the organisational and social context that shapes how news are produced (e.g. free press or more or less controlled press; economic pressure; political bias). Research has examined the production process of news and how this process follows an own logic of newsworthiness that influences which issues enter the media and which not. There is also awareness that there are events and dimensions of change which are not reported in the media. Not everything is newsworthy and what is selected follows the own media production logic of news. In this respect media reporting is selective and it is difficult take stock of the aspects which have not been reported without looking beyond the media. These issues must be identified and approached differently. For example, it is important for linguistic research of texts alone to acknowledge that such approaches may not be able to consider what drives the media agenda and which kinds of texts might be systematically included/excluded as a result of unobserved institutional and contextual factors.

However, since the media are part of social change, it reflects as much as influences social changes, and, accordingly, can be used to examine long term social change. Since many risk issues are newsworthy, we can expect to find a lot risk communication, which allows us to examine the changing practice of risk reporting and the use of the risk semantic. Broad changes in the relationship between news institutions and risk communication (e.g. which risks are considered, how they are reported, etc.) are so general and part of more generally changing discourses and linguistic practice that they will affect newspapers as well since they have to appeal to the public.

Given the novelty of Big Data and Big Data methods, investigations such as ours involve the development of theoretical frameworks for linking instantiated language to discourse-semantics. In our case, this involved a thorough investigation of the lexicogrammar of risk language in news journalism. In this report, we map out strategies for engaging with the systemic functional notion of experiential meaning primarily through complex querying of constituency parses. In terms of the systemic functional conceptualisation of the Mood system as a resource for making interpersonal meanings, as well as the notion of arguability, we demonstrate novel strategies of exploiting dependency parsing provided by the Stanford CoreNLP toolkit. Though existing automated parsing generally cannot provide the level of depth necessary for full systemic annotation of language, the partial account that can be provided still proves sufficient for connecting lexicogrammar to discourse-semantics in a rigorous and systematic fashion.

As these new methods involve automated analysis via computer programming, our project also contributes to methodology via a repository of code for manipulating large and complex linguistic datasets. This repository, though designed for our particular investigation, is readily reusable by other researchers interested in how language is used as a meaning-making resource. Our methodological work is available open source at `https://github.com/interrogator/risk`. Documentation and code used to build and annotate the NYT corpus is also freely available there.

Chapter 3

The case study: *The New York Times*, 1963, 1987–2014

1. Selecting *The New York Times* as a case study

There is good evidence that the risk semantic has become more common in societal discourses and practices. A direct count of articles which contain a risk token at least once showed how the dynamic of risk developed in many countries after WW2 (Zinn 2011). It clearly shows how risk is mainly a phenomenon that developed a particular dynamic after WW2 in particular in the late 1980s. It also shows that the risk semantic had been around for quite a while without a clear dynamic. This is interesting and invites more long term investigations.

With the current study we wanted to examine in much more detail whether during a historical relatively short period from 1987 to 2014 (we used a sample of the 1963 volume to contrast with the later years) significant shifts can be observed using much more sophisticated research strategies than used in earlier corpus based approaches on the risk semantic (e.g. Hamilton et al, 2007).

Therefore we selected only one newspaper—The New York Times—as a case study after careful consideration of other available resources. We aimed to find a resource that allows longitudinal analysis of long term social change with a limited number of intervening factors. We were looking for a paper which provided a high quality digitised archive and a central news institution over the centuries. The (London) Times and the NYT seem suitable because of their important social role within a society. They also fulfil further selection criteria such as wide circulation (not just regional), good accessibility and high data quality. However the NYT has been finally selected because of the central role of the US in the world and the prestige and clout of the NYT. The NYT is a historically central institution of media coverage (Chapman 2005) with a continuously high status and standard of coverage. It is influential, highly circulated and publicly acknowledged news media. It contains extensive coverage of both national and international developments, its digital archive covers all years since WWII and is relatively easy to access. Available Australian Newspapers such as The Australian or The Age offer similar digitised archives only for recent decades and at higher cost. Long term historical analyses are much more complicated and will be pursued when we have proven our methodology. The project concentrates on a single newspaper and follows a reproduction logic (Yin 1989) for four reasons:

1. The 'historical change of concepts' (Koselleck 2002) is so general that it can be identified even in

specific newspapers though newspaper specific factors have to be considered.

2. A detailed analysis of available newspapers archives by the CI has found that, in the US, only the Washington Post provides a comparable archive. While both show no significant differences in the general increase of the usage of the risk semantic (Zinn 2010, p. 115), access and data management has proven easier and more reliable with the NYT.

3. The case study allows a more detailed analysis of how the change of the newspaper might have influenced the use of risk. A collection of newspapers, as in many linguistic text corpuses would not lead to representative results but would create uncontrolled biases. Instead, the case study of a specific newspaper allows a much more detailed analysis of how change of the newspaper itself, such as a change in leadership or style of news reporting, might have influenced the use of risk.

4. The study limits the amount of data and restricts costs without losing significant outcomes. Originally we wanted to compare the volumes 1963, 1988, 2013 of The New York Times. We soon found out about the availability of a high quality data resource, The New York Times Annotated Corpus (http://catalog.ldc.upenn.edu/LDC2008T19) which covers all articles published from 1987–mid-2007 and includes substantial metadata and contains 1,130,621,175 words. We complemented this dataset with articles from the NYT online archive up to 2013/14. In order to further validate our results, future research has been planned that will compare our results with more recent data from other US newspapers. Though in the US many newspapers are digitised the main issue is that some papers are strictly PDF while some of these PDFs have the plain text version also available. We identified major newspapers which are suitable for comparative purposes in future research.

2. Building the *Risk Corpus*

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).

We relied on two sources for our data. *The New York Times Annotated Corpus* 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.

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]

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 3.1: Metadata tags and content

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.[2]

Article text and any available metadata were extracted from this unstructured source content using *Python*'s *Beautiful Soup* library 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). These metadata fields provided both human-readable information for use during qualitative analysis of texts, and machine-readable information that could be used to restructure the corpus in future investigations.

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.[3]

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 a specific field of discourse. As a case study, we decided to focus on health articles. Librarian-added metadata concerning article topic/category (MC metadata field) was used to locate all articles tagged with the case-insensitive Regular Expression `\ bhealth.*`.[4]

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.

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**

Table 3.2: 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 data-source.

After all data processing, we had a 150 million word corpus of nearly 150,000 unique 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 health subcorpus contained a subset of 8,524023 words, 6,944 articles and 36,547 risk words. A breakdown of the size and composition of each annual subcorpus is provided in Table 3.2. During analysis, when conducting absolute frequency analysis, 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.

```
<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 3.1: Example file: NYT-1995-12-30-10.txt

3. Tools and interface used for corpus interrogation

Special tools needed to be developed to work with the very large dataset of both raw NYT articles and parsed paragraphs containing a risk word. Given a well-established history of use within humanities and social sciences, as well as a particular strength in working with linguistic data, we developed a Python-based toolkit for querying our data and visualising query results. Our purpose-built toolkit provided the ability to quickly search each subcorpus of our data, edit the results from our searches, perform concordancing and thematic categorisation, and generate visualisations of results. Though many parts of the toolkit were designed with more general Digital Humanities projects in mind, certain components of the toolkit were designed exclusively to aid in our particular investigation (projection of counts from 1963 and 2014; automatically stripping names and titles from U.S. politician names, etc.). The most important functions and their purpose are outlined in Table 3.3, with a simple example of a function shown in Figure 3.2. More detailed explanations and demonstrations are provided at `http://nbviewer.ipython.org/github/interrogator/risk/blob/master/risk.ipynb`; the repository of code itself is available via *GitHub* (`https://github.com/interrogator/risk`), where it can freely be downloaded, or duplicated and modified.

Function name	**Purpose**
`interrogator()`	interrogate parse trees, find keywords, collocates, etc.
`plotter()`	visualise `interrogator()` results
`quickview()`	view `interrogator()` results
`editor()`	edit `interrogator()` results
`conc()`	complex concordancing of subcorpora
`collocates()`	get collocates from corpus/subcorpus/concordance lines
`quicktree()`	visually represent a parse tree
`searchtree()`	search a parse tree with a Tregex query

Table 3.3: Core Python functions developed for our investigation

Finally, we developed an IPython Notebook based interface for using these functions to investigate the NYT corpus (also available via our GitHub URL above). This served not only as our main platform for interrogating the dataset, but also as a means of dynamically disseminating results without being

```python
def ngrams(data,
           reference_corpus = 'bnc.p',
           clear = True,
           printstatus = True,
           n = 'all',
           **kwargs):
    """Feed this function some data and get its keywords.

    You can use dictmaker() to build a new reference_corpus
    to serve as reference corpus, or use bnc.p

    A list of what counts as data is available in the
    docstring of datareader().
    """

    import re
    import time
    from time import localtime, strftime
    import pandas as pd
    try:
        from IPython.display import display, clear_output
    except ImportError:
        pass
    from corpkit.keys import keywords_and_ngrams, turn_input_into_counter
    from corpkit.other import datareader
    from dictionaries.stopwords import stopwords as my_stopwords

    loaded_ref_corpus = turn_input_into_counter(reference_corpus)

    time = strftime("%H:%M:%S", localtime())
    if printstatus:
        print "\n%s: Generating ngrams... \n" % time
    good = datareader(data, **kwargs)

    regex_nonword_filter = re.compile("[A-Za-z-\']")
    good = [i for i in good if re.search(regex_nonword_filter, i)
            and i not in my_stopwords]

    ngrams = keywords_and_ngrams(good, reference_corpus = reference_corpus,
                                 calc_all = calc_all, show = 'ngrams', **kwargs)

    out = pd.Series([s for k, s in ngrams], index = [k for k, s in ngrams])
    out.name = 'ngrams'

    # print and return
    if clear:
        clear_output()
    if printstatus:
        time = strftime("%H:%M:%S", localtime())
        print '%s: Done! %d results.\n' % (time, len(list(out.index)))
    if n  == 'all':
        n = len(out)
    return out[:n]
```

Figure 3.2: Python function for getting n-grams from corpus data

limited by considerations of space. In being open-source, and in explicitly showing the exact queries used to generate findings, the Notebook ensures both reproducibility and transparency of the entirety of our investigation. At the same time, it provides a framework for sophisticated corpus-assisted discourse analysis using cutting-edge digital research tools. Researchers are encouraged to run the Notebook in conjunction with this report, so that they can generate and manipulate our key findings as they see fit.

Chapter 4

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.[5] 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:

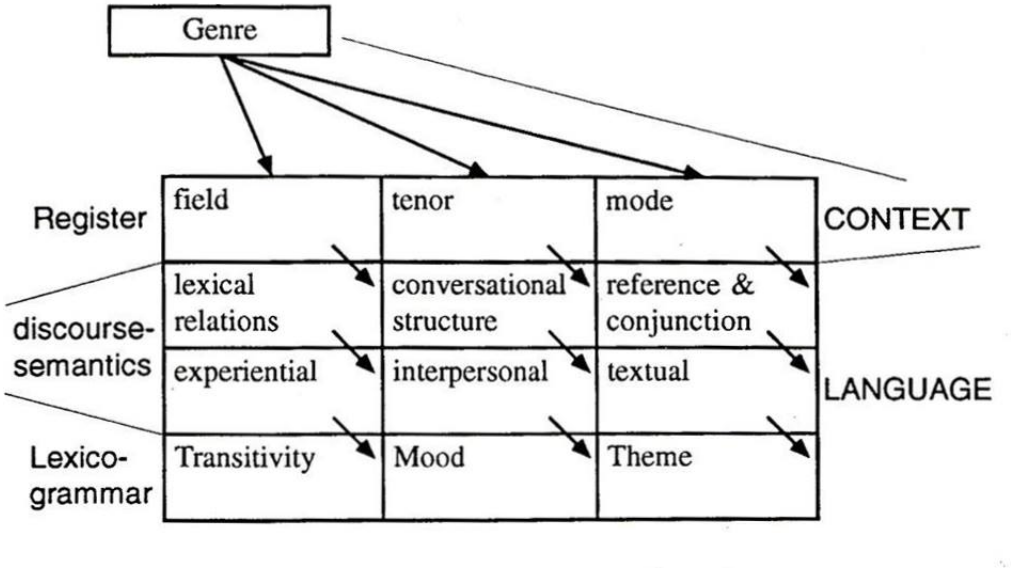


Figure 4.1: Strata and metafunctions of language (from Eggins, 2004)

'*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).[6]

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 being different ends of the same stratum of language: 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 (generally simultaneous) searches of both grammatical and lexical features of sentences containing one or more risk words.

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) (?, ?). 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.[7]

2.1. Risk and the experiential metafunction

In SFL, experiential meanings are made via the transitivity system. Transitivity analysis of a clause involves breaking it down into its *process*, *participants* and *circumstances*, realised congruently 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, and is responsible for selecting the ways in which these participants are realised: 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 4.2) 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 4.2: 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'.

Clause complex
Clause
Group/phrase
Word
Morpheme

Table 4.1: Rank Scale in SFL

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. 365). First, through *grammatical metaphor*, semantic processes may be realised grammatically as participants ('I accepted *the invitation*') for the purpose of 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 4.5 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.

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.

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 4.3). 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.[8] 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 4.3, 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 4.3: 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,

Role	Arguability	Example
Subject	Very high	*For Mobic, the **risks** of heart attack and stroke rose 37 percent, Dr. Graham's study showed.*
Finite/ Predicator	High	*But candid talk about job prospects and debt obligations **risked** the wrath of management, she said.*
Complement	Medium	*This approach holds some **risk** for a union boss.*
Adjunct	Low	*The wire is stretched very tautly, and we are at some significant **risk** it will snap from overload.*

Table 4.2: Arguability of risk words in differing mood constituents

Role	Arguability	Example
Head	Higher	'*So far, pregnancy **risk** does seem to come with this class of drugs*,' Ms. Glynn said.
Non-head	Lower	*They purchased billions of dollars in **risky** subprime mortgages.*

Table 4.3: Arguability of risk words as either head or non-head

p. 117). In the contexts of proposals and commands, it is the one who is supposed to perform the action (*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).

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 and Predicator positions would indicate that risk is discussed and argued about. A shift from Mood Block to Residue (especially Complement and Adjunct positions) 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.

The ways in which we operationalise the notion of arguability while interrogating the parsed data are outlined in Section 10.

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).

Our study is in the latter domain.[9] 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 (approximately 300/over

150,000). Fourth, they relied on collocation (without lemmatisation[10]), 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, as in Figure 4.4). 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 pre-existing 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.

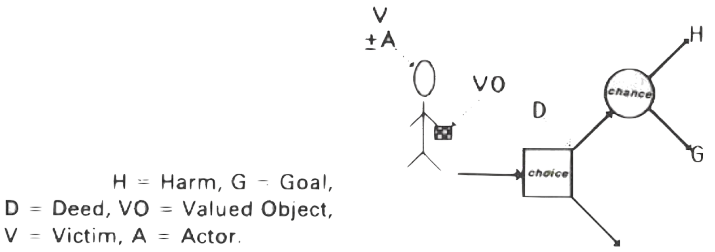


Figure 4.4: Risk frame (from Fillmore & Atkins, 1992)

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 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.[11] 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 our investigation did not typically involve common CADS practices such as keywording, clustering, collocation, or the use of stopword lists. Our reasons for avoiding these practices are varied. In the case of keywording, for example, we found the notion of using reference corpora comprised of 'general' language to be inherently problematic. 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. When the size of our corpus permitted, 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. In the case of smaller subcorpora, when lexicogrammatical queries typically did not yield large sets, though keywording was performed, we searched for keywords in each subcorpus, and used the entire corpus as the reference corpus, to avoid potential epistemological issues of non-representative or unbalanced data.

`__ >># (/(NP|VP|PP)/ > (VP <<# processes.relational $ (@NP <<# /(?i)\brisk.?/)))`	In relational processes in which a risk word is the Token/Carrier, what is the head of the Value/Attribute?

Figure 4.5: *Tregex*-based search query and gloss

Clustering and collocation, though mainstays of CADS, are also largely 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 between lexical items 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 4.5). 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, in absolute terms, by word class, and by experiential and interpersonal role. 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 discourse-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[12] 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 (see Matthiessen, 1995).

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. Because our analysis began with broader sites of change and progressed toward more micro-features upon discovery of interesting initial results (e.g. from the increasing frequency of risk as a modifier to the frequency of the modifier *at-risk*), it is possible that some micro-level features were obscured by the lack of significant change at broader levels.

Chapter 5

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.

Summary: an example

Summaries of each major finding will be presented in highlighted text boxes.

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, are available online, at `https://github.com/interrogator/corpkit` and `https://github.com/interrogator/risk`. A non-interactive version is available at `http://nbviewer.ipython.org/github/interrogator/risk/blob/master/risk.ipynb`. This Notebook does not suffer from spatial limitations, and thus contains additional information, including the exact Tregex queries used in interrogations, as well as complete lists of the concordance lines discussed only briefly here. Tools and results from other kinds of corpus linguistic analysis, such as keywording and collocation, are also available there, but have not been described here.

1. How frequently do risk words appear?

The first point of interest was the overall frequency of risk words in the NYT (Figure 5.1) and the distribution of risk words by word class (nominal, verbal, adjectival/adverbial), absent any consideration of surrounding grammar (see Figure 5.2). In terms of the relative frequency of risk words, we note a general upward trend, with a number of peaks and troughs worthy of further investigation. In terms of word classes of risk, we found that not only are nominal forms by far the most common in the NYT, but that it is nominal risk words that vary the most in frequency, with the other categories remaining more or less stable. Interestingly, in the span for which we have no data (1964–1986), adjectival forms overtake verbal forms of risk in frequency.

Figure 5.1: Relative frequency of risk words

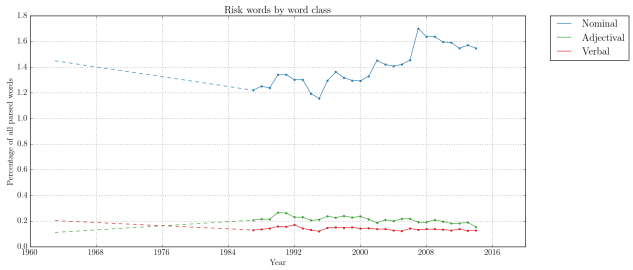


Figure 5.2: Relative frequency by word class

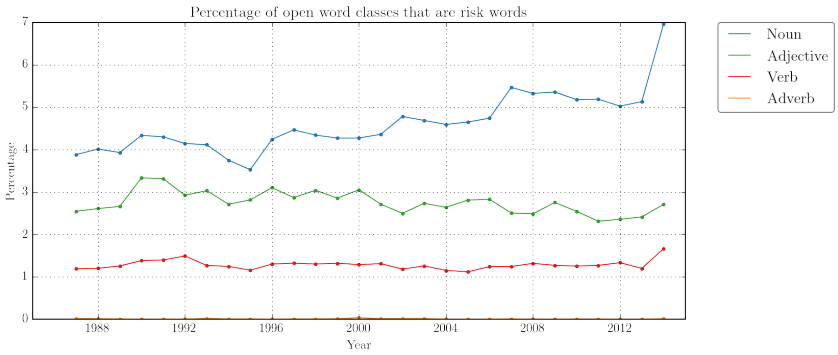


Figure 5.3: Percentage of each open word class that are risk words

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 our dataset more generally (Figure 5.2). This showed that even when compared to potential trends toward nominalisation generally (Figure 5.3), nominal risks are becoming more frequent in later NYT editions.

This is an important preliminary finding: verbal groups form the nucleus of experiential meanings, the shift toward nominal risk is in effect a shift toward risk discourse where risk is not the central event being depicted, but is instead a player within a broader range of events. Furthermore, given that nominal risk is closer to synonymous with harm or threat (see Section 2.1, shifts toward nominal risks are indeed evidence for Lupton's claim that the semantics of risk increasingly reflect negative outcomes.

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*, it functions as a modifier, rather than a participant. Accordingly, in our analysis, functional categories are typically preferred over formal categories where possible.

A similar question is the number of unique risk words appearing per year. Figure 5.4 demonstrates that there is a general increase in the relative number of unique risk words over time.[13]

Summary: frequency of risk words

Risk words appear to be increasing in relative frequency, with modest increases in the number of unique risk words per year.

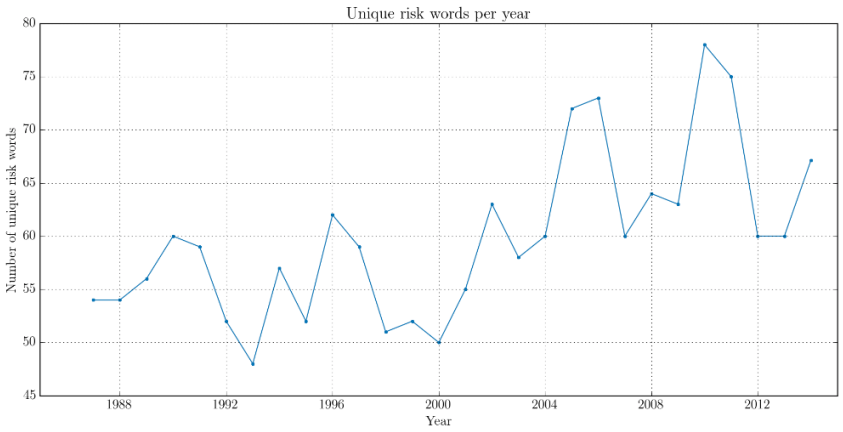


Figure 5.4: Unique risk words

2. Which experiential roles do risk words occupy?

In a systemic-functional conceptualisation of the experiential metafunction of language (that is, the ways in which language is used to communicate who did what to whom), risk words may take the form of a Participant (*The risk was there*), Process (*I risked it*) or a Modifier (*a risky encounter*). Though these pattern to some extent with word classes (e.g. *participant = noun, process = verb, modifier = adjective*), word classes on the whole are a poor indication of functional role, especially in genres such as print news journalism, which rely heavily on nominalisation and grammatical metaphor to pack large amounts of experiential information into each clause. As shown in Table 5.1, for example, nominal risks commonly perform Modifier functions, and adjectival functions often perform Participant functions.

Using Stanford CoreNLP's dependency parses, we counted the frequency of risk words within these three functional roles (Figure 5.5). In line with the results from word-class based searching, we find that risk as a Process is declining in use. Risk as Modifier, patterning in part with adjectival risk, appears to be increasing. That said, we can also see here the affordances of a functional grammar in corpus assisted discourse research: in this case, much richer evidence of changing usage of risk can be found through an understanding of its semantic function rather than its word class alone.

The decline of risk as a Process can be read as a shift away from the 'risk frame' identified by Fillmore and Atkins (1992). Their frame is essentially a mapping of the possible kinds of participants that can occur when risk is used as a process. In a very typical risk process, such as *He risked everything*, *He* is the Actor and everything is the *Valued Object*. Risk as participant is less likely to explicitly index the major components of the risk frame: in *The risk must be weighed against the benefit*, we are not given any specific information about the configuration of the risk scenario. (Risk as modifier encompasses a number of functional roles. As such, we leave this discussion for sections 8 and 9.)

Like with our analysis of the word classes of risk, we can also understand the shifts in the experi-

Example	**Word class**	**Experiential role**
It was risky	Adjective	Participant
A risking of lives	Noun	Nominalised process
Risk management	Noun	Modifier

Table 5.1: Key differences between word class and experiential role

ential roles of risk words as evidence for an increasing implicitness of risk in NYT news discourse. As mentioned earlier, in a systemic-functional conceptualisation of language, the process is the locus of experiential meaning—it selects the kinds of participants that can occur as its arguments. Modifiers and circumstances are the least consequential kinds of experiential meaning, as they provide ancillary or supplementary kinds of meaning, regarding the manners in which processes were performed, or characteristics of participants.

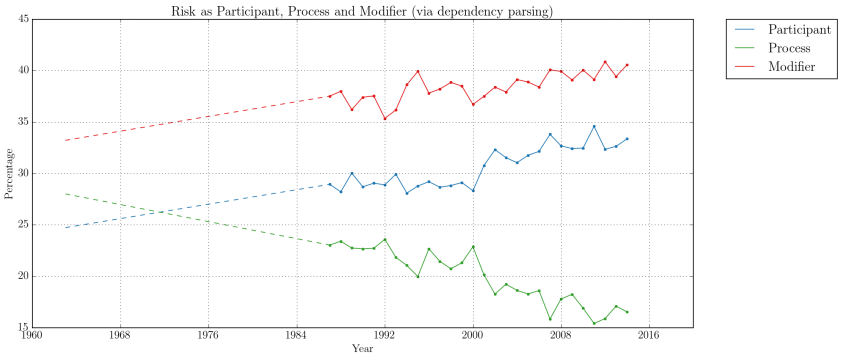


Figure 5.5: Experiential roles of risk words

Summary: experiential function of risk words

Risk as a process is declining in use, and has been overtaken in frequency by risk as a participant.

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.[14] Our first area of interest was the proportion of each, with respect to general trends in the NYT. As shown in Figure 5.6, 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.

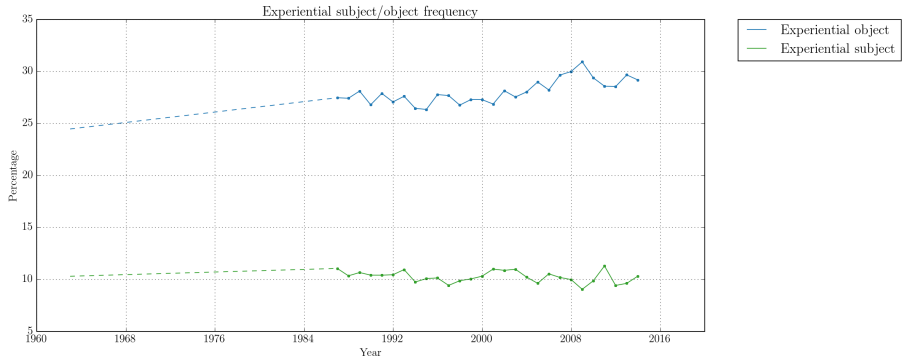


Figure 5.6: Risk as experiential subject and object as percentage of all risk roles

As subject	As object
1. But the most prevalent **risk** for the average traveler to Peru is the high altitude of the Andes	1. The company has resolved accounting problems, he said, and stabilized profit margins, while new management has reduced the company's **risks**
2. The **risk** would be that the stock would recover during the period that the investor was out of the stock	2. But an empty village is a big **risk**.
3. But the **risk**, though very small, that a man facing execution could win a new trial raises the question why this rule has proved so hard to follow	3. They said there was only a little **risk**, and now he 's not with us anymore

Table 5.2: Examples of risk as experiential subject and object in 2001

Summary: risk as experiential subject/object

Risk is more often an experiential object than an experiential subject. The gap has widened considerably over time, pointing to a decreasingly agentive role of risk as participant.

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 5.3 and 5.4 show the top twenty processes for risk as experiential subject and object, taking passivisation into account.[15]

Interesting here is the dominance of processes seeking to quantify risk (*increase*, *outweigh*, *rise*, *grow*). Also salient is the presence of a large set of mental processes (*seem, appear, assess, understand, accept*). This may be seen as evidence for an increased demand for the control of risk (see discussion in sections 9; contrast with section 5)).

Summary: processes with risk participants

When risk is a participant, quantification is often at the centre of the experiential meaning. The high proportion of mental processes highlights a portrayal of risks as perceived.

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 5.3: 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 5.4: 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 5.5 and 5.6 show the most common pre- and post-head modifiers of risk as a participant.

Some of these modifiers are undergoing longitudinal trajectory change. As can be seen in Figure 5.7, *calculated risk* has an outbound trajectory, decreasing steadily. The large number of occurrences projected for 1963, however, is partially 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 increasing in frequency. Also interesting is the spike in the *high risk* construction between 2002–2004. Here, concordancing reveals links to particular events. *High risk*, peaking in 2004, is asociated with the outbreak of the H5N1 avian flu outbreak:

1. Mr. Johannessen said health care providers had a moral obligation to ensure - through direct questions and, if necessary, medical records - that people who asked for flu shots were at high risk.
2. Dr. Anthony S. Fauci, director of the National Institute of Allergy and Infectious Diseases, said that nearly 90 million Americans had a high risk of catching flu, with half of that number usually seeking vaccinations.
3. Nearly 90 million Americans are at high risk to contract a potentially fatal case of influenza.
4. Dr. Hinds said his county had about 90,000 people at high risk for flu.

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 5.5: 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 5.6: Pre-head modification of participant risk

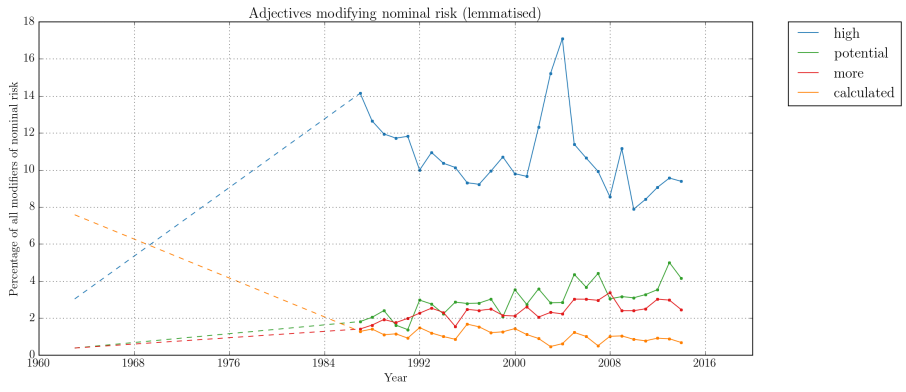


Figure 5.7: Selected modifiers of participant risk as percentage of all risk modifiers

Summary: modifiers of risk as participant

Calculated risk has been overtaken by *potential risk* in overall frequency. *High-risk* spikes in frequency in references to H5N1.

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 five 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). Fourth is *putting somebody/something at risk*, which involves an obligatory nominal object argument and a prepositional-phrase complement. Finally, we have *pose risk*, which differs again in terms of participant roles: the entity who may suffer a negative outcome is realised circumstantially (*The legislation poses no risk to federally licensed facilities*), with the nature of the potential harm generally left implicit.

Other *process + nominal risk word* constructions sit on the cusp of being recognisable risk processes: *to carry risk*, for example, is frequent in the data, but we have not included it because we feel that the semantic burden of this process still lies in *carry* (unlike *pose* in *to pose risk*).[16] Our first interest is

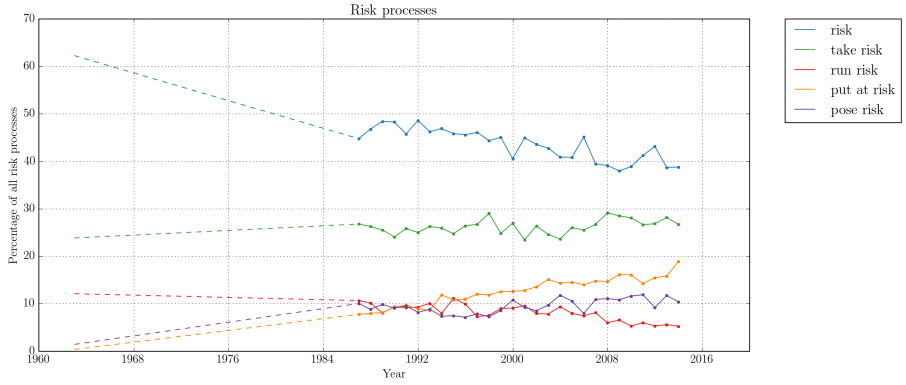


Figure 5.8: Risk processes as percentage of all parsed processes

the overall frequency of these five risk processes. Figure 5.8 charts the trajectory of the five identified risk processes. Most interesting here are that the 'standard' (i.e. predicatorial) risk process is steadily decreasing, in favour of the other processes, each of which seems to provide additional connodations of the agency of the risker as well as his/her/its understanding of the level of risk.

The second notable finding here is that *putting at risk* has overtaken *running risk* in frequency. Concordancing revealed that in 2014, *putting at risk* is used in cases where the potential harm is either implicit (left) or explicit (right):

1. Ultimately, there is a price to pay: If you attack our soldiers, you're putting yourself at risk.
2. But addicted health care workers need not be physicians to put patients at risk.
3. While obviously no airline or company deliberately puts people at risk, 'sometimes new risks are identified and steps have to be taken,' Mr. Koch said.

1. The auction houses deny that they are trimming profits with givebacks or putting themselves at financial risk.
2. Rather, such tax status is generally put at risk when groups stray from their mission.
3. They had handled her body, putting them at serious risk of infection.

That said, we also noted that there seems to be some evidence for lessening agency in recent *risk running* processes. Compare 1963 (left) and 2014 (right) results:

1. However, if adults decide to run a risk, this is up to them, and anyway, Switzerland adequately handles American affairs in Havana.
2. In Washington at the weekend it was pretty welt agreed that the MIG incident was not deliberate provocation; the feeling was that, even with the Russian presence, Castro would not wilfully run the risk of American retaliation.
3. If he sticks to the more-or-less official Republican position against off-track betting, he runs the risk of losing thousands of New York City votes, which he needs.

1. Fans see this revolving door of injuries with so much regularity that they run the risk of becoming desensitized
2. 'One runs the risk of falling for a voice.'
3. 'I would run the risk of having two boys,' she said.
4. On the other hand, if Argentina does default, it runs the risk of more lawsuits, said Siobhan Morden, head of Latin America strategy at Jefferies.
5. And, like an overdressed beachgoer, a classic cocktail served straight up runs a high risk of wilting in the sunshine.

Overall, the shift in both the semantics of risk running and the increasing preference for *putting at risk* can be seen as evidence for decreasing agency in risk, as well as an increasing implicitness of the potential harm. This finding is especially significant, given that the existing descriptions of risk (Fillmore & Atkins, 1992), as well as the current FrameNet database, include accounts of *running risk* as a frame, but not *putting at risk*.

Summary: types of risk processes

Both *pose risk* and *put at risk* have overtaken *run risk* in frequency. Use of the prototypical risk process, *to risk* is declining. Finally, there is some evidence for reduced agency the *run risk* process.

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. In frame semantic terms, experiential subjects

of risk processes can often be mapped to *actors*.[17] Experiential objects are either *valued objects* or *harm* (*they risked their lives/death*). Table 5.7 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 5.8 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 5.7: 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 5.8: 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. Even embedded processes lacking explicit connotations of power are typically deployed in the contexts of government, industry or society. Table 5.8 shows concordance results for *risk alienating* in 2013, which appears 14 times.

There is some evidence everyday people are less commonly riskers in later editions: if we sort the 'riskers' (Figure 5.9) by those that are decreasing at the fastest rate over the course of our dataset, *man/men*, *woman/women* and *person/people* are all in the top seven. Conversely, when sorting by those on the most upward trajectory, the lemmas *bank*, *company*, *firm* and *agency* are in the top seven (Figure 5.10). Further findings in this vein are presented in more depth in Zinn & McDonald, 2015.

stoked further concerns that unemployment risked	becoming	endemic and could eventually cause social upheaval
franchise, the stage scene could have risked	becoming	an embarrassment for the brand, but Mr. Timbers'
with locally, or else the Vatican offices risk	becoming	institutions of censorship
on which the experience depends – or risk	becoming	irrelevant to future generations, Mr. Staggs said
restart growth, warning that the euro area risked	becoming	mired in the same kind of economic stagnation that
If left unaddressed, such practices risk	becoming	more and more entrenched, Ann Harrison of
without serious savings in this area, we risk	becoming	an unbalanced force, one that is well compensated
Switzerland risks	becoming	one of the most restrictive places for management
What was the exception before now risks	becoming	the standard practice
the pope's new remarks that the church risked	becoming	a 'small chapel' overly fixated on sexual
hailed the step as significant, it risks	becoming	the latest of many tentative moves toward talks
Rather than race the clock to Bed-Stuy and risk	becoming	an early bike-share casualty, I stopped at a
increasingly turning to what, strangely, risked	becoming	the most marginalized group of all: the bosses
and currency crisis in the European Union risks	becoming	a crisis of liberal democracy itself

Table 5.9: *To risk becoming* in 2013 subcorpus

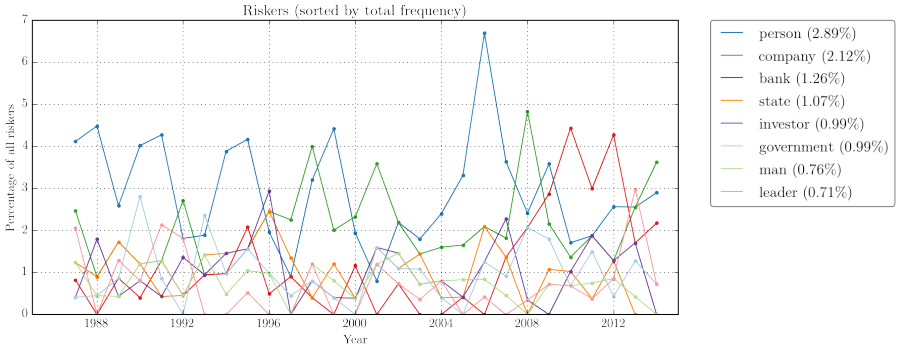


Figure 5.9: Relative frequencies of common riskers

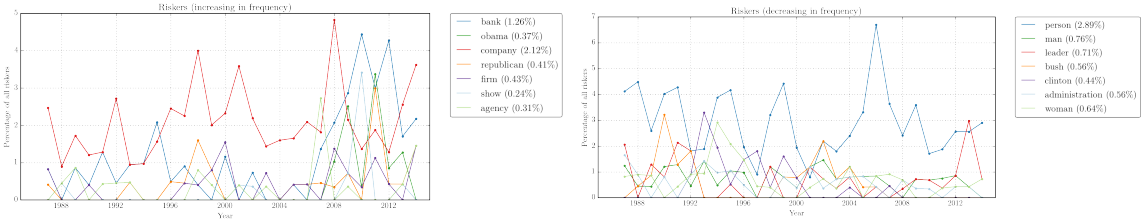


Figure 5.10: Riskers, sorted by most increasing (left) and most decreasing (right)

Summary: participants in risk processes

When risk is a process, risked things/potential harms often pertain to individual health (*to risk life, death, health, etc.*). This contrasts with processes as potential harm, which generally relate to people in positions of power (*to risk alienating voters*, for example).

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

There are many different kinds of risk as modifier (see Table 5.10 for a non-exhaustive list of examples). Our first interest was in gauging the prevalence of the different forms. From this query, we noted that pre-head nominal modifiers are increasing in frequency. A good example is *risk factor* (see Figure 5.12, Table 5.11), which typically occurs within fields related to health.

Modifier type	Example
Adjectival pre-head	*a risky move*
Post-head	*A person at risk*
pre-head nominal	*risk management*
Adverbial	*to riskily act*
Circumstance head	*to be at risk*

Table 5.10: Types of risk-as-modifier

0	and many had no obvious	risk	factors
1	states have periodontitis , and it is a known	risk	factor for atherosclerosis , the buildup of plaque
2	large sugary drinks are not the main	risk	factor for obesity
3	it makes sense , she said , to treat this	risk	factor as early as possible , even if not everyone
4	smoking and drinking were considered the dominant	risk	factors for cancers of the throat
5	inflammation leads to atherosclerosis , a known	risk	factor for stroke and dementia
6	bones after age 50 or those with significant	risk	factors for fracture
7	about capitalization or operations , no ‘	risk	factors ’ - the sort of thing one typically
8	selling stock to make some disclosures about	risk	factors and debt that were not explicitly required
9	is dementia in his family , and cardiovascular	risk	factors are also risk factors for dementia

Table 5.11: Randomised concordance lines for *risk factor* in 2013

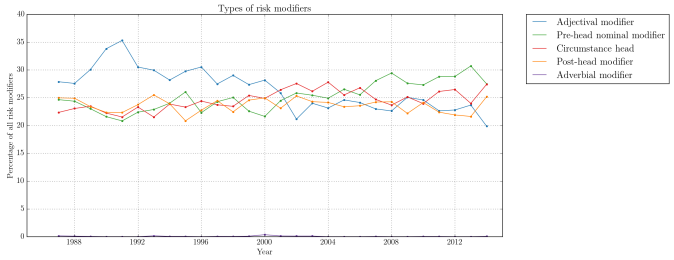


Figure 5.11: Types of risk modifier

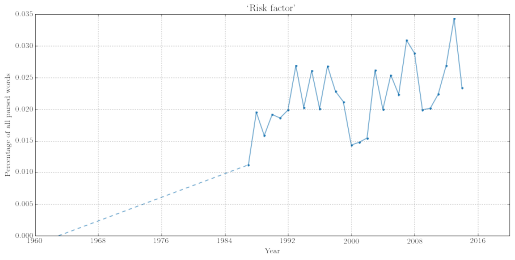


Figure 5.12: Relative frequency of *risk factor*

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 5.13). 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.

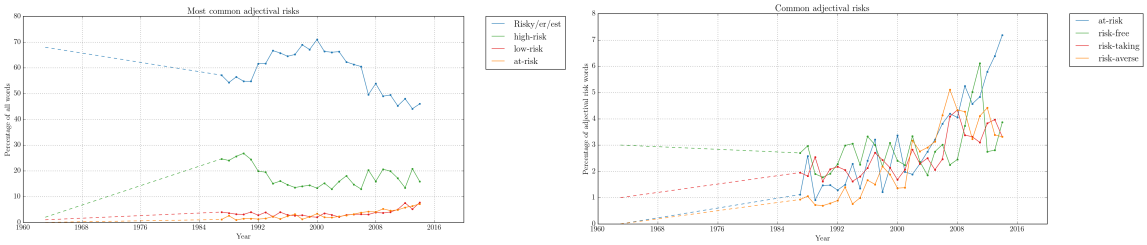


Figure 5.13: Common adjectival risk words as percentage of all adjectival risks

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.

Summary: frequencies of modifier risk

Common risk modifiers (*risky*, *riskier*, *riskiest*) are gradually being displaced by a number of less common constructions (e.g. *low-risk*, *at-risk*, *risk-averse*, *risk-free*). This reflects increasing nuance and complexity in the way risk is used to modify participants in discourse.

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 5.12 and 5.13, 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 mainly comprised of vulnerable populations of people (*women*, *children*, *students*). In contrast, as can be seen in the concordance lines presented in Table 5.14, risk modifiers attaching to financial participants are generally either bare adjectival forms (*risky*, *riskier*, *riskiest*), or quantified variants (*low-*, *high-*, *higher-risk*).

In need of further research is whether or not the list of entities that can sensibly be modified by *at-risk* is beginning to grow: since the U.S. subprime mortgage crisis (beginning in 2007), references to *at-risk homeowners* appear to be on the rise. Results from 2011, for example, show that *nations* and even *economic sectors* are being modified with *at-risk*:

1. `Mr.  Obama asked for $400 million for the World Bank's clean technology fund, $95 million for the bank's program to prevent deforestation and $90 million for its program to help at-risk nations cope with the effects of a warming planet by, for instance, developing drought-resistant crops.`
2. `The most at-risk sectors included auto components and automobile companies, which generate nearly 30 percent of their sales in Europe, as well as food and tobacco firms.`

Note that it is difficult to reconcile the semantic meaning of *at-risk* constructions with the semantic frame of risk provided by Fillmore and Atkins (1992). Though elements of both the VICTIM and VALUED OBJECT appear to be at work, neither provides an adequate label for *at-risk people, children, homeowners or nations*. Rather than being an oversight during the articulation of the risk frame (recall Figure 4.4), in light of the increased use of these kinds of constructions since the mid 1990s, we hypothesise that *at-risk* constructions (as well as *to put at risk*) are demonstrative of a broader shift in risk discourse toward general clusters of negative outcomes, rather than specific and measurable potential harms. Connection between this shift and sociological theory is made in the following chapter.

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 5.12: 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 5.13: Most common at-risk participants in the corpus

Summary: participants modified by risk

While risk as a modifier is often used in the context of finance/commerce, *at-risk* typically attaches to vulnerable human demographics.

0	the big banks to take on riskier and riskier	business	that could end up destabilizing the financial
1	that banks internalize the costs of their risky	business	rather than have them borne by the rest of society
2	any such risks , broadway being a risky enough	business	as it is
3	facing pressure from lawmakers to discourage risky	behavior	by bankers that might fuel another financial
4	global markets , driving money toward less risky	investments	like treasury securities
5	investors are feeling cautious and want low-risk	investments	that produce steady income
6	's long-stated intention to shrink its riskier	businesses	in favor of the steadier , and more consistently
7	adolescents are more likely to engage in risky	behavior	when in groups
8		investments	relatively unpalatable
9	returns , a pension fund that is full of risky	investments	will swing heavily , soaring in value when the
10	mr. kane said bailouts should be viewed as equity	investments	whose risks deserve a return to taxpayers of at
11	regulators to reduce their involvement in riskier	businesses	and to have more capital on hand to weather
12	-quarter profit rose 7 percent as it reshaped its	business	to reduce risk and better navigate future
13	families to put their money into more risky	investments	like real estate , stocks and lightly regulated
14	the prevalence of anxiety disorders and risky	behavior	-lrb- both of which reflect this developmental
15	the masks can fog up , making a risky	business	of , say , inserting an intravenous line or
16	the campaign behind the new report , called risky	business	, is funded largely by three wealthy financiers
17	and children have become a high-profit , low-risk	business	for mexican narcotics cartel bosses who , chief
18	population that most often engages in high-risk	behaviors	like driving fast
19	have not banned banks from engaging in higher-risk	businesses	like money transfers to certain countries , they

Table 5.14: Randomised instances of *investment(s)*, *business(es)* and *behavior(s)* modified by *risk* in 2014

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. Figure

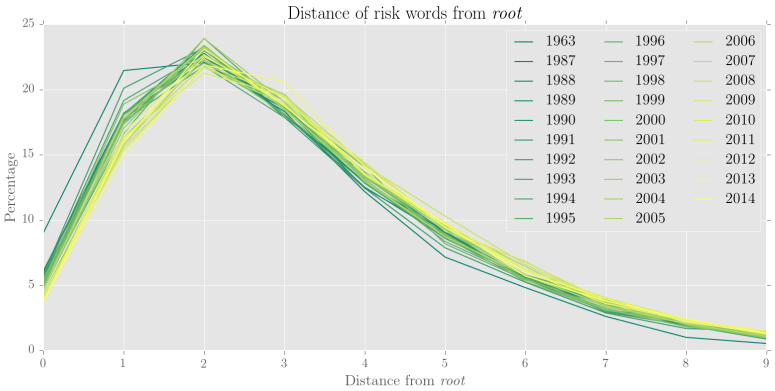


Figure 5.14: Distance from *root* of risk words by year

5.14 shows that in early samples, risk occupies core roles within the dependency hierarchy, and thus sits closer to the core part of the meaning being exchanged within the clause. In later samples, risk more commonly occurs later in the dependency structure, in less focial positions. As explained earlier, though this experimental method is not a perfectly reliable indicators of arguability, it does indicate an

Dist. from *root*	Example
0	By failing to get on board with social networking sooner, Google **risked** being left behind.
0	John Paul **risked** his life attending an underground seminary as priests he admired were killed with the Jews.
0	In such cases, the woman **risks** a prison sentence of up to two years.
0	The release of high-level Taliban leaders from Guantanamo would certainly **risk** a political backlash in an election year.
0	Doctors who use the word 'obese' in their notes may **risk** alienating patients.
1	'Generally all vehicles have some **risk** of fire in the event of a serious crash.'
1	It may do so again despite its misgivings, because the alternative of an uncontrolled default is too **risky**.
1	There is **risk**, because Montero, for all his defensive questions, could be a star.
1	But with all the amenities that modern N.F.L. sidelines have these days, the **risk** of frostbite for players is minimal.
1	But Latino political leaders say the **risk** in changing the questions could create confusion and lead some Latinos not to mark their ethnicity, shrinking the overall Hispanic numbers.
11	The use of air power has changed markedly during the long Afghan conflict, reflecting the political costs and sensitivities of civilian casualties caused by errant or indiscriminate strikes and the increasing use of aerial drones, which can watch over potential targets for extended periods with no **risk** to pilots or more expensive aircraft.
12	An article on Saturday about a moratorium on research involving a highly contagious form of the H5N1 avian flu virus misstated the professional affiliation of Dr. Anthony Fauci, who said the scientific community needed to clearly explain the benefits of such research and the measures taken to minimize its possible **risks**.
13	They are even at odds with Pope Benedict XVI, who has approved the use of condoms 'in the intention of reducing the **risk** of infection.'
14	The United Nations Convention Against Torture prohibits the transfer of a detained person to the custody of a state where there are substantial grounds for believing that the detainee is at **risk** of torture.
15	The trading blowup that followed has now become a flash point in the fierce debate over the Volcker Rule, which would ban banks from trading with their own money in an effort to prevent them from placing **risky** wagers while enjoying government backing.

Table 5.15: Examples of risk words in near and far from *root*

increasing preference to position risk as non-core, ancillary information, rather than as the main thing which is under discussion (See Table 5.15).

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 long list of functional roles. Of key interest, however, are risk words at the head of each major component of the Mood system—Subject, Finite/Predicator, Complement and Adjunct (CoreNLP parses unfortunately do not distinguish between Finite and Predicator in a reliable way, so the categories are collapsed here). From Figure 5.15, we can see that risk is shifting from Subject and Finite/Predicator to Complement and Adjunct roles. This is an important result: risk words in more arguable roles are steadily decreasing, while risk in less arguable roles are becoming more common. Like earlier findings, this suggests an increasing implicitness of risk in NYT discourse, with less talk actually *about* risk, but more talk where the relationship between risk and the subjects of the talk is assumed to be more or less common knowledge.

Summary: risk and arguability

Longitudinally, risk words are shifting to less focal parts of clauses. We can approximate these changes using both indices or semantic function information within dependency parses.

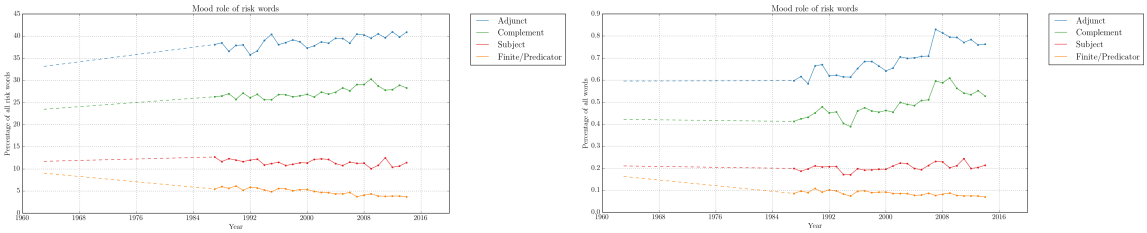
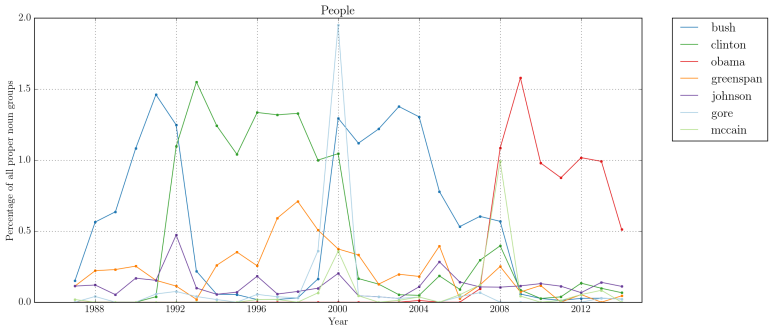


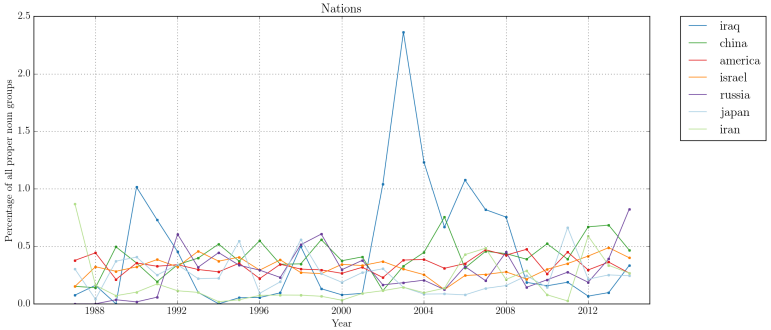
Figure 5.15: Frequency of risk words for each Mood component as percentage of all risk words/all parsed data

11. Risk words and proper nouns

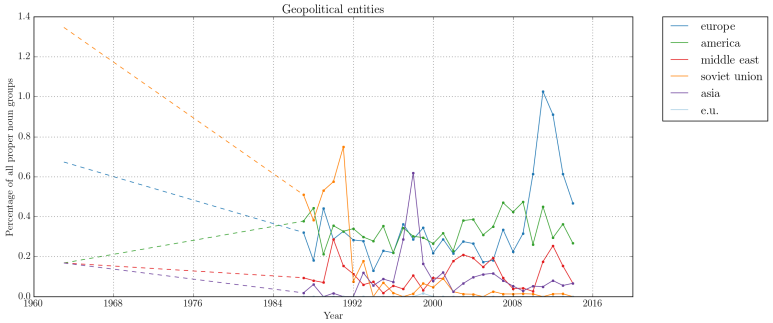
We searched for proper noun groups in parse trees containing a risk word. This is a departure from many of our earlier queries, as here we are looking only at which entities co-occur with risk words, rather than determining how risk words and non-risk words relate to other another lexicogrammatically. The result of this query was 68891 unique 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: *People*, *Nations*, *Geopolitical entities*, *Companies*, *Organisations* and *Medical themes*. The results were then plotted (see Figure 5.16).



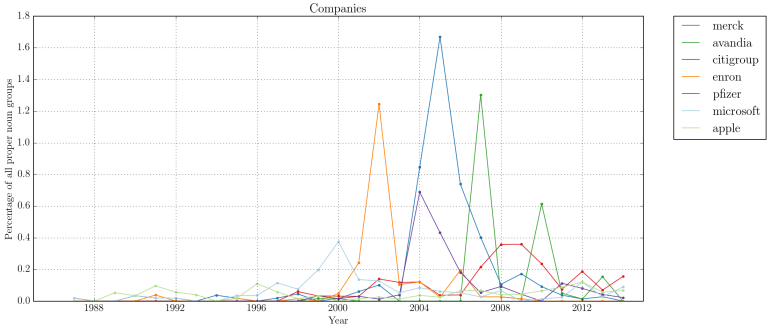
(a) People



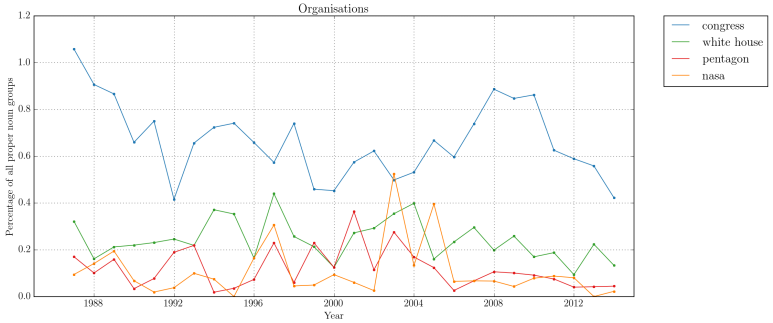
(b) Nations



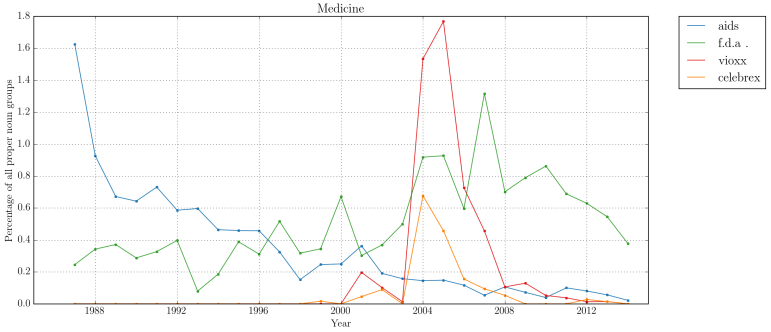
(c) Geopolitical entities



(d) Companies



(e) Organisations



(f) Medical terms

Figure 5.16: Proper noun groups co-occurring with risk

A number of historical events were easily recognisable within the peaks and troughs of these charts. Key events represented through these interrogations include:

1. US presidents and presidential candidates[18] (Figure 5.16a)
2. The First Persian Gulf War (Figure 5.16b)
3. The Iraq Wars (Figure 5.16b)
4. September 11 and the War in Afghanistan (Figure 5.16b)
5. The beginning of the 2014 Crimean crisis (Figure 5.16b)
6. The Asian financial crisis (Figure 5.16c)
7. The breakup of the Soviet Union (Figure 5.16c)
8. The Eurozone crisis (Figure 5.16c)
9. The Space Shuttle Colombia Disaster (Figure 5.16d)
10. The collapse of Enron (Figure 5.16e)
11. The U.S. subprime mortgage crisis (Figure 5.16e)
12. The U.S. outbreak of HIV and the AIDS crisis (Figure 5.16f)
13. The recall of Vioxx (Figure 5.16f)

This area of our investigation is perhaps the most promising as a means of connecting risk language to particular people and events.

Spatial considerations have precluded a full treatment of the charting of risk language to specific events, despite the fact that enough data exists for detailed analyses of any number of potential foci. Future research that centres on detailed exploration of health domains (including the Vioxx recall) is planned.

Summary: risk and proper nouns

We can use proper nouns to see which people, places and things co-occur with discussion of risk.

12. Summary of key findings

We found that the behaviour of risk words has changed longitudinally in a number of key respects:

1. Risk words appear to be increasing in relative frequency, with modest increases in the number of unique risk words per year.
2. Risk as a process is declining in use, and has been overtaken in frequency by risk as a participant.
3. Risk is more often an experiential object than an experiential subject. The gap has widened considerably over time.
4. *Calculated risk* has been overtaken by *potential risk* in overall frequency. *High-risk* spikes in frequency in references to H5N1.
5. When risk is a participant, quantification is often at the centre of the experiential meaning. The high proportion of mental processes highlights a portrayal of risks as perceived.

6. Both *pose risk* and *put at risk* have overtaken *run risk* in frequency. Use of the prototypical risk process, *to risk* is declining. Finally, there is some evidence for reduced agency the *run risk* process.
7. When risk is a process, risked things/potential harms often pertain to individual health. This contrasts with processes as potential harm, which generally relate to people in positions of power.
8. Common risk modifiers (*risky*, *riskier*, *riskiest*) are gradually being displaced by a number of less common constructions (e.g. *low-risk*, *at-risk*, *risk-averse*, *risk-free*)
9. While risk as a modifier is often used in the context of finance/commerce, *at-risk* typically attaches to vulnerable human demographics.
10. Longitudinally, risk words are shifting to less focal parts of clauses. We can approximate these changes using both indices or semantic function information within dependency parses.
11. Proper nouns co-occurring with risk words highlight the close relationship between risk and health discourse.

As will be discussed in Chapter 7, many of these shifts appear to be a part of a broader discourse-semantic trend of implicitness and inarguability of risk.

Chapter 6

Risk in health articles in the NYT

This chapter is currently being prepared for publication, and cannot be included here. See:

Zinn, J. O. & McDonald, D. (2015). *Changing Discourses of Risk and Health-Risk: a corpus analysis of the usage of risk language in The New York Times.* Chamberlain, M. (ed.) 2015/6: Medicine, Discourse, Power and Risk. Routledge

Chapter 7

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*).[19] 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[20], 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 (See Zinn & McDonald, 2015). 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 social situations (i.e *registers*) and different 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 (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*. Consider the following two sets of examples. The first, from 2012, shows examples of the embedded process as negative outcome.

1. Some Democrats are saying the White House set itself up for the charges by making a vow that was bound to be difficult to keep and that would **risk alienating** its business supporters.
2. Some speculated that this partnership **risked alienating** other big retailers, like 7-Eleven, by giving Starbucks influence over how Square 's payment system was developed.
3. And campaigning on behalf of members of Congress could **risk alienating** swing voters, many of whom seem to prefer bipartisan government and dislike one-party rule.

The second contains grammatical subjects modified by *at-risk* (from 2008):

1. He secured nearly $100,000 for a program at the Sephardic Community Center in Brooklyn that seeks to help 'at-risk immigrant youth successfully acculturate' into American society
2. Through the years, he said, more than 1,000 at-risk young people have arrived at his doors.
3. The document signed off on a $1.5 million grant to World Vision, a group that hires only Christians, for salaries of staff members running a program that helps 'at-risk youth' avoid gangs.

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 the following chapter.

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.

2.1. Word class and experiential role of risk words

First, we must consider the trend toward nominal risk words (recall Section 1; Figures 5.2 & 5.3). There is a major functional-semantic difference between risk as a participant and risk as a process, to which this shift congruently corresponds.[21] As a process, when the direct object is the valuable object, *risk* may be nearly synonomous with *jeopardise* (*I risk my life*). When the direct object is the negative outcome, risk as a process means 'may potentially face/incur/suffer', with a negative connotation. As in Filmore and Atkins' frame, the end result of a risk process is either the goal/benefit or the harm/negative outcome. When risk is a participant, however, it generally stands in for only the harm/negative outcome: in a clause like *the risk was outweighed by the benefit*, the negative outcome is realised by the risk word:

1. An array of new techniques, each with its own risks and potential benefits, makes for bewildering options for women.

Risk/reward ratio is another example seen commonly in the NYT data: again, in this case, risk *is* the potential harm:

1. Coughlin did not care to defend his decision or to discuss the risk/reward calculation of leaving a talented and highly paid leader in a game in which athletes sometimes are injured.
2. So the franchise made the ultimate risk-reward play this summer.
3. addressing these concerns does not impede the fiduciary responsibility because in today's marketplace, there are many alternative investments with similar risk-reward profiles.

When risk is a process, however, *benefit* and *reward* do not behave in a similar way. Instead, *benefit* refers to specific institutional schemes:

1. `The Kaziyevs were told they had 10 months to become citizens or risk losing Medicaid benefits.`
2. `Once the marriage ban in New York State is lifted, domestic-partner couples, both gay and straight, will risk losing access to health care and other benefits if their employers treat marriage as the only ticket for entitlement to these benefits, which are increasingly expensive.`

Because of this difference, a shift toward nominal forms is in itself a shift toward a semantic conceptualisation of risk as more interchangable with harm itself. Risk as a process can contain within its adjuncts (or, less likely, within a complement) the goal, or positive outcome. Risk as a participant cannot. When it co-occurs with the goal, the relationship between them is through coordinated conjunction.

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 the 1963 data, the things at risk of loss are 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

2.2. Domains of risk disourse

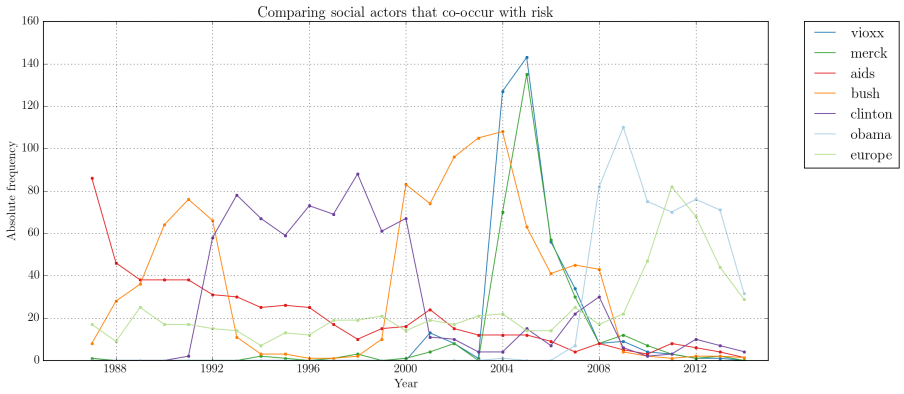


Figure 7.1: Comparing social actors that co-occur with risk

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 (See Figure 7.1). 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.[22] In hundreds of co-occurrences of risk words, Vioxx and Merck, we uncovered a mere handful where the potential harm was to the manufacturer. Though we found one solid example in the 2006 subcorpus (*'The verdict highlights the **risks** that Merck faces as the number of lawsuits over Vioxx continues to grow.'*), this same article containined four other risk words, each of which positions the consumer as being subject to potential negative outcomes:

1. Mr. Escobedo said that Vioxx was especially dangerous to Mr. Garza because of his other **risk** factors and that he should never have been prescribed the drug.
2. 'Mr. Garza was the last person in the world that should have been taking Vioxx,' said Mr. Escobedo, who told the jury that Merck had known since 2000 that the drug posed heart **risks** but continued selling it for four years.
3. About 20 million Americans took Vioxx from 1999 to 2004, when Merck withdrew the drug after a clinical trial showed that it increased the **risk** of heart attacks and strokes compared with a placebo. Earlier clinical trials had also shown that Vioxx appeared to be much riskier to the heart than naproxen, an older painkiller.
4. But recently the tide has seemed to turn abruptly against the company, as its lawyers struggle to explain a raft of documents that show its scientists were concerned about Vioxx's heart **risks** several years before Merck stopped selling the drug in 2004.

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.

2.3. Implicitness of risk

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 further examples of the *at-risk* construction:

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. Some of the proceeds from tickets sales for the event [...] will go to support local arts programs in Washington Heights and the Broadway League's Family First Nights, which the League describes as 'a nationwide program specifically designed to encourage **at-risk families** to attend theater on a regular basis.'
3. Mr. Tepfer noted that Mr. Douglas, who was in the neighborhood when the body was found and was interviewed by the police at the time, 'preyed on **at-risk women**, on prostitutes, and he engaged in sex and strangled them to death.'

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.[23]

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*. Below are examples of *calculated risk* in 1963, contrasted with *potential risk* in 2008.

1. It is, of course, a **calculated risk** that Mr. Kaye is taking.
2. Kennedy has taken a **calculated risk** here.
3. A spokesman for the group acknowledged that granting a 10 per cent discount before a study in depth had been made was a calculated risk.

1. One was to make health care providers and caregivers of infected children aware of the **potential risk** of pre-chewing.
2. At issue were the **potential risks** of having government-run funds in China and other foreign countries make big investments in American businesses.
3. Rat pups exposed to BPA, through injection or food, showed changes in mammary and prostate tissue, suggesting a **potential cancer 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. This aligns with the idea that risk (sociological reference) has come to be simply *threat*.

2.4. Low-risk, moderate-risk, high-risk

1. Hemophiliacs, at **high risk** of AIDS, have been hard hit by the disease.
2. Another 25 percent are **at moderate risk**.
3. But why on this isolated campus, where no AIDS cases have been reported among students at **low risk** of catching the disease, are students so concerned?

During the first years of the U.S. spread of HIV, 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.

2.5. Risk as modifier

1. At JPMorgan Chase, the **risk models** hid--and were used to hide--risks from the traders and top executives.
2. After a rogue trader cost MF Global $141 million, Promontory came in to bolster certain areas of the firm's **risk controls**.
3. He was a total **risk junkie**.
4. The programs are all based on the concept of risk management, rather than the unattainable goal of total **risk elimination**.

Risk occurs within many different modifier positions (see Table 5.10). 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*—more examples above), we can see increased social significance of risk as a concept through the evolution of specific jobs whose central concern is risk (see Section 8 for discussion of the emergence of *risk factor* in particular). Pre-head nominal modification reflects the 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 (*a risky move*, *the riskiest option*) is not a prerequisite for meaning to be understood.

Increasing nominalisation and 'participantification' of risk are also indicative of decreased arguability. The key affordance of nominalisation is that it reduces the need to make meanings through constellations of Participants and Processes: instead,

Using other tests, such as counting the dependency parse indices of risk words, longitudinal change in the arguability of risk words is consistent. In earlier editions, risk words more commonly occupy the more arguable positions of Subject and Finite. In later editions, risk more commonly occurs in heavily dependent, ancilliary positions. Thus, less often does a risk word form the central component being discussed; more and more often, it is used as a modifier of one of these components, or as a part of a supporting, subordinate clause.

Given that Processes in the Transitivity system pattern with the Finite/Predicator in the Mood system, nominalisation facilitates clauses with larger amounts of less arguable information. This discursive function of nominalisation is well-acknowledged both within SFL and outside of it. The increased experiential information density is paid for with 'the interpersonal price of decreasing negotiability' (Halliday & Martin, 1993, p. 41). At the same time, nominalisation 'allows the writer to give the required flavor of objectivity to his or her statements and claims' (Holes 1995, p. 260). Nominalisation disengages the speaker/writer from commitment to the truth of his/her statements by allowing him/her to make 'unattributable claims' (Quirk et al, 1985: p. 1289); it also has the capacity to blur/mystify agency, thus 'masking real intentions' (Hatim, 1997: p. 114).

We are limited, however in our ability to interpret our approximation of arguability through dependency indices. 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[24] 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.[25]

3. People and risk

A particular strength of our approach is that it is possible to understand nuanced distinctions in the ways in which social actors are related to risk. Looking specifically at words pertaining to normal, everyday people (*person/people*, *man/men*, *woman/women*, *childs/children*, etc.), we can see that everyday people are associated more and more with risk discourse: this class of nouns is becoming more frequent, overtaking words related to institutions. When looking specifically at who is the actor in risk processes, however, a different picture emerges: banks, agencies and companies are becoming more frequent, while everyday people become steadily less prominent. This finding paints a rich image of the changing discourse-semantics of risk, whereby risks are created through the actions of powerful people and institutions, but, increasingly, suffered and endured by people. This aligns with neoliberal sociological conceptualisations of the distinction between late and reflexive modernities.

4. Sociological perspectives

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. Spikes in specific terms may correspond with particular events, but general shifts in meaning change place concurrently, and thousands of datapoints overlap and intertwine with the points of interest.

When focussing on the subcorpora of health risks, linguistic reactions to real-world events were 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.

A final point of interest is that adjectival risk words in some respects behaved contrary to our expectations. Simple adjectival risks as modifiers of participants (*the risky manoeuvre*, *riskier choices*) appear to be decreasing in frequency. Furthermore, though there is a very large variety of adjectival risks, and though there is a general trend toward a greater number of risk adjectives, this increase is a slight and gradual one. 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.

From positive risk taking to exposure to risk

- It is difficult to automatically identify positive and negative portrayals of risk through querying or concordancing, as much of this sentiment may be carried in the clauses preceding or following a risk word. That said, we can nonetheless note a number of findings which point toward a reduced agency in participation within risk scenarios.
- exposure to risk is implied in the *at-risk* construction: very rarely are things at-risk due to a conscious choice by everyday people.
- Especially notable is that many constructions that are rising in relative frequency exist without any explicit, or even imaginable, potential goal/positive outcome in the risk scenario. Those at risk of disease are not taking a gamble that has any potential pay-off.
- Money may be risked, but is not put at risk by investors.

Expected risk taking but lacking control

- This hypothesis is partially confirmed by our data. The increasing use of risk language within everyday life world concerns suggests ... through journalism, people are educated about which factors are within and outside of their control.
- Given that factors beyond the control of individiduals play a significant role in their likelihood of developing heart disease or cancer,
- A complicated picture emerges, where individuals are indeed charged with the task of managing their health risks, and yet, the kinds of negative outcomes they are attempting to prevent are in large part caused by factors beyond their control.
- The key semantic distinction between different risk processes is the extent to which the risker is aware of the potential for a negative outcome, or able to choose to participate in the process of risking or not.

The increasing salience of at-risk status in risk reporting

- Though a key way in which increased salience may manifested itself in language is through increasing use of *at-risk* as the head of the Subject or the Finite/Predicator of a clause, this is rarely grammatical in English. Accordingly, our test for increased salience is centred on increasing usage generally.

Individualisation winners and losers—risk-takers versus at-risk groups

-
-
-

5. Summary

-
-
-

Chapter 8

Limitations and future directions

Broadly, our project synthesised *corpus assisted discourse studies* as a **methodology**, *systemic functional linguistics* as a **theory of language**, and *sociological accounts of risk* as **a set of related assumptions about risk**. Our overarching aim was to combine the theories and methods of these areas to provide an empirical account of the ways in the discourse-semantics of risk have undergone longitudinal change, determining in the process which kinds of methods are suitable for further corpus-driven research into risk.

Methodologically, our study has demonstrated the potential for large annotated corpora from a single source to provide empirical evidence for key claims in sociological literature. Though ours is not the first study of discourse using linguistic corpora, it is novel in two key respects. First, by creating a corpus comprised of annual subcorpora, and by interrogating each subcorpus in turn, we can learn not only about how language is used to communicate risk, but also how risk language has changed longitudinally. Second, our study takes advantage of major developments within computational linguistics/natural language processing. Most obvious is the use of constituency and dependency parsing, but also in this category is the development of an IPython interface for manipulating the corpus and visualising the output of queries. These characteristics both allow far more nuanced kinds of investigation than the practices generally seen in CADS (keywording, collocation, concordancing, etc.), and assist in creating reproducible, transparent, open-source humanities/social science research.

1. Limitations of scope

In order to fulfill these aims, we necessarily limited our investigation's scope. The first main issue of scope was our choice of print news data from a single publication. The advantages of this kind of text are many: the NYT is a widely read, well-known, and influential publication. The homogeneity of its language and its consistent structure in many ways facilitate longitudinal and quantitative analysis. The drawback of this kind of data, however, is that we can say little about how risk language works within other kinds of communication. Accordingly, we have made no effort to measure the importance or weight of print journalism against other text types in which risk language occurs, such as film and television, various online media or casual spoken/written conversation.

There were also import constraints imposed both by our chosen theory of language. Though SFL has proven useful as a framework for analysing how language is drawn upon as a resource for making particular kinds of meanings, it is also a theory which has little to say about language and cognition,

for example. This was suitable for our particular investigation, as we cannot possibly determine either the authors' intent behind, nor the readers' interpretations of, the thousands of articles being analysed. Accordingly, we did not attempt to draw links between risk language in the NYT and the ways in which risk is cognitively understood by writers or readers. We suggest that the various strengths of different functional accounts of language can work in tandem, however, and thus welcome future insights from cognitive approaches to risk.

A third and final constraint is our selection of linguistic phenomena for detailed analysis. Primarily, we focus on the experiential and interpersonal dimensions of risk language. The third key component of language is its textual dimension: how language is reflexively organised into meaningful, coherent units. Though our decisions here were guided by the fact that risk as a word does not tend to play important roles in building cohesion and coherence in narratives, we readily admit that more detailed analysis of the role of risk within this dimension may yield important insights that we have not uncovered. Tracking whether risk words shift longitudinally within the textual dimension (between given and new information within a clause, for example) may also be able to show us the extent to which people are aquainted with the notion of risk itself.

2. Shortcomings in natural language processing tools

A major issue in our study relates to the performance and epistemological consequences of digital tools used during the investigation. In short, available digital tools may not perform as desired. Parsers remain far from perfect, and innumerable mistakes in parsing are present in our dataset. What was missing in our results 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, and millions of features extracted and counted. Accordinglyt, oversights and mistakes are unfortunately bound to remain.

A related issue is that the parser used here—*Stanford CoreNLP*—relies on phrase structure and dependency grammars, rather than systemic functional grammar (for which fewer computational resources are presently available). We were thus left with the task of translating systemic-functional concepts into phrase structure grammar and dependency grammar. This process was often time-consuming and counter-intuitive, laden with currently unsolvable ambiguities, as well as theoretically difficult to reconcile.

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 basic treatment of health risks given here.

3. The limits of lexicogrammatical querying

A major issue we faced during our investigation, and have yet to deal with directly, is the potential for similar discourse-semantic meanings to be made via a number of different kinds of lexicogrammatical arrangements. Consider the following invented examples:

1. They risked their money
2. Risked money was lost
3. They risked their savings
4. The risk of money loss was there
5. She took her money from her purse and risked it.
6. The money, which they risked, was lost.
7. They had money. They risked it.

Each of these examples communicates the same kind of semantic meaning—that money was risked—but through different grammatical strategies, ranging from the group level (Ex. 1) to the clause-complex (Ex. 7). Our analyses typically dealt with the most common, or *congruent*, kinds of realisations, but at the expense of meanings made incongruently, or above the level of the clause. With great difficulty, we could perhaps construct a query that matches every one of these results, or merge the results of a number of separate queries. As the queries grow in complexity, however, undesirable results may creep in: a query matching *money* in the above cases would also likely match *death* in *They risked death*, despite the fact that one is the risked object and one is the potential harm. Determining the proper functional role in the cases above is very simple for human coders, but the number of results in need of categorisation is often far from trivial. Limited by both the ability of current parsers and by constraints of scope, we found ourselves largely unable to devise methods for accounting for incongruence in risk language during automated querying. As a result, our analysis was restricted for the most part to meanings that were being made in the most probable, normative ways.

The second major issue is the exact converse scenatio: counted together in many our automatic queries are many examples with contradictory semantic meanings. Continuing our example of money loss, consider the following:

1. They would have risked their money
2. They didn't risk their money
3. Risking money was a terrible idea, so they didn't do it.
4. Don't risk their money.

In each of these cases, money was not necessarily lost. Lexicogrammatical querying, however, would simply count *money* as the *risked/lost thing*. Though we were careful not to conflate our abstracted results with occurrences of particular events (money loss), we did not attempt to determine whether certain things were more often either hypothetically or really risked. Our approach is not unique in its lack of engagement with incongruent meanings, hypotheticals, and the like: few corpus-based studies of discourse have attempted to distinguish between these meanings automatically. Indeed, querying of these kinds of features could uncover disparities between real and imagined sites of risk in news discourse: do politicians risk more than everyday people, or do they only *potentially risk* more? Given the extensive work on meanings made by mood and modality systems in English, as well as the accuracy of automatic

annotation for these kinds of lexicogrammatical features, we suggest that more constrained studies of mood features of risk could plausibly be undetaken using tools and methods developed here.

Finally, it must also be noted that any study of text corpora necessarily involves removing text from the actuality in which it was produced. Though we can be attuned to the nature of written news journalism, we have not been able to account for meanings made multimodally (through adjacent images, advertisements, etc.). Though perhaps not a critical issue in studies of print news corpora (in comparison to television advertisements, casual conversation, etc.), it is nontheless important to acknowledge that in some sense we have been studying *text*, rather than *texts*. Synthesis of corpus findings with in-depth analyses of individual articles, or of the influence of the media production process, would no doubt improve our ability to generalise our results. Indeed, future research incorporating these perspectives is planned.

4. Research agenda

The tools and methods here could easily be operationalised with new datasets, allowing for research into changes in risk semantics at differnet points in time, or into regional/multilingual differences. Of course, further studies need not be limited to *risk*: though we limited our analysis to risk words and their co-text, resources such as *The New York Times Annotated Corpus* (Sandhaus, 2008)) present the possibility of analysis of any other lexical items and/or grammatical structures. Sociological hypotheses concerning the relationship between risk and close synonyms, such as danger or threat, or between risk and key events, such as terrorism, could be studied with a great deal of specificity with a larger corpus. Such a resource could also serve as a more authoritative reference corpus.

Theoretically, there are far more avenues worthy of exploration: key sociological claims need to be addressed in detail, with sustained description and analysis. As just one example, though the increasing synonymy of risk and threat cannot be accurately measured in a corpus constructed of risk words and their co-text, examination of this link would be readily observable with a different corpus. A number of metrics exist for determining the extent to which two words are interchangeable (functionality provided by both NLTK and Sketch Engine)—this could be another fruitful method for uncovering shifting discourse-semantics of risk.

5. Conclusions

We have both provided empirical support for a number of major sociological hypotheses concerning risk, and demonstrated that novel combinations of interdisciplinary methods can offer new ideas and new kinds of evidence about risk. A number of key findings aligned with theories of neoliberalism and reflexive modernity, but at the same time, key claims within this body of literature were problematised: while we saw a dramatic increase in the extent to which science and research terms co-occur with risk, we saw little evidence that the science and research were being actively contested. Whether this reflects increasing trust and reference for scientific method, or whether it simply reflects changing journalism practices, requires further research. Indeed, we expect that contestation of science is more likely to occur in everyday discourse of ordinary citizens than in news journalism, for which science writing is a cost-effective means of generating authoritative news content.

Though our study has perhaps raised more questions than it has answered, it has also demonstrated the feasibility of quantitatively and empirically observing key sites of sociological change.

Notes

1. This was operationalised through the case-insensitive regular expression **\brisk.*?\b**, where **\b** acts as a word-boundary marker.
2. Efforts are underway to digitise all risk words in 1963 additions, as well as editions from years between 1963 and 1987.
3. 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* are currently being turned into an identically structured, though dramatically larger, cloud-hosted corpus. In planned future research, interrogation of this corpus will be 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.
4. 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 3.1 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.
5. 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.
6. 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
7. 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.
8. The corpus contained too few modalised risk predicators for analysis of longitudinal change in modalisation.
9. 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.
10. 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.
11. 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
12. Modification through embedded clauses (*the children who were at risk*) has been left out for reasons of scope.
13. 1963 is excluded from analysis here, as poor quality OCR created a number of non-word results such as *risks-wnrk*, *risks.North* and *risks.With*.
14. These are not particularly meaningful distinctions in SFL, as the role of participants depends on the process type. Though process type identification is becoming more and more feasible (see OâĂŹDonnell, 2008), at the time of our investigation, we lacked resources to automatically distinguish between process types, and thus relied on borrowed notions of experiential subject- and objecthood.
15. *Take* and *run* are removed from the object column here, as *take risk* and *run risk* are considered risk processes.
16. Further research devoting more time to which examples constitute risk processes and which constitute experiential objects of other processes seems timely.
17. A limitation of frame semantics is its ability to deal with other kinds of risk processes: in *The site posed risks*, *pose* is the experiential subject, but not the actor or risker.

18. Though the filtering out of titles and given names collapses the distinction between Bushes and Clintons, we can still reasonably infer which was being spoken about at which, and doubt can be eliminated by concordancing

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

20. 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.

21. Though SFL practioners will remind us that nominal does not equal participant, this does not change the importance of the congruent realisation, especially when using very large datasets.

22. Our forthcoming chapter presents this discussion in more depth.)

23. Indeed, this is aligned with recent changes to the frame semantic conceptualisation of risk. At the time of writing the FrameNet entry for *run risk* included the following caveat: '*NOTE: This Frame is currently in the process of being changed so that some instances of at risk.n will be moved to the Being_at_risk frame, and some will be moved to the Risky_situation frame. In the Being_at_risk frame, risk is almost always supported with at, and its external argument is the Asset*' (see Baker, Fillmore, & Lowe, 1998).

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

25. Newer releases of `corpkit` are oriented more toward dependency than constituency parses.

References

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1* (pp. 86–90). Association for Computational Linguistics. Retrieved 2014-02-04, from `http://dl.acm.org/citation.cfm?id=980860`

Baker, P. (2004). Querying Keywords Questions of Difference, Frequency, and Sense in Keywords Analysis. *Journal of English Linguistics*, *32*(4), 346–359. (00071)

Baker, P., Gabrielatos, C., Khosravinik, M., KrzyÅijanowski, 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.

Beck, U. (1992). *Risk society: Towards a new modernity*. Sage.

Beck, U. (2009). *World at risk*. Polity.

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. (Cited by 0034)

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.

Dean, M. (2010). *Governmentality: Power and rule in modern society*. SAGE Publications, Inc.

Douglas, M. (1986). *Risk acceptability according to the social sciences*. Russell Sage Foundation.

Douglas, M. (2013). *Risk and blame*. Routledge.

Duguid, A. (2010, November). 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.

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. Retrieved from `http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-17.pdf` (Cited by 0001)

Giddens, A. (2002). *Runaway world: How globalisation is reshaping our lives*. Profile books.

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, March). 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. (00000) doi: 10.1515/text-2013-0028

JOHNSON, S., & SUHR, S. (2003, January). From 'Political Correctness' to 'Politische Korrektheit': Discourses of 'PC' in the German Newspaper, Die Welt. *Discourse & Society*, *14*(1), 49–68. Retrieved from http://das.sagepub.com/content/14/1/49.abstract doi: 10.1177/0957926503014001929

Luhmann, N. (1993). *Risk: A Sociological Theory.* New York: Walter de Gruyter.

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). Retrieved from http://www.aclweb.org/anthology/P/P14/P14-5010

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. Retrieved from http://books.google.com.au/books?hl=en&lr=&id=_cA9AAAAQBAJ&oi=fnd&pg=PA235&dq=%22cline+of+arguability%22&ots=EmVZ29HG-U&sig=O6AQs9awRnhTO1RSblkkLNq2Ac8&redir_esc=y#v=onepage&q=%22cline%20of%20arguability%22&f=false

Matthiessen, C. M. (1995). *Lexicogrammatical cartography: English systems.* International Language Science.

Matthiessen, C. M. (2013). Applying systemic functional linguistics in healthcare contexts. *Text & Talk*, *33*(4-5), 437–466. (00000)

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). *MALLET: A Machine Learning for Language Toolkit.* Retrieved from http://mallet.cs.umass.edu

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

O'Malley, P. (2012). *Risk, uncertainty and government.* Routledge.

OâĂŹDonnell, M. (2008). The UAM CorpusTool: Software for corpus annotation and exploration. In *Proceedings of the XXVI Congreso de AESLA.* (00026)

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). *SPINDLE Automatic Keyword Generation: Step by Step.* Retrieved from http://blogs.it.ox.ac.uk/openspires/2012/09/12/spindle-automatic-keyword-generation-step-by-step/

Rose, N. (1999). *Powers of freedom: Reframing political thought.* Cambridge university press.

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.

Skolbekken, J.-A. (1995). The risk epidemic in medical journals. *Social Science & Medicine*, *40*(3), 291–305.

Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied linguistics*, *21*(1), 3–25.

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