Chapter 9
Visual Analytic Observatory of Scientific Knowledge

Abstract A conceptualization of research on uncertainties in scientific knowledge is presented. Several common sources of uncertainties in scientific literature are characterized, notably, retracted scientific publications, hedging, and conflicting findings. Semantically equivalent uncertainty cue words and their connections with semantic predications are identified and visualized as the first step towards a systematic study of uncertainties in accessing and communicating the status of scientific assertions.

Introduction

As new discoveries and advances are made, scientific knowledge, conveyed through the content of scientific literature, is subject to constant changes. These changes could be revolutionary as well as evolutionary (e.g., Kuhn 1962; Fuchs 1993; Shneider 2009). Despite the tremendous growth in terms of scholarly metrics to measure various aspects of scientific activities and the growing efficiency in retrieving relevant scientific publications in general, accessing scientific knowledge to meet our needs for assessing the state of the art of a research area and making various decisions remains a major challenge (Chen 2016).

Today, we still have to build our understanding of the state of the art of science through painstakingly time-consuming and cognitively demanding processes. We still have to piece together sporadically distributed information and transform it to a cohesive conceptualization of our own. The knowledge acquisition process from the vast volume of scientific literature remains the most challenging bottleneck not only for scientists and researchers, but for everyone seeking to obtain an accurate picture of the state of the art. Although increasingly sophisticated techniques

C. Chen and M. Song, *Representing Scientific Knowledge*,
https://doi.org/10.1007/978-3-319-62543-0_9

emerge to address one or more specific aspects of the knowledge acquisition bottleneck, the scientific community as a whole is still limited by the lack of integrative and widely accessible options to increase the throughput of the bottleneck and in turn to increase the efficiency and effectiveness of the transformation from information to knowledge. Furthermore, the development and evaluation of such tools is hindered by the lack of accessible and persistently maintained resources such as classic cases and training materials of in-depth studies of representative high-impact research, contemporary and innovative metrics and analytic tools, metadata and gold standards for comparative and evaluative studies.

Visual Analytic Observatory of Scientific Knowledge

We envisage a widely accessible and persistently maintained community resource —a *visual analytic observatory of scientific knowledge* (VAO). The central idea of the VAO is that the essence of scientific knowledge can be captured by a set of semantically organized assertions along with their status of uncertainty and that knowledge represented in this way can fundamentally increase the efficiency and accuracy of our understanding of scientific knowledge. As a result, many existing analytic methods will be fruitfully extended to the new level of granularity. A sketch of the architecture is illustrated in Fig. 9.1. The development of the VAO[1] is supported by the Science of Science and Innovation Policy (SciSIP) program of the National Science Foundation.

In this ambitious framework, unstructured text in a scholarly publication will be transformed to a semantic network of assertions along with their epistemological status and the provenance of their evolution. A set of scientific articles will be represented by a more extensive but organizationally equivalent semantic network. Ultimately, the body of scientific literature of a scientific domain can be represented in this framework. This framework will eventually enable us to transform how we communicate and keep abreast with the advances of science.

A unique focus in the VAO development is the role of uncertainty in advances of scientific knowledge. The goal of the VAO is to improve the clarity of the representation of scientific knowledge substantially and especially improve the clarity of the uncertainties associated with particular areas of scientific knowledge. Ultimately, the VAO will make scientific knowledge easy to access with the level of clarity that one can communicate efficiently to address Heilmeier's series of questions regarding the planning, execution, or evaluation of scientific inquiries.

[1] https://www.researchgate.net/project/Research-A-Visual-Analytic-Observatory-of-Scientific-Knowledge-VAO.

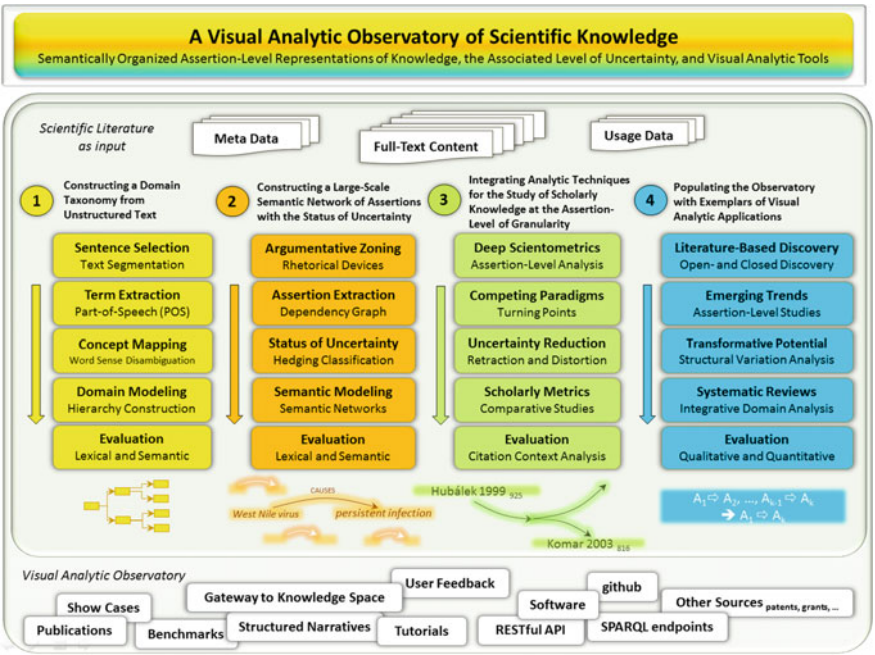

Fig. 9.1 Architecture of a visual analytic observatory of scientific knowledge

Many efforts in representing scientific knowledge attempt to reduce scientific knowledge to a set of propositions. For instance, in Semantic MEDLINE, scientific assertions are extracted from unstructured text of published articles and expressed as propositions in a generic form of (*Subject*)-(*Predicate*)-(*Object*) as in (*West Nile Virus*)-(*Causes*)-(*Persistent Infection*) or (HIV)-(Causes)-(AIDS). The negation of an assertion is also an assertion. An assertion itself can be embedded as the subject or the object of another assertion, e.g. (Assertion_1)-(is a)-(Assertion_2). On the one hand, scientific knowledge is represented by propositions that have been aggregated and mapped to standard vocabularies such as UMLS concepts and semantic types. The complexity of the diverse expressions in natural languages is considerably reduced and it is easier to handle the represented knowledge computationally. On the other hand, much of the meta-knowledge is lost, notably the epistemic status of a scientific assertion in terms of how scientists try to communicate the subtlety as precisely as possible through carefully chosen words, and, perhaps more importantly, how scientists handle conflicting and contradictory reports in the scientific literature on the exact same topic. We believe that scientific knowledge should be represented and communicated along with its epistemic status and the provenance of its status. Such meta-knowledge is an integral part of an expert's domain expertise. The knowledge of uncertainty is an expert's expertise!

The study of uncertainties in scientific knowledge should pay particular attention to two sources of uncertainty: (1) hedging and (2) contradictions. If a scientific claim is modified by hedging devices, then it indicates that the researcher who is making the assertion evidently has reservations to the truth of the claim. For instance, a statement that HIV causes AIDS leaves no doubt to its audience about the firm belief of its author. In contrast, one may become doubtful when reading a more carefully crafted statement: *a recently published study suggest that X might be responsible for Y if condition Z is met*. Hedging may become necessary when information is incomplete or entirely missing.

Intuitively, the level of uncertainty is higher when it is evident that contradictory information prevents scientists from making a positive and absolute assertion. Conflicting, contradictory, and controversial results must be reconciled before speculations and hypotheses can be accepted as part of scientific knowledge. How often do we come across topics or research areas that are puzzled by conflicting information? How important is it for scientists to reconcile contradictory findings?

The VAO aims to provide an integrative, extensible, and shared platform for the study of scientific knowledge and for the research and development of new tools. As a community resource, the VAO will enable scientists, analysts, and the general public to accomplish several types of analytic tasks that have been so far cognitively demanding and time consuming. It will enable the study of scientific knowledge to reach a deeper level of granularity and, more importantly, a potentially more efficient and effective way to understand critical information in scientific discovery and in the public understanding of science. It has the potential to increase the productivity of research at a reduced cost.

Types of Uncertainties in Scientific Literature

Scientific knowledge is never free of uncertainty. It is difficult to communicate uncertainty clearly, especially on issues with widespread concerns, such as climate change (Heffernan 2007) and Ebola (Johnson and Slovic 2015). The way in which the uncertainty of scientific knowledge is communicated to the public can influence the perceived level of risk and the trust (Johnson and Slovic 2015). A good understanding of the underlying landscape of uncertainty is essential, especially in areas where information is incomplete, contradictory, or completely missing. For instance, there is no information on how long the Ebola virus can survive in a water environment (Bibby et al. 2015). If surrogates with similar physiological characteristics can be found, then any knowledge of such surrogates would be valuable. Currently, finding such surrogates in the literature presents a real challenge (Bibby et al. 2015).

According to sociological views of scientific change, competition leads to scientific change (Fuchs 1993). Three types of scientific change are likely to emerge: permanent discovery, specialization, and fragmentation. The severity of competition is the strongest in settings that lead to permanent discovery. A lighter degree of

competition is associated with specialization. The least competitive environment is associated with fragmentation.

Scientists compete for recognition and reputation. Many other tangible or intangible benefits may come with established reputation and authority. Publishing novel and interesting discoveries is one of the long established traditions in science. The threshold of publishing a scientific article has been lowered over the years. New journals are launched at a high speed.

From the competition point of view, novel, interesting, and controversial ideas are likely to attract more attention than commonly known, trivial, and expected results. Sociologists suggest that the interestingness of a topic depends on whether it challenges our current beliefs. If we know the information that we are about to learn is contradict to our current belief, then we can expect that the gain from understanding the new information is likely to be the highest.

Table 9.1 presents some examples of sentences from MEDLINE articles. These sentences indicate some common types of uncertainties in biomedicine. The first column is a list of terms that indicate some types of uncertainty—we call them cue words of uncertainty, for example, the term unknown in the first sentence "The mechanism is unknown." The uncertainty is high when the mechanism of a disease is unknown. Contradictions are another type of uncertainty. One must validate each of the contradicting components before making selections. Similarly, controversial and inconsistent results in published articles all represent a degree of uncertainty. In summary, if there are competing alternative interpretations, then we are dealing with uncertainty.

Hedging and Speculative Cues

Hedging is a particularly relevant concept for characterizing the tentative and context-dependent nature of scientific claims (Hyland 1996). Hedging is a rhetorical means, or a communicative technique, to convey the degree of uncertainty associated with a statement or an assertion (Behnam et al. 2012; Clark et al. 2011; Di Marco et al. 2006; Horn 2001; Kilicoglu and Bergler 2008). The presence of hedge words can mitigate an otherwise overstated scientific claim such that the status of the knowledge is documented more accurately. Reinstating hedging information surrounding an assertion can help us to understand precisely what is currently known about the assertion. Introducing hedging information provides an additional and important means to characterize the role of an assertion in the context of the domain knowledge as a whole. Furthermore, it will enable us to understand not only the current status of a scientific assertion, but also the trajectory of the evolution of its status over time. We will be able to better understand how the uncertainty associated with a scientific assertion changes as new information, e.g. new discoveries, becomes available. We will be able to better assess the potential of a research program in terms of the extent to which it reduces the uncertainty of the scientific knowledge of a particular area.

Table 9.1 Sentences that indicate uncertainties in scientific knowledge

Terms on uncertainty	Type	Instances	PMID	Sentence ID	Sentence
Unknown	ab	300800	165704	10667452	The mechanism is unknown
Suspect	ab	165545	12351994	77704397	An immunopathology is suspected
Unclear	ab	164034	7260869	10419608	The etiology is unclear
Unusual	ab	141237	3629081	33402065	Such cases are unusual
Controversial	ab	122406	2499131	34124598	The results are controversial
Consensus	ab	113464	23979725	152414767	There is no consensus on treatment
Incomplete	ab	95914	2419361	29557765	This association is incomplete
Conflicting	ab	91371	11433428	68757263	These reports are conflicting
Contrary	ab	68059	8324612	45236707	On the contrary it is increasing
Debatable	ab	64233	860951	13010435	Possible causes are being debated
Inconsistent	ab	53353	10434263	63377317	The results are inconsistent
Uncertain	ab	48831	3585876	3573539	The etiology is uncertain
Unexpected	ab	46336	2260033	53665387	This result is unexpected
Confusing	ab	39363	2250070	44822246	This was confusing and misleading
Paradoxical	ab	38218	7635297	51365510	This leads to a paradox

Uncertainty cues in scientific writing in general come from adjectives, adverbs, auxiliaries, verbs, conjunctions, and nouns. Szarvas et al. (2012) identified uncertainty cues in each of these categories. For instance, probably, likely, and possible are uncertainty cues in the adjective and adverb category. Examples of auxiliaries as uncertain cues include may, might, and could. Speculative verbs include suggest, seem, and appear. Nouns include speculation, proposal, and rumor.

Researchers have developed heuristics that can be used to detect propositions with uncertainty based on uncertainty cues. For example, based on the suggestions of Kilicoglu and Bergler (2008), one can derive the following heuristics to identify propositions that are likely to involve uncertainties:

- If a proposition has an uncertain verb, noun, preposition, or auxiliary as a parent in the dependency graph of the sentence, then the event is regarded as uncertain.
- If a proposition has an uncertain adverb or adjective as its child, then it is regarded as uncertain.

Figure 9.2 shows a log-log plot of hedging words appeared in the conflict (x) and non-conflict (y) axes. Words appear below the dashed line appeared more often in the conflicting set than in the non-conflicting set. The conflict set consists of MEDLINE articles with sentences containing words such as conflict, contradictory, and inconsistent, whereas the non-conflict set does not contain such cue words.

Uncertainty cue detection has mostly been developed in the biomedicine domains (Szarvas et al. 2012). Researchers studied the distribution of hedging cues in scientific writings of different domains. Rizomilioti (2006) studied publications from three domains, namely, archeology, literacy criticism, and biology and found that uncertainty cues were the highest in archeology and the fewest in literacy criticism. Hyland (1998) found that writers in humanities use hedging devices significantly more than writers in sciences. Falahati (2006) compared psychology, medicine, and chemistry and found that hedges are more often in psychology than in medicine and chemistry.

Table 9.2 shows the uncertainty levels of scientific disciplines in Elsevier's full text repository Consyn as of August 13, 2015. The uncertainty level of each scientific discipline, or subject area, in Consyn is estimated by the proportion of items containing any of the five words: *conflicting*, *contradictory*, *inconsistent*, *discrepant*, and *irreconcilable*. These five words are useful indicators of controversial and unresolved alternative interpretations. They indicate the lack of clarity of the status of scientific knowledge.

The top four disciplines with the highest rates of uncertainty items are all social sciences. Mathematics, physics and astronomy, chemical engineering, along with material science and chemistry are the five disciplines with the lowest level of uncertainty word use. Psychology has 32%—the highest rate of items characterized

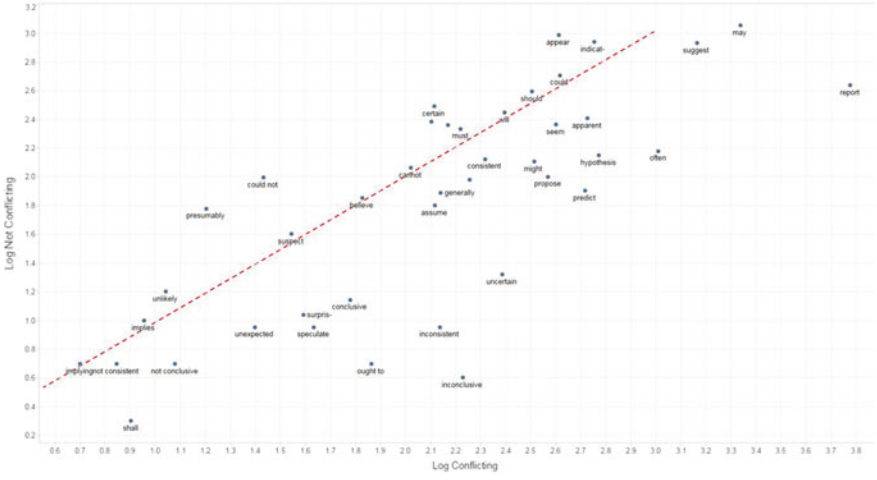

Fig. 9.2 Hedging words in the conflicting set versus the non-conflicting set

Table 9.2 The uncertainties of scientific disciplines

1	Subject area (as of 8/13/2015)	Journal items only	Subtotal items in area	Rate (%)
2	Psychology	70,096	220,250	32
3	Business, management and accounting	26,717	97,083	28
4	Social sciences	74,835	283,598	26
5	Economics, econometrics and finance	27,920	113,083	25
6	Neuroscience	99,908	434,270	23
7	Medicine and dentistry	423,391	2,093,102	20
8	Veterinary science and veterinary medicine	24,390	126,768	19
9	Pharmacology, toxicology and pharmaceutical science	56,441	305,601	18
10	Nursing and heal professionals	39,692	218,124	18
11	Arts and humanities	14,470	78,844	18
12	Environmental sciences	56,594	328,192	17
13	Immunology and microbiology	51,184	310,404	16
14	Agricultural and biological sciences	63,010	400,272	16
15	Biochemistry, genetics and molecular biology	120,012	800,766	15
16	Computer science	32,040	252,366	13
17	Decision sciences	17,500	144,119	12
18	Earth and planetary sciences	24,393	225,816	11
19	Engineering	45,281	510,624	9
20	Energy	18,253	235,489	8
21	Mathematics	17,737	239,676	7
22	Physics and astronomy	28,507	498,418	6
23	Chemical engineering	17,434	355,512	5
24	Material science	24,038	608,991	4
25	Chemistry	20,585	52,2442	4

Source Consyn

by the five words of uncertainty. In contrast, material science and chemistry have 4%—the lowest.

It has been estimated that 11% of sentences in MEDLINE abstracts contain speculative terms (Light et al. 2004). The purpose of hedge classification is to determine whether a sentence is speculative or factual (Medlock and Briscoe 2007). Machine learning techniques such as Support Vector Machines (SVMs) have been used to classify sentences into speculative or non-speculative groups (Light et al. 2004).

HypothesisFinder is a good example of detecting speculative statements in the domain of Alzheimer's diseases (AD) (Malhotra et al. 2013). HypothesisFinder uses a dictionary of speculative patterns. Their study identified three groups of speculative patterns and their ability to detect speculative sentences accurately. For

example, the strongest signals are given by phrases such as "*might be involved*," "*hypothesized that*," and "*raising the possibility that*." The medium-strength signals include "*seems to*," "*appears to be*," and "*can be anticipated*." Weak patterns include "*presume*," "*suppose*," and "*would*." HypothesisFinder is available online[2] as part of the information retrieval system SCAIView Academia. A precision of 0.91 and a recall of 0.73 were reported for their evaluation based on the BioScope corpus (Szarvas et al. 2008).

Finding Semantically Equivalent Uncertainty Cues

We are developing a new method for uncertainty cue word recognition (Chen et al. 2017). Unlike earlier studies that commonly used hand-crafted rules and dependency graphs to identify cues of uncertainty, we found that recent advances in deep learning and distributional semantics have the potential to make substantial improvements (McDonald, and Ramscar 2001).

The distributional hypothesis is that words appearing in the same contexts tend to have similar meanings (Harris 1954). They are likely semantically equivalent. Word2vec (Mikolov et al 2013) is one of the most popular word embedding models in the recent years (see Chap. 6). Using a Word2vec model training on Google news, we expanded a list of hand-picked uncertainty cue words to obtain many more semantically equivalent uncertainty cue words.

The seed list is shown in Table 9.3. The selection of the initial uncertainty cue words was based on our own heuristics of how an uncertain can be directly characterized or indirectly inferred. For example, words in the original seed list include words such as inconsistent, ambiguous, debatable, bizarre, and surprising. When these words are found in a scientific publication, one can expect that the statement implies some degree of uncertainty. For example, inconsistent results may imply that a research question involves uncertainties because researchers cannot settle it yet and extra efforts are required to clarify the current inconsistency. Similarly, if a study has produced surprising results, then the underlying theory is questionable because it was not capable of predicating the results correctly.

The word2vec expansion increased the number of semantically equivalent uncertainty cue words by almost 10 times with a total of 469 words combined. The expanded words represent 83.37% of the combined set. The original seed list represents 16.63 of the combined set. Figure 9.3 visualizes the combined set of uncertainty cue words. Words from the original seed list are shown in red, including prominent words such as inconsistent, contrary, ambiguous, bizarre, and debatable.

[2] http://www.scaiview.com/scaiview-academia.html.

Table 9.3 A seed list of uncertainty cue words

Ambiguity or -ous	Irreconcilable
Baffling	Misbelief
Bizarre	Misconception
Conflicting	Misleading
Confusing	Mystery or -ious or -ies
Consensus	Paradox or -ical
Contentious	Perplexity
Contradictory	Puzzling
Contrary	Skeptic
Controversial	Surprising or surprise
Debatable	Suspect
Deceptive	Suspicion
Dispute	Unanticipated
Doubtful	Uncertain
Dubious	Uncertainty
Fallacy	Uncharted
Flaw	Unclear
Implausible	Unconvincing
Impossible	Undetermined
Improbable	Undiscovered
Incoherent	Unexpected
Incompatible	Unexplained
Incomplete	Unidentified
Incomprehensible	Unknown
Inconceivable	Unpredictable
Inconclusive	Unrecognized
Incongruity	Unreliable
Inconsistent	Unusual

In contrast, words expanded from the word2vec model are shown in green, including words such as misguided, inaccurate, tricky, muddled, and contradictive.

Figure 9.4 shows the network of 469 uncertainty cue words colored in 11 communities, i.e. semantically equivalent classes. The size of a label is proportional to the eigenvector centrality of the corresponding node in the network. For instance, inconsistent has the highest eigenvector centrality, followed by contrary and ambiguous, all of which belong to the same class.

Uncertainty cue words can be used to select sentences that may involve a degree of uncertainty. Furthermore, uncertainties surrounding semantic predications can be identified.

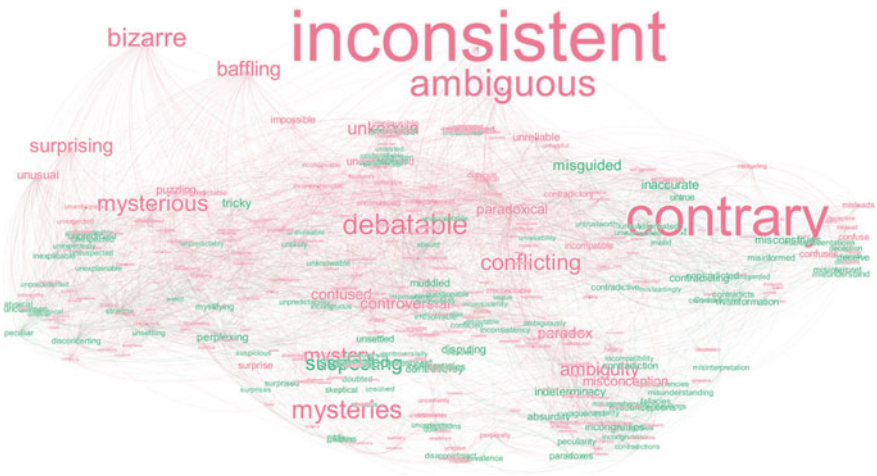

Fig. 9.3 Uncertainty cue words from the original seed list (in red) and expanded (in green)

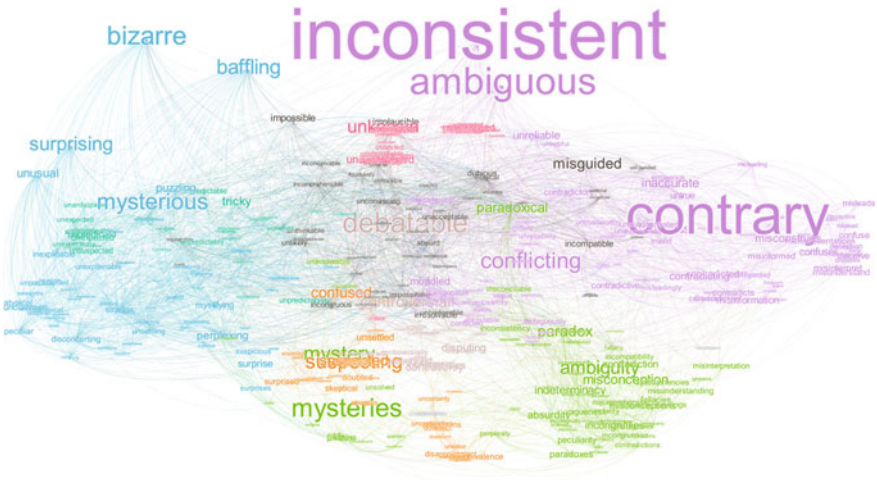

Fig. 9.4 Uncertainty cue words colored in 11 semantically equivalent classes. The size of a label is proportional to the eigenvector centrality of the corresponding word

Citation Distortion and Provenance of Evidence

Citation distortion is a significant source of uncertainty. The scholarly impact of a scientific discovery is often considered in terms of how it is cited in subsequent publications. The strength of the influence of a single published statement on subsequent interpretation reveals interferences in chains of collective reasoning (Rzhetsky et al. 2006). However, as authors of citing articles commonly paraphrase

the original work in citation contexts, citations may distort the intended interpretation of the original source. Such distortions deviate the true epistemic status of the original finding.

Greenberg (2009) demonstrated how citations are overwhelmingly biased towards citing supportive as opposed to refuting papers of a specific claim and the important role of review papers in directing the flow of citations concerning a scientific claim. His study shows that primary data that weakened or refuted claims were ignored and citations exponentially amplified supportive claims over time. Greenberg's analysis also found evidence of how the status of a scientific hypothesis may be distorted in subsequent citations and a hypothesis was incorrectly referred to as a matter of fact—"*This subclaim had transformed from hypothesis to "fact" through citation alone, a process that might be called citation transmutation*" on page 5 of (Greenberg 2009).

Greenberg found that citation biases, amplifications, and citation diversions appeared not only in scientific literature, but also in nine grant proposals funded by the NIH. His investigation raises an important question that the science of science and innovation needs to address concerning the trustworthiness of how scientific knowledge is stated, paraphrased, and quoted.

Citation contexts of a published article refer to the sentence containing an instance of citation along with surrounding sentences. Citation contexts and hedging are connected in an interesting way. Research shows that hedging is more frequently seen in citation contexts than other sentences of a scientific article (Di Marco et al 2006). Semantic predications extracted from citation contexts of an article will provide additional insights into the semantic relations extracted from the original statements. These semantic predications and corresponding information of their uncertainties form a chain of evidence of how the original work impacts subsequently published studies. Taken together, the provenance of evidence is valuable for us to develop a good understanding of scientific knowledge and its dynamics.

Retraction

If hedging and citation distortions indirectly indicate the possibility of uncertainties involved in scientific knowledge, the retraction of a published article sends direct signals that some claimed scientific knowledge must be re-examined and re-validated (Chen et al. 2013). In such situations, the uncertainty should increase and scientific knowledge as a collective belief system should be rolled back to the point prior to the publication of the retracted article. Notorious examples of retracted studies include the highly controversial study on the connection between MMR vaccines and autism by Wakefield et al. (1998), the Bell Lab physicist's

forging data (see Service RF 2002), the high-profile retraction of Hwang (Kakuk 2009), and the rise and fall of STAP.[3]

We may all have heard a variant version of the same story. A hen lays a golden egg every day. However, that is not good enough for its owner, who would rather to have all the golden eggs all at once. So the owner killed the hen to retrieve the golden eggs, but to his surprise, he ended up with no golden eggs not only for that day but forever. To a scientist, a high-profile breakthrough would be a golden egg. Under the intensive competition, the more golden eggs he could produce, the better. Unlike the hen that can produce a golden egg every day, a researcher may not guarantee when he can deliver a golden egg. In fact, no one can plan for the delivery of a golden egg in his entire scholarly career.

Imagine two scientists are competing for recognition in a high-profile area of research. The one who makes a breakthrough first is likely to receive all the attention and all the resources. In contrast, his competitor is likely to suffer a great deal of loss in terms of attention and resources. The two scientists not only have to publish, but also maximize the chance that what they publish will attract the attention of the field.

A retraction is a step that can undo the process of a publication. Retractions are most common in areas that are advancing very fast. Publications in such areas have a relatively low half-time expectation. Chen et al. (2013) found that the most active and fast-moving areas of research have a higher rate of retraced articles. This is the type of scientific change that is resulted from the highest degree of competition, namely permanent discoveries. As Fuchs explained, scientists with high mutual dependence in the research fronts and working on research with a high degree of uncertainty have the highest stake.

Severe competition and pressure is not an excuse for compromising the integrity of one's scholarship. It is, however, something that one can anticipate as a result of the interplay of a broad spectrum of social, psychological, and behavioral factors.

The retraction of a published article is a mechanism for restoring compromised scientific knowledge. Figure 9.5 shows numerous highly cited retracted articles. The fact that a retracted article has been highly cited requires investigations at a deeper level. Why has it been cited? Did authors cite the article before its retraction or afterwards? If they cited after the retraction, are they aware of the fact that is has been retracted? What difference will a retraction make as far as the contemporary scientific knowledge is concerned? Each node labeled in the visualization represents a retracted publication. The size of a node is proportional to its citation counts. The larger size a node, the more citations it has. In Fig. 9.6, the node labeled as Nakao N has a large size. In fact, the retracted article (Nakao et al. 2003) is among the top 10 most cited retracted articles in the Web of Science.

Figure 9.7 illustrates various information of the retracted article by Nakao et al. The title indicates that this article is about a randomized controlled trial of a combination treatment in non-diabetic renal disease. The sentences highlighted in

[3]http://www.nature.com/news/stap-1.15332.

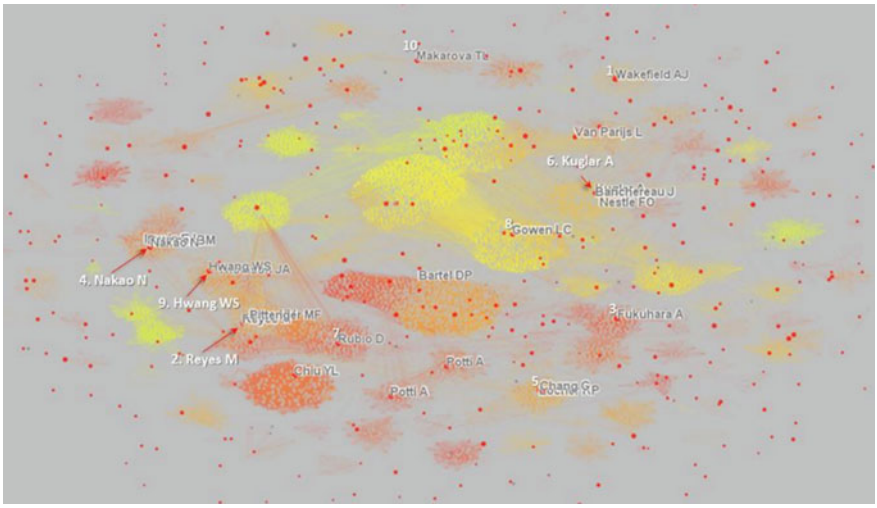

Fig. 9.5 Retracted articles (red dots) in a co-citation network. *Source* Chen et al. (2013)

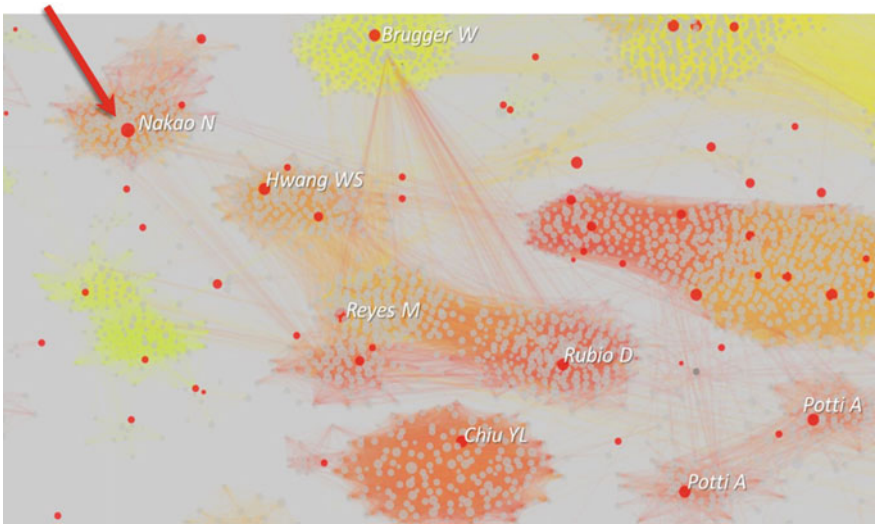

Fig. 9.6 A retracted article by Nakao et al

yellow are sentences from which semantic predications are extracted. For example, one semantic predication is extracted from the title of the article (the first row in Table 9.4). The subject of the predication is Angiotensin-Converting Enzyme Inhibitors. The object is Diabetic Nephropathy. The predicate is TREATS. The diagram shown in the lower right corner of Fig. 9.7 depicts how these predications are connected. The publication of the article imposes this small network of

Fig. 9.7 The first page of the retracted article by Nakao et al. (2003) along with semantic predications extracted from the article

predications over the existing scientific knowledge, which can be represented by semantic predications from all published articles. The retraction removes such a network of predications from the current global network of predications. Thus publications and retractions both alter the structure of scientific knowledge. More specifically, the publication of Nakao et al. establishes or strengthens the claim that combined modality therapy TREATS diabetic nephropathy (Predication: 7762151). The retraction of the article weakens the semantic relation.

A retraction increases the uncertainty of scientific claims. If the affected predications have a unique position, it may in turn influence the uncertainty of a much larger area in scientific knowledge. We will introduce other types of uncertainties associated with scientific knowledge.

Distributions of Uncertainty Cues

The advances of science, engineering, and technology have considerably pushed the boundaries of scientific knowledge. At the same time, the integral role of uncertainty in science has been very much overlooked.

Table 9.4 Semantic predications extracted from Nakao et al. (PMID: 12531578)

Sentence	Predication	Subject	Predicate	Object
62520136	3281525	Angiotensin-converting enzyme inhibitors	TREATS	Diabetic nephropathy
62520293	878505	Maximum	PROCESS_OF	Patients
62520293	1111567	Diabetic nephropathy	PROCESS_OF	Patients
62520293	2931514	Pharmaceutical preparations	INHIBITS	Angiotensin-converting enzyme inhibitors
62520293	4958352	Pharmaceutical preparations	INHIBITS	Receptor, angiotensin II
62520685	1111567	Diabetic nephropathy	PROCESS_OF	Patients
62520804	2653028	Trandolapril	ISA	Angiotensin-converting enzyme inhibitors
62520804	4958352	Pharmaceutical preparations	INHIBITS	Receptor, angiotensin II
62521248	581854	Renal function	COEXISTS_WITH	Excretory function
62521478	7762151	Combined modality therapy	TREATS	Diabetic nephropathy

Table 9.5 shows distributions of uncertainty cue words in the most representative and most comprehensive sources of scientific publications. Google Scholar, the Web of Science, PubMed contain meta-data of scientific publications, whereas ScienceDirect, Springer, Mendeley, and Core are sources of full text articles. For each word, its frequency in a data source is compared with the frequency of the word knowledge. For example, the frequency of the word unknown in Google Scholar is 99% of the frequency of the word knowledge, whereas the frequency of the word contrary only appears 52.59% of the frequency of the word knowledge. Within the same data source, we can compare the popularity of an uncertainty cue word. Between different data sources, we are able to compare the relative frequency. For example, the word unknown is relatively more popular in the Web of Science (132.94% of knowledge). The term uncertainty is most frequently found in Google Scholar (69.52%).

Table 9.6 lists distributions of uncertainty cue words in non-scientific publications. Non-scientific sources including U.S. Supreme Courts, patents and applications, New York, Google, and NSF.

The leading cue words in the U.S. Supreme courts include words such as contrary and controversial. Interestingly, both USPTO and the New York Times are led by the word impossible. The NSF award abstracts are led by the word uncertainty.

Contradictory Claims

These observations have two implications: one on the interestingness and the other on the uncertainty. The interestingness explains the motivations behind the dynamics of the discourse of the argumentation. According to a theory proposed by sociologist Murray Davis (1971), the best way to attract people's attention is to convince them that you can show them that what they believe is questionable. This is the first and the most critical step to get their attention. Davis even suggested that it is possible to routinize this strategy such that one can systematically respond to the current beliefs of a group of people. He identified 12 dialectical relations regarding hypotheses and their antitheses (Table 9.7). For example, if everyone believes that A and B are not connected, then its antithesis argument that A and B are connected is likely to be interesting. If everyone believes that A is changing, then one would be interested in an argument that A is constant. Davis warned that if one takes this strategy too far, it may backfire. The antithesis may sound too ridiculous to retain anyone to listen. Davis' framework is in fact a classification of patterns in our knowledge. If our current belief is in one form of knowledge, an antithesis pattern may provide an alternative interpretation. If it is believed that "A is a B," then one is likely to find it interesting why "A is not a B" is even possible. Similarly, causal relations are an important type of knowledge. Which one should we believe: "HIV causes AIDS" or "HIV does not cause AIDS"?

A large degree of differences between claims on related topics may reflect a degree of uncertainty concerning the status of underlying knowledge. The higher

Table 9.5 Distributions of uncertainty cue words in scientific publications (with reference to the term knowledge)

Google scholar		ScienceDirect		Web of science		Springer		Mendeley		Pubmed		Core	
Unknown	0.990	Unknown	0.6021	Unknown	1.3294	Unknown	0.3416	Unknown	0.3875	Unknown	0.6526	Uncertainty	0.4748
Incomplete	0.755	Conflicting	0.3613	Unusual	0.5966	Conflicting	0.2516	Unclear	0.2166	Undetetmined	0.6203	Unknown	0.4224
Impossible	0.7251	Surpris*	0.3309	Suspect	0.5217	Contrary	0.1998	Uncertainty	0.1806	Unclear	0.3784	Impossible	0.3895
Consensus	0.6992	Contrary	0.3308	Surpris*	0.4091	Impossible	0.1972	Consensus	0.1582	Unusual	0.3269	Surpris*	0.3685
Uncertainty	0.6952	Uncertainty	0.3225	Uncertainty	0.3527	Surpris*	0.1783	Unusual	0.1376	Consensus	0.2398	Contrary	0.3226
Unexpected	0.6394	Unclear	0.2945	Controversial	0.3373	Unclear	0.1625	Contrary	0.1187	Uncertain	0.1965	Consensus	0.2473
Supris*	0.6016	Impossible	0.2901	Contrary	0.3341	Uncertainty	0.1552	Controversial	0.1121	Controversial	0.1747	Incomplete	0.2440
Uncertain	0.5896	Suspect	0.2426	Unclear	0.2754	Incomplete	0.1312	Incomplete	0.1020	Incomplete	0.1490	Ambigu {ity|ous}	0.2266
Unusual	0.5319	Incomplete	0.234	Conflicting	0.2712	Suspect	0.1296	Uncertain	0.0979	Contrary	0.1403	Unusual	0.2164
Contrary	0.5259	Unusual	0.2285	Unexpected	0.2401	Consensus	0.1189	Unexpected	0.0735	Conflicting	0.1200	Inconsistent	0.2039

* is a wildcard, e.g., surpris* including surprises, surprisingly, and surprised

Table 9.6 Distributions of uncertainty cue words in non-scientific publication sources

Supreme		USPTO		NYTimes		Google		NSF	
Contrary	1.3001	Impossible	1.1443	Impossible	1.0749	Unknown	0.9633	Uncertainty	0.1980
Controversial	1.0170	Contrary	1.1167	Unusual	0.8905	Supris*	0.6358	Unusual	0.1161
Dispute	0.9794	Unknown	0.7607	Dispute	0.7309	Dispute	0.6064	Debatable	0.0868
Inconsistent	0.7455	Incomplete	0.3731	Contrary	0.5375	Myster*	0.5284	Conflicting	0.0817
Impossible	0.4704	Unexpected	0.3715	Unknown	0.5128	Impossible	0.3972	Surpris*	0.0672
Ambigu*	0.2825	Supris*	0.3029	Suspect	0.3472	Unusual	0.2358	Incomplete	0.0540
Conflicting	0.2678	Unusual	0.23S0	Unexpected	0.3306	Unexpected	0.1771	Uncertain	0.0516
Doubtful	0.2226	Incompatible	0.1970	Uncertain	0.3192	Suspect	0.1752	Impossible	0.0510
Unusual	0.217S	Inconsistent	0.1898	Suspicion	0.3004	Bizarre	0.1523	Unexpected	0.0400
Uncertainty	0.1844	Unreliable	0.1545	Controversial	0.2822	Controversial	0.1431	Consensus	0.0388

Table 9.7 12 dialectical relations identified by Murray Davis

Phenomenon		Dialectical relations		
Single	Organization	Structured	⟷	Unstructured
	Composition	Atomic	⟷	Composite
	Abstraction	Individual	⟷	Holistic
	Generalization	Local	⟷	General
	Stabilization	Stable	⟷	Unstable
	Function	Effective	⟷	Ineffective
	Evaluation	Good	⟷	Bad
Multiple	Co-relation	Interdependent	⟷	Independent
	Co-existence	Co-exist	⟷	Not co-exit
	Co-variation	Positive	⟷	Negative
	Opposition	Similar	⟷	Opposite
	Causation	Independent	⟷	Dependent

the uncertainty, the more discrepant results there are. The uncertainty reduces as we learn more and more about our topic. Research is driven by the uncertainty in that once a topic has revolved much of its uncertainty, the research of the topic will lose its attraction to researchers. Competing on a settled topic is pointless. A topic with much of its uncertainty resolved would become a good topic for a textbook. The knowledge is codified.

Figure 9.8 depicts the distributions of two contradictory semantic predications found in each year's MEDLINE records. The predication "HIV Causes AIDS" is overwhelming in terms of its volumes (shown in purple). The predication "HIV is not the cause of AIDS" appears almost every year, but its volume is much smaller. The co-existence of contradictory claims indicates a considerable degree of uncertainty. Active researchers are likely to be aware of such uncertainties in their areas of expertise. In fact, one can claim that the domain expertise is the knowledge uncertainty.

The predication "HIV Causes AIDS" has the second strongest burstness (38.7063) among MEDLINE records published between 1900 and 2014. In particular, the predication first appeared in 1984 and it began to burst from 1991 till 2000.

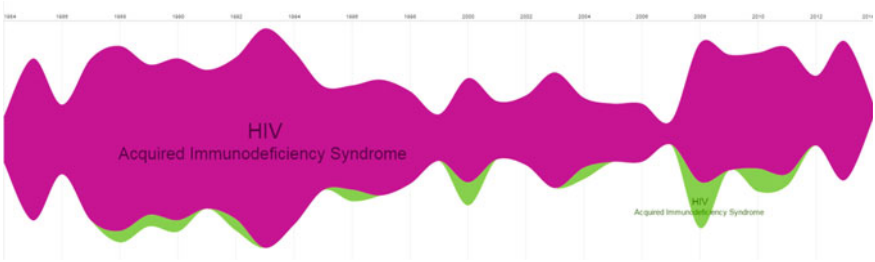

Fig. 9.8 Contradicting semantic predications extracted from MEDLINE records on causal relations between HIV and AIDS

Table 9.8 contains two opposing semantic predications regarding the causal relation between HIV and AIDS. The first four sentences contain the positive predication: HIV causes AIDS, whereas the second four sentences contain the negation of the predication: HIV is not the cause of AIDS. The two predications are contradictory. In the first example, a study suggests that "HTLV-III is the primary cause of human AIDS." The semantic predication of "HIV causes AIDS" partially preserves the meaning of the original sentence. First, the use of hedging word suggest modifies the status of the simplistic predication. Second, HTLV-III in the original sentence is abstract to the broader concept HIV. The more specific term human AIDS is mapped to the broader concept AIDS. Furthermore, the "primary cause" is simplified by the semantic type CAUSES.

If a knowledge system contains contradictory claims, then it is important for a researcher to be able to identify the status of these claims precisely. Furthermore, researchers would often need to take into account the provenance of evidence associated with each of the claims and how such evidence is validated and assessed. If Fuchs' theory is correct, resolving contradictory claims is most likely to play a central role in the work of research fronts because, as a type of competition, resolving contradictory claims would be critical for re-allocating recognition and resources. If contradictory claims appear within the boundary of a specialized area of research, resolving them is unlikely to have a greater degree of impact than that from the first scenario. The specialization effectively shields off much competition. The matter would be even less impactful if contradictory claims are limited to an area of research that is already fragmented off the main stream. To Fuchs, competition leads to scientific change.

Table 9.8 HIV causes AIDS (with the green background) and HIV is not the cause of AIDS (with the pink background)

SID	PMID	Sentence
35528335	6145881	The results strongly indicate that the antibodies to HTLV-III are diagnostic of AIDS or indicate significant risk of the disease, and suggest that **HTLV-III is the primary cause of human AIDS**
34893490	6200936	These results and those reported elsewhere in this issue suggest **that HTLV-III may be the primary cause of AIDS**
35618164	6095415	A transmissible agent especially a retrovirus (HTLV, LAV), is now widely considered in AIDS etiology.
30287966	6100647	HTLV-I is etiologically associated with adult T-cell leukemia-lymphoma (ATLL), HTLV-II has been isolated from a patient with hairy T-cell leukemia, and HTLV-III is the cause of acquired immune deficiency syndrome (AIDS). |
20897139	3399880	HIV is not the cause of AIDS.
33396961	2644642	(iii) pure HIV does not cause AIDS upon experimental infection of chimpanzees or accidental infection of healthy humans.
40872383	8906995	Furthermore, Cys-138 was found in chimpanzee immunodeficiency virus (CIV), a lentivirus that is similar to HIV but does not cause AIDS in chimpanzees.
49995531	1342726	Molecular biologist Peter Duesberg's argument that HIV is not the cause of AIDS is analyzed in light of his contention that a version of Koch's postulates has not been satisfied.

The existence of contradictory claims may potentially lead to the recognition of anomalies, which may in turn overthrow a well established paradigm. The key to determine whether contradictory claims may have the potential for a Gestalt Switch depends on why and how these claims differ. For example, if we consider HIV Causes AIDS, HIV Causes AIDS in human, and HIV does not cause AIDS in chimpanzees as different claims, then there will be no contradiction. On the contrary, if we use the same semantic predication HIV Causes AIDS or the negation of the predication to represent these claims, then our interpretations of these claims are contradictory. The 8th sentence in the table explicitly indicates that the contradiction exists at both levels of the extracted predications and the original writings.

The Reduction of Uncertainty

Table 9.9 demonstrates how the uncertainty associated with a scientific topic may be reduced over time as we learn more about the topic. In 1987, dementia is common in patients with AIDS, but its mechanism was unknown. In 1993, the cause of the AIDS dementia was still unknown, but there was some progress. Radiological and pathological studies have focused on subcortical white matter. In 2000, while the cause of neuronal damage in AIDS was still unclear, its relationships with HIV dementia remain debatable. In 2004, the search narrowed down to HIV-1 transactivating factor Tat. A sequence like this demonstrates how the uncertainty of scientific knowledge can be reduced over time.

A meta analysis is a study of studies that address a set of research questions. A meta analysis statistically normalizes various discrepancies in the findings of studies with equivalent or comparable designs. Ioannidis and Trikalinos (2005) conducted a meta meta-analysis, which means a study of meta-analytic studies. They attempted to answer two questions:

1. How is the between-study variance for studies on the same question changed over time?
2. When did the studies appear with the most extreme results?

They found that the between-study variance appears to decrease over time. They also found that the most extreme results are likely to appear at the beginning period of the research. As shown in Fig. 9.9, the results swung widely with reference to the results published immediately before them. The magnitude of the differences decreases over time.

Table 9.9 The knowledge of the cause of dementia in patients with AIDS

P-VLLD	Year	Subject	Predicate	Object	Sentence
3039662	1987	Dementia	PROCESS_OF	Patients	Dementia is common in patients with AIDS, but the mechanism by which the human immunodeficiency virus type 1 (HTV-1) causes the neurological impairment is *unknown*
3039662	1987	Acquired immunodeficiency syndrome	PROCESS_OF	Patients	Dementia is common in patients with AIDS, but the mechanism by which the human immunodeficiency virus type 1 (HTV-1) causes the neurological impairment is unknown
7689819	1993	White matter	L0CATI0N_0 F	Diagnostic radiologic examination	The cause of acquired immunodeficiency syndrome (AIDS) dementia, which is a frequent late manifestation of human immunodeficiency virus (HIV) infection, is unknown but radiological and pathological studies have implicated alterations in subcortical white matter
10871764	2000	HIV infections	COEXISTS_WITH	Acquired Immunodeficiency Syndrome	Neuronal apoptosis has been shown to occur in HIV infection by a number of in vivo and in vitro studies, however, the cause of neuronal damage in AIDS is still unclear and its relationships with the cognitive disorders characteristic of HIV dementia remain a matter of debate
15361847	2004	AIDS dementia complex	PROCESS_OF	AIDS patient	The HIV-1 transactivating factor, Tat, has been suspected of causing neuronal dysfunction that often leads to the development of HTV-associated dementia in AIDS patients

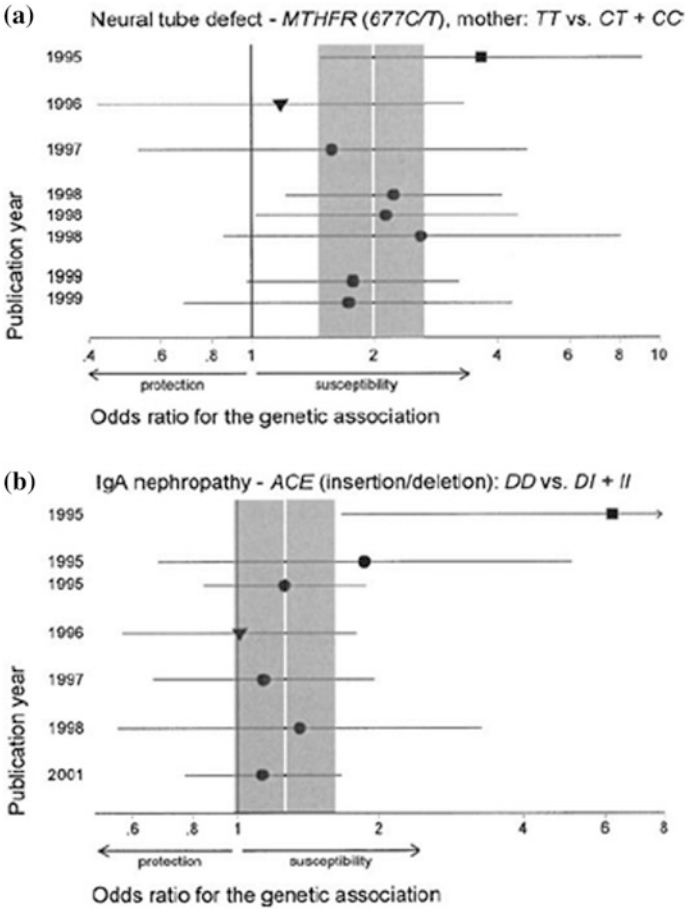

Fig. 9.9 The diversity of published claims decreases over time. *Source* Ioannidis and Trikalinos (2005)

Propositions and Their Epistemic Status

The term meta-discourse in philosophy refers to a discussion about a discussion, as opposed to a simple discussion about a given topic. It also refers to a word or phrase that serves as a guide for the reader on the importance of an example or the role of the text to follow in the discourse. Strictly speaking, meta-discourse is not the subject itself; rather, it provides additional information about the subject. Meta-discourse includes phrases such as "on the other hand," "after all," and "to our best knowledge." In scientific writing, meta-discourse may serve multiple purposes (Table 9.10). It is generally advised that technical, academic, and other non-fiction writers should use meta-discourse but bear in mind not to bury the subject itself.

Table 9.10 Purposes served by meta-discourse

Purpose	Meta-discourse
To denote the writer's confidence	May, perhaps, certainly, must
To denote the writer's intentions	In summary, in a nutshell
To give directions to the reader	Therefore, however, finally
To organize the text	First, second, therefore

Much of scientific assertions found in scholarly publications share a generic structure that consists of two parts: the core of the assertion and a modifier or a descriptor about the assertion. The modifier serves a similar role as the meta-discourse. We can think of many structures that share the same composite pattern in which one part serves the central role and the other part characterizes the central part.

- Analysis = Meta-Analysis + Analysis Proper
- Data = Meta-Data + Data Proper
- Message = Meta-Message + Message Proper
- Discourse = Meta-Discourse + Discourse Proper
- Knowledge = Meta-Knowledge + Knowledge Proper
- Statement = Epistemic Status (Meta-Proposition) + Proposition.

What we are interested is the last one on the list: a statement is seen as a proposition and its epistemic status. For instance, given the statement that *the mechanism of the disease is unknown*, the statement that mechanism is unknown conveys the epistemic status of the subject. Consider another example, *there is currently no consensus on what causes the disease.* The *what causes the disease* is the core message, whereas the lack of consensus is the epistemic status, or the meta-knowledge.

Research on representing scientific knowledge has overwhelmingly focused on the Proposition part of a statement. For instance, Semantic MEDLINE's semantic predications essentially correspond to the Proposition part of the pattern. Given a semantic predication of HIV CAUSES AIDS, none of its epistemic status nor the provenance of its evolution is preserved—the meta-knowledge is not accessible in association with the plainly expressed semantic predication. There is no trace of its original context. There is no indication how confidently the claim was made. There is no sign of any controversies involved. Thus we refer to this type of information as propositions, which form part of scientific knowledge but they are not complete in that one cannot make any meaningful inference or reasoning just based on propositions without knowing to what extent they are considered true and to what extent they are still unknown.

The Epistemic Status part of the statement is largely overlooked with notable exceptions in the study of hedging in scientific writing (Hyland 1996). The Epistemic Status part is meta-discourse in nature because it guides the reader about how to interpret the Proposition part. The use of hedging words is a sign of

uncertainty at least from the position of the writer. A clause that contains suggestions of incomplete, conflicting, or contradictory information presents evidence that the certainty of a proposition is questionable.

The following MySQL query highlights the two-part structures in scientific writing. In particular, the query finds sentences that contain a specific claim and a meta-discourse that qualifies the claim. The query searches for the phrase "claim that" as the anchor and shows the text before and after the anchor phrase. This is a commonly used information search method known as the Keyword in Context (KiC) method. The table contains a text field—context—of paragraphs from scientific articles.

```
SELECT article_id, substring(context, if(locate('claim that',
      context)>30, locate('claim that', context)-30, 1), 60) As
      KiC
FROM fulltext
WHERE project='sample' AND context LIKE '% claim that %'
LIMIT 20;
```

Table 9.11 shows examples of sentences that are joined by the anchor phrase 'claim that'—the text before the anchor is serving the role of a meta-discourse, whereas the text follows the anchor is the actual claim the authors are making. For instance, several cases are indirect quotations from published articles. In two of the examples, authors exclude a claim rather than make a claim.

Table 9.12 shows examples of the contexts in which the word 'uncertainty' are used. The level of uncertainty varies from 'entirely uncertain,' 'in part, fragmentary and uncertain,' 'at best difficult and uncertain,' to 'the extent of … is uncertain,' and 'the ultimate role of … is uncertain.'

Separating sentences into such two parts allows us to study the dynamics of uncertainty and its role in the development of scientific knowledge. The absence of the epistemic status part commonly implies that the proposition is considered true or valid. For instance, HIV causes AIDS is equivalent to a statement: *research has long established that HIV causes AIDS*. The length of the epistemic status part may serve as a simplistic indicator of the level of uncertainty—the longer the string length of this part, the higher the likelihood of the uncertainty. Of course, it is quite conceivable that one can express a high level of uncertainty concisely.

A useful device to analyze groups of words rather than individual words is a dependency graph. Since we need to effectively separate the proposition from a description of its epistemic status or other types of modifiers and wrappers, dependency graphs lend us graph-theoretic properties as well as linguistic and semantic relations. In the following examples, we will illustrate how we can identify a proposition and its epistemic status from a corresponding dependency graph. Furthermore, we will search for patterns that can be computationally processed and synthesized.

Table 9.11 Examples of claims and leading meta-discourse

#	Article ID	KiC
1	2007057	with their triple helix model **claim that** the contribution of
2	2007068	On this basis they claim that the technology reflected in th
3	1994398	of study and Hannam's (2009) claim that tourism studies is
4	2418115	give evidence to support the claim that improvements in edu
5	1930043	roper use of the term, we can claim that research broadly re
6	2131039	Moya-Anegon et al. [3] also claim that 85% of the journals in
7	2139416	1988), and Granovetter (1973) claim that individuals' person
8	2055620	We make no claim that the resulting sample is by any means a
9	2416554	Sipido et al.20 claim that the average life expectancy of pa
10	1982348	Legl23 corroborated Davis's24 claim that library instruction
11	1982356	s corroborated Davis's (2003) claim that library instruction
12	1909729	urrent discussion, we make no claim that DEA suffers from su
13	1953896	Third, we claim that the really destructive critique to the
14	2355226	iz-Baños, and Courtial (2005) claim that power laws are not
15	2346377	y, Baten and Muschalli (2012) claim that since the 1990s eco
16	2199679	(1998) claim that personality varies with structural holes a
17	2199787	Finally, we claim that the emergence of strategic roles can
18	1965013	One could claim that the quality incentive is embedded in th
19	2078654	imulations in Japan and China claim that the reduction impac
20	2078783	Many analysts claim that the use of green roofs is an effici

Table 9.12 Sentences containing the word 'uncertainty' in MEDLINE articles

PMID	Sentence
5321391	The duration of function of individual grafts is entirely uncertain at present
5940637	Severe osteomalacia of uncertain etiology was observed in a 44-year-old woman
11526856	The behavioral role of these response sequences is uncertain
11881655	All three approaches are beset with uncertainties, and it is important to state at the outset that no completely convincing evidence exists for extraterrestrial life
12056428	On the basis of these data that are, in part, fragmentary and uncertain, upper and lower limits of rad doses under different amounts of mass shielding are estimated
13118110	The extent of the uptake, however, is uncertain, again because of the liberation of chromogenic substances
13561107	The ultimate role of these agents in the treatment of major emotional disorders, such as schizophrenic reactions, still is uncertain
13684978	The value and risks of the procedure have been examined in 20 patients with obstructive jaundice of uncertain origin and in one further patient with a post-cholecystectomy syndrome
14287175	The assays indicated 1.2–2.6% RNA, similar to previously published work, but only 0.0–1.0% DNA, near enough the sensitivity limits to render the presence of DNA in the preparations uncertain
14792375	Prognosis in pancreatitis is at best difficult and uncertain

Dependency Graphs

Given a sentence, the dependent relations derived from the sentence can be represented in a dependency graph as shown in Fig. 9.10. The original sentence "A transmissible agent especially a retrovirus (HTLV, LAV), is now widely considered in AIDS etiology." is from an article published in 1984 (PMID: 6095415). The dependence graph divides the sentence into a few groups of words. For example, the semantic predication "HIV CAUSES AIDS" is extracted from the segment "(HTLV, LAV) is now widely considered in AIDS etiology."

The dependency parser from the Stanford NLP library identifies HTLV and LAV as the subject of this segment (nsubjpass). The word considered/VBN-12 means that it is a verb at the 12th position of the sentence. The text "is now widely" modifies the word considered, thus in the dependency graph, they are shown as the three nodes below the considered node. By retaining words with specific dependency types, we can computationally simplify a sentence by retaining the most salient message. For example, instead of considering the entire sentence, we can focus on the key message: HTLV and LAV are considered in AIDS etiology.

It is intuitively easy to separate a proposition from its conditional or contextual wrapper from a dependency graph because it is straightforward to identify sub-graphs that correspond to the two parts. For example, in the dependency graph shown in Fig. 9.10, the core proposition is represented by the sub-graph located at the lower right part of the graph, whereas the sub-graph on the left represents a modifier of the former sub-graph. The number [1] in Fig. 9.10's caption means that this is the first sentence in the abstract of the MEDLINE article.

The dependency graph shown in Fig. 9.11 represents a long and complex sentence from a 1984 article (PMID: 6100647). This is the 4th sentence from the abstract of the article:

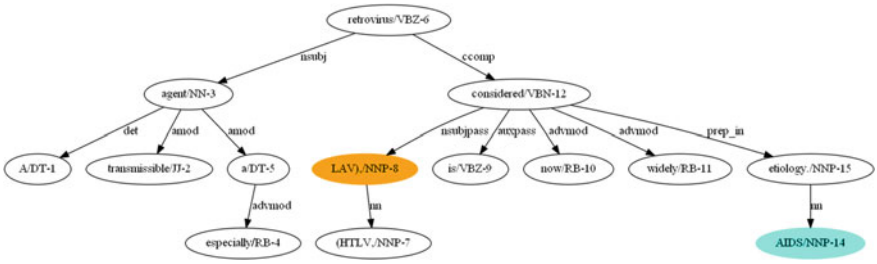

Fig. 9.10 The first appearance of the predication "HIV CAUSES AIDS" in 1984 (PMID: 6095415; SID: 35618164 [1]: "A transmissible agent especially a retrovirus (HTLV, LAV), is now widely considered in AIDS etiology.")

HTLV-I is etiologically associated with adult T-cell leukemia-lymphoma (ATLL), HTLV-II has been isolated from a patient with hairy T-cell leukemia, and HTLV-III is the cause of acquired immune deficiency syndrome (AIDS).

The sentence contains three statements. The HIV-CAUSES-AIDS predication is extracted from the last statement, which is represented by the sub-graph of the word cause/NN-25. By filtering the dependency types, we can simplify the sub-graph to a much simpler graph: HTLV-III—cause—AIDS. The complexity of the sentence is clearly reflected in the complexity of the dependency graph. The dependency graph provides a sense of context for the predication of our interest as well as other predications.

In Fig. 9.12, the predication is derived from the sub-graph at the lower right of the graph under the word cause: HTLV-III—cause—AIDS. The sub-graph as a whole is the object of the verb suggest/VBP-22, which is the verb at the 22nd position of the sentence. Words such as suggest are considered as hedging words. Writers often use hedging words to express a degree of uncertainty of their statements. A statement expressed with hedging words implies that the writer does not

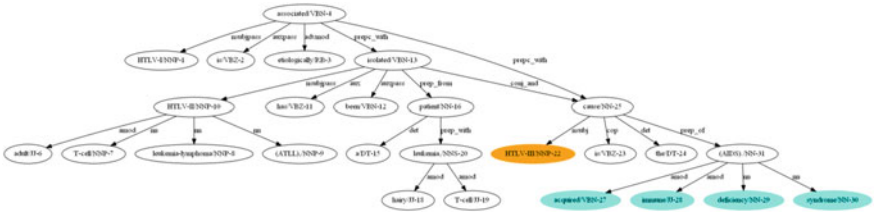

Fig. 9.11 The dependency graph of a sentence in a 1984 article (PMID: 6100647; SID: 30287966 [4]). HTLV-I is etiologically associated with adult T-cell leukemia-lymphoma (ATLL), HTLV-II has been isolated from a patient with hairy T-cell leukemia, and HTLV-III is the cause of acquired immune deficiency syndrome (AIDS)

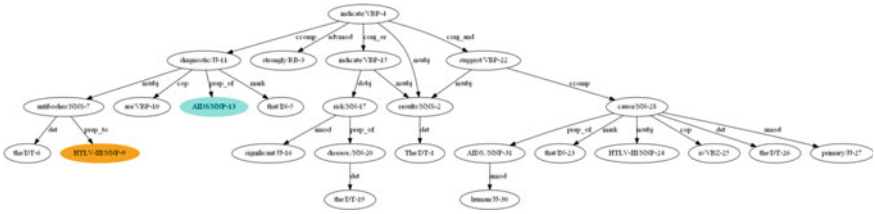

Fig. 9.12 The dependency graph of a sentence from a 1984 article (PMID: 6145881; SID: 35528335[6]). This is the 6th sentence in the abstract: *The results strongly indicate that the antibodies to HTLV-III are diagnostic of AIDS or indicate significant risk of the disease, and suggest that HTLV-III is the primary cause of human AIDS*

rule out the possibility of exceptions. Without hedging, the predication sounds like "HIV causes AIDS, period!" With hedging, it conveys that the status of the statement may be conditional on other factors, for example, "To our best knowledge, HIV causes AIDS."

The example shown in Fig. 9.13 also contains a hedging word suggest. In addition, there is another layer of hedging—may be—in the core statement: HTLV-III may be the primary cause of AIDS. It is reasonable to perceive that this sentence has a higher degree of uncertainty than the one in the previous example because of the presence of two levels of hedging. The word cause is modified by the word primary, which can be seen as another level of hedging because it does not rule out other possible causes. The three levels of hedging make the statement as precise as the writer wants to convey his/her best knowledge about this matter. The writer only needs to do that when the real status of the proposition is still uncertain. Therefore, the presence of hedging is an indicator that the scientific assertion in question is associated with a considerable degree of uncertainty.

The dependency graph shown in Fig. 9.14 is complicated. The predication in the complex sentence boils down to a short statement re-constructed from the graph: A direct role of PBM in the pathogenesis of AIDS is postulated.

The Length of Uncertain Statements

The dependency graph in Fig. 9.15 contains a segment that led to the extraction of the predication "HIV CAUSES AIDS." The subject was Barbara Hogan. The text includes a segment that she affirmed that HIV causes AIDS. The sub-graph that represents the assertion is very simple, as colored in the graph. This observation leads us to propose another way to measure the uncertainty of a scientific assertion: the longer an assertion is in terms of the total number of words, the more uncertain it is likely to be. In other words, if one has to state a claim with uncertainty, then he/

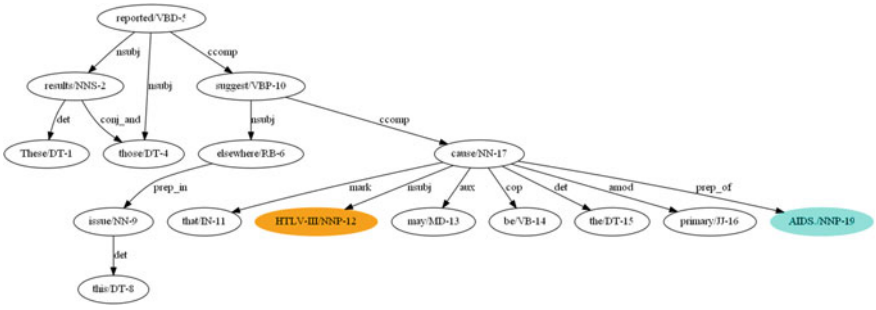

Fig. 9.13 The dependency graph of a sentence from a 1984 article (PMID: 6200936; SID: 34893490[7]). This is the 7th sentence in the abstract: *These results and those reported elsewhere in this issue suggest that HTLV-III may be the primary cause of AIDS*

Fig. 9.14 The dependency graph of a sentence in a 1990 article (PMID: 2104787; SID: 18493183 [16]). This is the 16th sentence in the abstract: *The results of this study clearly indicate that PBM from HIV+ individuals are endowed with the capacity to mediate ADCC against HIV-infected/coated cells and thus, we postulate that PBM may play a direct role* in vivo *in lysis or suppression of HIV-coated/infected cells and in the pathogenesis of AIDS*

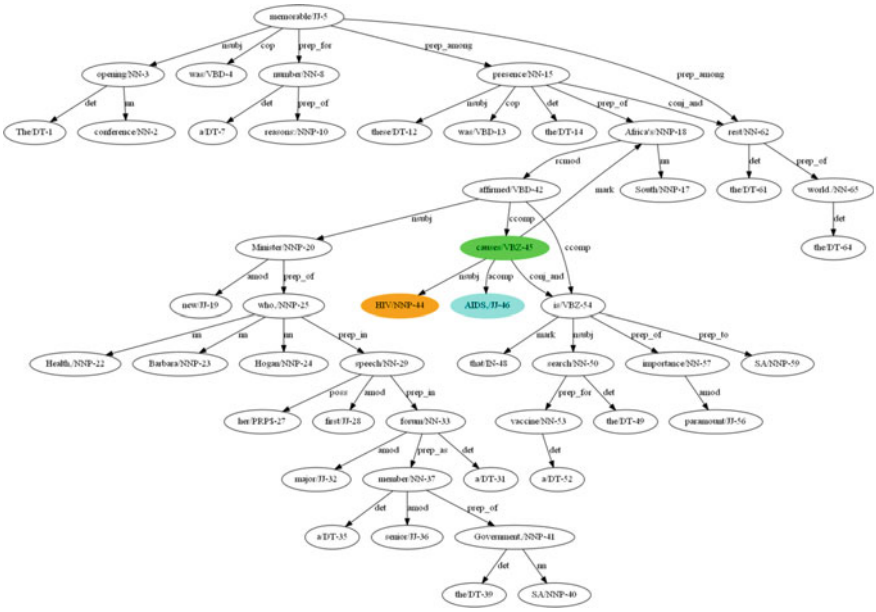

Fig. 9.15 A dependency graph of a sentence in a 2009 article (PMID: 19202348; SID: 120435934[12]). This is the 12th sentence in the abstract: *The conference opening was memorable for a number of reasons: among these was the presence of South Africa's new Minister of Health, Barbara Hogan who, in her first speech in a major forum as a senior member of the SA Government, affirmed that HIV causes AIDS, and that the search for a vaccine is of paramount importance to SA and the rest of the world*

she is likely to include hedging, restrictions, limiting conditions, exceptions, or other factors or sources of uncertainty.

The dependency graph in Fig. 9.16 represents the title of an article (PMID: 3399880). This is a very short statement: HIV is not the cause of AIDS. It is hard to imagine how it can be shortened any further. It has no hedging, no qualifying conditions, and no exceptions. As far as it is concerned as an assertion, it is absolutely certain; or, its uncertainty is zero. The author of the sentence is very confident.

The example shown in Fig. 9.17 is the 8th sentence from a 1990 article (PMID: 1980675). A negative predication is extracted from the sentence. The sentence as a whole simply states the position of Duesberg. There is no hedging or other indicator of uncertainty. How should we assess the uncertainty of an indirect quote of a statement? We have two options: 1. Assume that the paraphrasing statement should have the same level of uncertainty as the uncertainty of the original statement, or 2. Consider that the paraphrasing statement has a higher level of uncertainty. We believe the latter makes more sense because its writer is not really taking the responsibility for the core claim.

Figure 9.18 shows another example of the negation of the predication "HIV CAUSES AIDS." This sentence makes two points: CIV is similar to HIV, but CIV does not cause AIDS in chimpanzees. Thus, the predication "HIV does not cause AIDS" extracted by SemMed does not preserve the original meaning of the text. This is another type of uncertainty. It is introduced in the process of mapping to a semantic type.

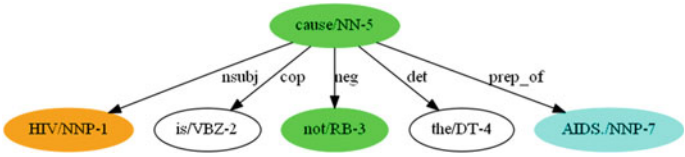

Fig. 9.16 The dependency graph of the title of a 1988 article (PMID: 3399880; SID: 20897139 [title]). The title is: *HIV is not the cause of AIDS*

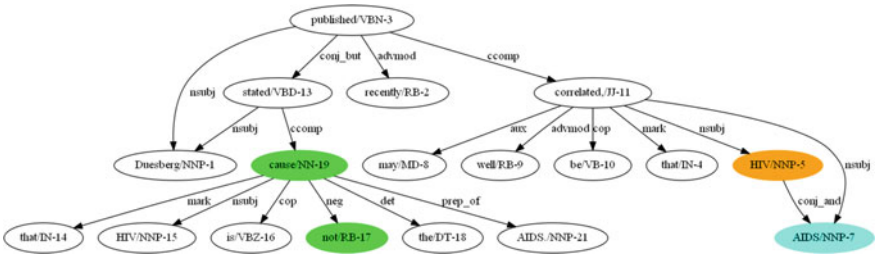

Fig. 9.17 The dependency graph of a sentence of a 1990 article (PMID: 1980675; SID: 51884237 [8]). This is the 8th sentence in the abstract: *Duesberg recently published that HIV and AIDS may well be correlated, but stated that HIV is not the cause of AIDS*

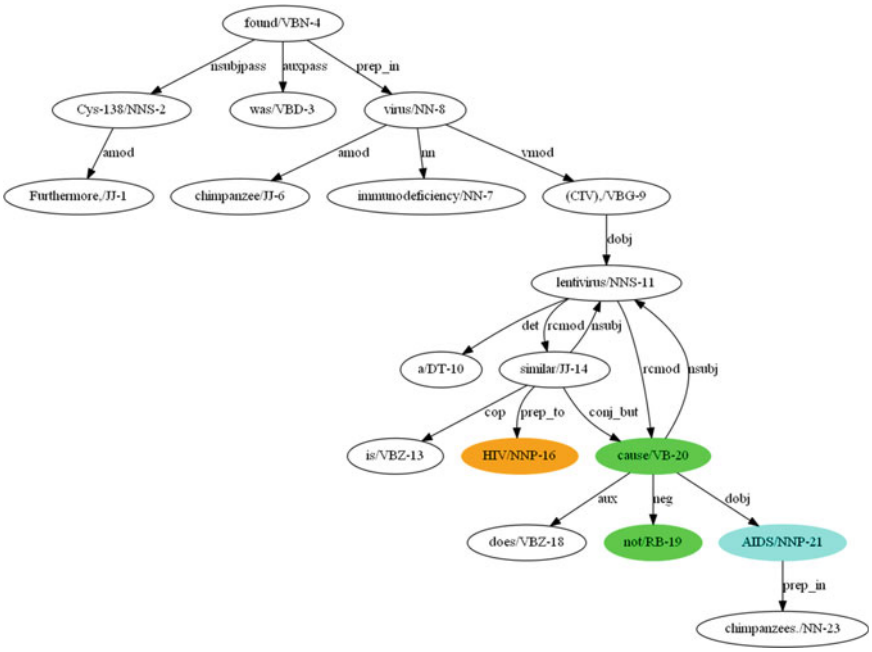

Fig. 9.18 The dependency graph of a sentence of a 1996 article (PMID: 8906995; SID: 40872383 [8]). This is the 8th sentence in the abstract: *Furthermore, Cys-138 was found in chimpanzee immunodeficiency virus (CIV), a lentivirus that is similar to HIV but does not cause AIDS in chimpanzees*

The example shown in Fig. 9.19 demonstrates another type of uncertainty. The core statement was "HIV could not cause AIDS simply through direct cytopathic mechanisms alone." Does it mean that HIV does not cause AIDS? Does it mean that HIV may cause AIDS through other mechanisms or a combination of multiple types of mechanisms? This type of uncertainty is resulted from the ambiguity that is unlikely to be resolvable at the level of individual sentences.

Figure 9.19 shows a streamgraph visualization. It depicts the volume of a stream of each semantic predication of causal relations found in SemMedDB. The width of a stream at a particular year is proportional to the number of articles in which the predication appears. Each stream is labeled by the subject and the object of the predication. The semantic type is not labeled because they are all causal relations. For example, the predication "HIV CAUSES AIDS", labeled as HIV/Acquired Immunodeficiency Syndrome in the streamgraph, emerged in 1984. It had the widest stream in 1985. In 1986, the most popular predication was "Retroviridae CAUSES Acquired Immunodeficiency Syndrome," but the predication "HIV CAUSES AIDS" became the most popular one again in 1987 and 1988. From this simple visualization, we learn that the research of HIV and AIDS was most active during 1984 and 1988 (Figure 9.20).

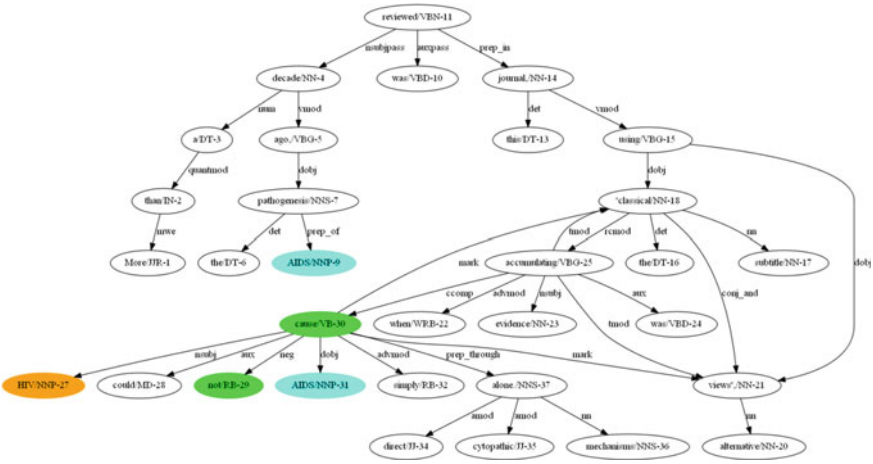

Fig. 9.19 The dependency graph of a sentence from a 2008 article (PMID: 18624032; SID: 111111060[1]). This is the opening sentence of the abstract: *More than a decade ago, the pathogenesis of AIDS was reviewed in this journal, using the subtitle 'classical and alternative views', when evidence was accumulating that HIV could not cause AIDS simply through direct cytopathic mechanisms alone*

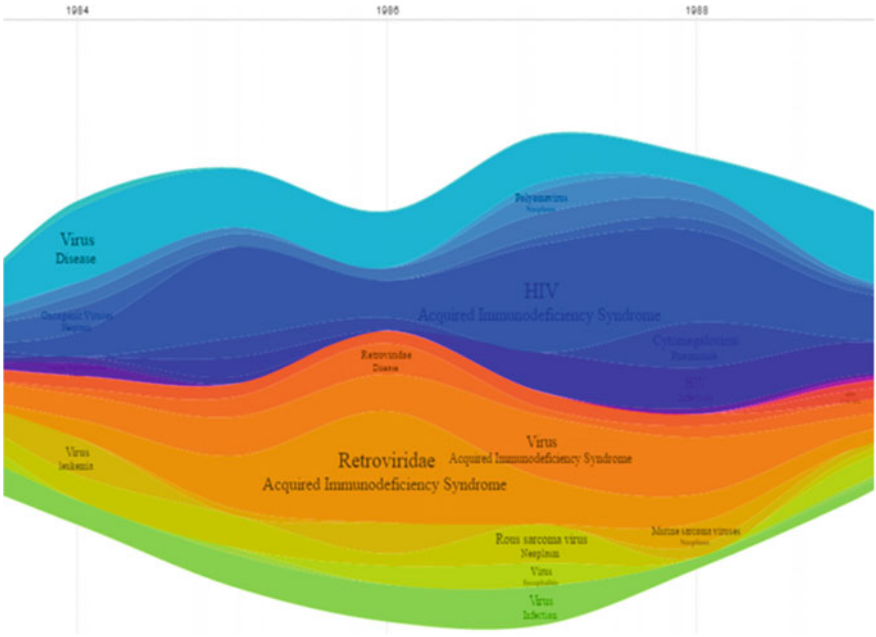

Fig. 9.20 A streamgraph visualization of semantic predications between 1984 and 1989

Summary

The understanding of the type and the degree of the uncertainty associated with a scientific proposition is about the epistemic status of scientific knowledge—it is meta-knowledge of science. Without the meta knowledge, a scientist would be like someone who only learns how to swim by reading books. Without the meta knowledge, scientists will have no way to differentiate codified knowledge from knowledge that is in the making.

The mainstream research of representing scientific knowledge has overwhelmingly focused on predications extracted from scientific literature. While representing scientific knowledge in a simplified form may serves important goals, in a long run, the omission of their epistemic status from the representation of scientific knowledge is likely to hinder the accessibility of scientific knowledge. Many problems with policy and administrative implications may not be adequately resolved. The public understanding of science may not offer the public with efficient and effective means to clarify controversies of scientific debates or reconcile contradictory results and interpretations in scientific literature.

Understanding the wide variety of types of uncertainties in science and their roles in the advance of science itself and in the broader context of everyday life is the first step towards a better understanding of how science works. A high level of uncertainty may attract more competitions because it may imply a potentially higher reward. A sudden increase of uncertainty may indicate the emergence of a new paradigm. Once the perceived level of uncertainty drops below a level, an area of research may lose its attraction. For funding agencies and peer reviewers of high-risk and high-reward programs, the perceived risk and the underlying uncertainty are the two sides of the same coin. They are integral part of innovative and competitive research. They should be treated as such.

Concluding Remarks

We began the book with Heilmeier's Catechism as a desired level of clarity and granularity in communicating scientific knowledge effectively. A competent scholar should be able to communicate complex scientific work that people without the relevant domain knowledge can understand to the extent it matters to them. For example, how many days of Ebola quarantine would be sensible? What is the key to help more people to understand controversies about climate change?

Rome was not built in one day. Many research programs' pragmatic values may not become clear for many generations. What are the arguments for or against supporting basic research as opposed to applied sciences? To put these questions in perspective, we introduced three major theories of scientific change at macroscopic levels from three distinct perspectives—philosophical, sociological, and evolutionary. The value of these theoretical visions is twofold: armed with these theories, we have a rich set of tangible properties that we can match and verify from different

perspectives, and we can start to construct a theory of our own that may connect predictions made by existing theories and reconcile inconsistencies across different expectations. Macroscopic theories of science focus on holistic properties of scientific domains. The notion of a scientific domain is a generic concept of a complex adaptive system, which may exist across multiple levels of granularity. It is valuable to develop a vision at this level to see the forest of scientific knowledge as well as the individual trees.

At lower levels of abstraction, we have reviewed a series of information metrics that measures the importance of information, semantic relatedness, and scholarly impact. An important issue concerning all the quantitative indicators is how to normalize a measurement to minimize bias and makes a comparison fair. Given the ever increasing enthusiasm in ranking increasingly diverse and heterogeneous targets, it is essential to be aware of the basic principles and implications of various normalization schemes.

Text mining techniques and applications in biomedical domains in particular are introduced. Pioneering, intermediate, and recent developments are outlined to highlight the major milestones in the course of development.

Semantic MEDLINE is a very valuable resource. It helps us understand many significant properties of semantic predications extracted from unstructured text. We illustrated how to utilize visual analytic functions in CiteSpace to explore semantic networks constructed from semantic predications. We outlined the development of an ambitious plan—a Visual Analytic Observatory of Scientific Knowledge (VAO) as the first step towards representing scientific knowledge that takes the uncertainty of science into account. We demonstrated two major sources of uncertainty in scientific literature, namely hedging and contradictory information. Finally, we illustrated a series of uncertainty types through dependency graphs of sentences of various complexity.

The uncertainty associated with a research question drives the research. The unknown or the uncertainty makes a competition meaningful because a competition needs a problem to solve. As the research advances, the level of uncertainty reduces and the competition becomes less motivated. Scientists either move elsewhere to challenge themselves with new problems or they proceed with specializations by using codified and routinized knowledge that has little room for uncertainty.

References

Angeles ADL, Ferrari F, Fujiwara Y, Mathieu R, Lee S, Lee S, Tu H-C, Ross S, Chou S, Nguyen M, Wu Z, Theunissen TW, Powell BE, Imsoonthornruksa S, Chen J, Borkent M, Krupalnik V, Lujan E, Wernig M, Hanna JH, Hochedlinger K, Pei D, Jaenisch R, Deng H, Orkin SH, Park PJ, Daley GQ (2015) Failure to replicate the STAP cell phenomenon. Nature 525(7570):E6–E9. doi:10.1038/nature15513

Bibby K, Casson LW, Stachler E, Haas CN (2015) Ebola virus persistence in the environment: state of the knowledge and research needs. Environ Sci Technol Lett 2(1):2–6

Behnam B, Naeimi A, Darvishzade A (2012) A comparative genre analysis of hedging expressions in research articles: is fuzziness forever wicked? Engl Lang Lit Stu 2(2):20–38

Chaomei Chen, Ming Song, Go Eun Heo (2017) A Scalable and Adaptive Method for Finding Semantically Equivalent Cue Words of Uncertainty. arXiv:1710.08327. https://arxiv.org/abs/1710.08327

Chen C (2016) Grand challenges in measuring and characterizing scholarly impact. Frontiers Res Metrics Analytics. doi:10.3389/frma.2016.00004

Chen C, Hu Z, Milbank J, Schultz T (2013) A visual analytic study of retracted articles in scientific literature. J Am Soc Inf Sci Technol 64:234–253. doi:10.1002/asi.22755

Clark C, Aberdeen J, Coarr M, Tresner-Kirsch D, Wellner B, Yeh A, Hirschman L (2011) MITRE system for clinical assertion status classification. J Am Med Inform Assoc 18(5):563–567

Clark M, Kim Y, Kruschwitz U, Song DW, Albakour D, Dignum S, Beresi UC, Fasli M, De Roeck A (2012) Automatically structuring domain knowledge from text: an overview of current research. Inf Process Manage 48(3):552–568. doi:10.1016/j.ipm.2011.07.002

Cross N (1997) Creativity in design: analyzing and modeling the creative leap. Leonardo 30 (4):311–317

Davis MS (1971) That's interesting! towards a phenomenology of sociology and a sociology of phenomenology. Philos Social Sci 1(2):309–344

de Knijff J, Frasincar F, Hogenboom F (2013) Domain taxonomy learning from text: the subsumption method versus hierarchical clustering. Data Knowl Eng 83:54–69

Di Marco C, Kroon F, Mercer R (2006) Using hedges to classify citations in scientific articles. In: Shanahan J, Qu Y, Wiebe J (eds) Computing attitude and affect in text: theory and applications, vol 20. The Information Retrieval Series. Springer, Netherlands, p 247–263. doi:10.1007/1-4020-4102-0_19

Falahati R (2006) The use of hedging across different disciplines and rhetorical sections of research articles. In: Proceedings of the 22nd NorthWest Linguistics Conference (NWLC22), Burnaby, February 18–19, 2006

Fuchs S (1993) A sociological theory of scientific change. Soc Forces 71(4):933–953

Greenberg SA (2009) How citation distortions create unfounded authority: analysis of a citation network. BMJ 339:b2680

Harris Z (1954) Distributional structure. Word 10(23):146–162

Heffernan O (2007) Clarity on uncertainty. Nature Reports, Climate Change, p 5

Horn K (2001) The Consequences of Citing Hedged Statements in Scientific Research Articles: When scientists cite and paraphrase the conclusions of past research, they often change the hedges that describe the uncertainty of the conclusions, which in turn can change the uncertainty of past results. Bioscience 51(12):1086–1093. doi:10.1641/0006-3568(2001)051[1086:tcochs]2.0.co;2

Hyland K (1996) Talking to the academy: forms of hedging in science research articles. Written Commun 13(2):251–281

Hyland K (1998) Boosters, heding and the negotiation of academic knowledge. Text 18(3):349–382

Ioannidis JPA, Trikalinos TA (2005) Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetic research and randomized trials. J Clin Epidemiol 58:543–549

Jensen JD (2008) Scientific uncertainty in news coverage of cancer research: effects of hedging on scientists' and journalists' credability. Human Commun Res 34:347–369. doi:10.1111/j.1468-2958.2008.00324.x

Johnson BB, Slovic P (2015) Fearing or fearsome Ebola communication? Keeping the public in the dark about possible post-21-day symptoms and infectiousness could backfire. Health, Risk & Society 17(5–6):458–471

Kakuk P (2009) The legacy of the Hwang case: research misconduct in biosciences. Sci Eng Ethics 15:545–562

Kilicoglu H, Bergler S (2008) Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. BMC Bioinformatics 9(Suppl 11):S10

Kuhn TS (1962) The structure of scientific revolutions. University of Chicago Press, Chicago

Lewandowsky S, Gignac GE, Vaughan S (2013) The pivotal role of perceived scientific consensus in acceptance of science. Nat Climate Change 3(4):399–404. doi:10.1038/nclimate1720
Light M, Qiu X, Srinivasan P (2004) The language of bioscience: facts, speculations, and statements in between. Paper presented at the HLT-NAACL 2004 Workshop, Biolink, 2004
Lippi M, Torroni P (2016) Argumentation mining: state of the art and emerging trends. ACM Trans Internet Technol 16(2):10:11–10:25
Malhotra A, Younesi E, Gurulingappa H, Hofmann-Apitius M (2013) 'HypothesisFinder:' a strategy for the detection of speculative statements in scientific text. PLoS Comput Biol 9(7): e1003117
McDonald S, Ramscar M (2001) Testing the distributional hypothesis: the influence of context on judgements of semantic similarity. Proceedings of the 23rd annual conference of the cognitive science society. pp 611–616
Medlock B (2008) Exploring hedge identification in biomedical literature. J Biomed Inform 41:636–654. doi:10.1016/j.jbi.2008.01.001
Medlock B, Briscoe T (2007) Weakly supervised learning for hedge classification in scientific literature. Paper presented at the proceedings of the 45th annual meeting of the association of computational linguistics, Prague, Czech Republic, June 2007
Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119
Nakao N, Yoshimura A, Morita H, Takada M, Kayano T, Ideura T (2003) Combination treatment of angiotensin-II receptor blocker and angiotensin-converting-enzyme inhibitor in non-diabetic renal disease (COOPERATE): a randomised controlled trial. Lancet 361(9352):117–124
Noorden Rv (2014) Publishers withdraw more than 120 gibberish papers. Nature. doi:10.1038/nature.2014.14763
Piffer D (2012) Can creativity be measured? An attempt to clarify the notion of creativity and general directions for future research. Thinking Skills Creativity 7(3):258–264. doi:http://dx.doi.org/10.1016/j.tsc.2012.04.009
Rizomilioti V (2006) Exploring epistemic modality in academic discourse using corpora. In: Macia EAo, Cervera AS, Ramos CR (eds) Information technology in languages for specific purposes of educational linguistics. Springer, New York, USA, p 53–71
Rzhetsky A, Iossifov I, Loh JM, White KP (2006) Microparadigms: chains of collective reasoning in publications about molecular interactions. PNAS 103(13):4940–4945. doi:10.1073/pnas.0600591103
Service RF (2002) Bell Labs fires star physicist found guilty of forging data. Science 298:30–31
Shneider AM (2009) Four stages of a scientific discipline: four types of scientists. Trends Biochem Sci 34(5):217–223
Summers-Stay D, Voss C, Cassidy T (2016) Using a distributional semantic vector space with a knowledge base for reasoning in uncertain conditions. Biologically Inspired Cogn Architectures 16:34–44
Szarvas G, Vincze V, Farkas R, Csirik J (2008) The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical text. BioNLP 2008: current trends in biomedical natural language processing. Association for Computational Linguistics, Columbus, Ohio, USA, pp 38–45
Szarvas G, Vincze V, Farkas R, Mora G, Gurevych I (2012) Cross-genre and cross-domain detection of semantic uncertainty. Comput Linguist 38(2):335–367
Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. Science 342(6157):468–472
van Raan AFJ (2004) Sleeping beauties in science. Scientometrics 59(3):461–466
Vincze V, Szarvas G, Farkas R, Mora G, Csirik J (2008) The BioScope corpus: biological texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 9(Suppl 11):S9
Wager E, Williams P (2011) Why and how do journals retract articles? An analysis of Medline retractions 1988-2008. J Med Ethics 37:567–570

Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, Berelowitz M, Dhillon AP, Thomson MA, Harvey P, Valentine A, Davies SE, Walker-Smith JA (1998) Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children (Retracted article. See vol 375, pg 445, 2010). The Lancet 351(9103):637–641

Zhu X, Turney P, Lemire D, Vellino A (2015) Measuring academic influence: not all citations are equal. J Assoc Inf Sci Technol 66(2):408–427