



UNIVERSIDAD POLITÉCNICA DE MADRID

**ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA
AGRONÓMICA, ALIMENTARIA Y DE BIOSISTEMAS**

DEPARTAMENTO DE BIOTECNOLOGÍA

***Detection and formal classification of certainty and
its application to text mining of chains of scholarly
statements.***

TESIS DOCTORAL

Mario Prieto Godoy

Licenciado en Biología

Madrid, 2019



**Programa Oficial de Doctorado en Biotecnología y
Recursos Genéticos de Plantas y Microorganismos
Asociados**

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRONÓMICA,
ALIMENTARIA Y DE BIOSISTEMAS

DEPARTAMENTO DE BIOTECNOLOGÍA

UNIVERSIDAD POLITÉCNICA DE MADRID

TESIS DOCTORAL

*Detection and formal classification of certainty and its
application to text mining of chains of scholarly statements.*

Autor:

Mario Prieto Godoy
Licenciado en Biología

Director:

Mark D. Wilkinson
B.Sc. Honors Genetics
Ph.D. Botany (Plant Molecular Biology)

Madrid, 2019

UNIVERSIDAD POLITÉCNICA DE MADRID

Tribunal nombrado por el Magfco. y Excmo. Sr. Rector de la
Universidad Politécnica de Madrid.

Presidente:

Secretario:

Vocal:

Vocal:

Vocal internacional:

Suplente:

Suplente:

In science, like in other moments of our lives, we find situations that make us hesitate. Hesitating on what we knew, even sometimes hesitating about our foundations. The progress arises in the science of wondering, why? or how? Sometimes we need to wonder what other says and ask ourselves “what if...?”, what if this outcome is not true? and if what this author wanted to say was something else?

La ciencia no se debería contar según lo que uno cree, sino pensando en cómo los demás van a interpretar aquello que tú les cuentas.

ACKNOWLEDGEMENTS

Bueno hora de los agradecimientos, en primer lugar, agradecer a Mark la oportunidad que me dio para hacer el doctorado. Además, agradecerle todo el apoyo, buen rollo y libertad en el lab para dirigir mi tesis, aunque en sus inicios fue algo agobiante tanta libertad, me ha dado una autonomía y perspectiva diferente de hacer las cosas que probablemente lleve conmigo siempre. Pero sobre todo he de agradecer la ayuda este último año, un año largo, pesado e intenso pero que al final ha dado sus frutos. Gracias.

También he de agradecer a mis compañeros de laboratorio: A Pablo y Miriam por sus comentarios y ayuda siempre que la he necesitado. A Alex jr., por todas las risas, buen rollo y consejos que hicieron falta tanto dentro como fuera del lab. A Marco por la toda la ayuda con la estadística y esos cafés tan buenos como necesarios a mitad de mañana. Pero sobre todo a Bea, por estar siempre disponible cuando lo he necesitado y sus estupendos comentarios sin los cuales esta tesis no sería lo que es. Han sido 4 años de ir al CBGP a gusto y era gracias a vosotros.

También agradecer al lab 285: Paul, Saray, Emilia, Chechu, Sandra, Andrea, Clara, Pablo y otros labs: Víctor, Patri, Andrea, Ana, Caixo, Elena (y alguno más que ya no está) por hacerme sentir uno más. Pese a haber pasado largas temporadas fuera volver sabiendo que estabais allí era siempre una alegría. Las comilonas en ciudad de la imagen, cenas de navidad, cerveceo, fiestas y salchichitas que siempre quedarán en el recuerdo. Espero que no sean la última.

A José, Juanma y Nacho, por los cenotes en el piso Coalición. A Jorge, por los buenos ratos y los consejos. A Javi y Carlos por el último año tan bueno en Guilileo, los últimos meses era un desahogo saber que estabais allí al volver.

Agradecer especialmente a mis hermanas Ana y Paula, por decirme siempre de bajar al Sánchez pese a que casi siempre digo que no, por ayudarme cuando que lo necesité, por esos ratos comiendo en la terraza, los bailes sorpresa en el salón y las risas yendo al gym. Estos 4 años no hubieran sido igual sin vosotras.

A Adriana, por ser siempre la primera en apoyarme y la primera en bajarme de las nubes, por toda la confianza y presión que ha hecho falta y por que sin ti probablemente esta tesis hubiera durado algún año que otro de más.

Y por último, a mis padres, por darme todo lo que necesité y más, por alentarme a seguir y porque sin vosotros no estaría donde estoy. Por mucho que os lo agradezca, nunca será suficiente, gracias. Esta tesis os la dedico a vosotros.

CONTENTS

ACKNOWLEDGEMENTS	VII
RESUMEN	XIII
SUMMARY	XV
1. INTRODUCTION	1
1.1. Overview	3
1.2. Natural Language Processing.....	4
1.3. Hedging	5
1.4. Certainty	8
1.5. Related Work and Prior Art	9
1.6. Basis.....	13
1.7. Citations & References	15
1.8. Hedging Erosion.....	16
1.9. Text Mining	17
1.10. Micropublications and Nanopublications	18
1.11. The emerging requirement for FAIR Data	19
1.12. Machine Learning (ML)	20
1.13. Ontologies	23
1.14. Identifying Reference Spans	24
2. DESCRIPTION OF THE RESEARCH PROBLEM	27
3. RESEARCH OBJECTIVES	31
3.1. Research Hypotheses.....	34
4. MATERIALS AND METHODS	37
4.1. Broad overview	39
4.2. Survey statement selection	39
4.2.1. Statements used	40
4.3. Survey design	47
4.4. Ranking study	53
4.5. Statistical Analysis of Survey Responses	54

4.6. Clustering.....	55
4.7. Identifying Reference Spans	56
4.8. Certainty Assignment of a New, Larger Corpus	57
4.9. Certainty Classification and Machine Learning Model	61
5. RESULTS	65
5.1. Certainty Surveys	67
5.2. Clustering of Statements by Survey Results	73
5.3. Comparison Between Surveys.....	79
5.4. Comparison Between Questions.....	82
5.5. Evaluating the respondents indication of ‘basis’	84
5.6. Basis and Certainty Correlation.....	86
5.7. Identifying Reference Spans	91
5.8. Machine Learning Model.....	91
5.9. Capturing Certainty in Formal Logics and Data Structures	95
5.10. Applying Automated Certainty Annotations	98
6. DISCUSSION	101
6.1. Evidence to support three levels of certainty in scholarly statements	103
6.2. The absence of a Low certainty category	106
6.3. Comparison between Questions and Surveys	107
6.4. Basis and Certainty Correlation	108
6.5. Machine learning	108
6.6. Application of this categorization system	109
6.7. Tools for researchers, authors, reviewers, and data miners.....	109
6.8. Future investigations to elucidate perceptions of certainty	111
7. CONCLUSIONS	113
8. REFERENCES	115
9. SUPPLEMENTAL INFORMATION	121
9.1. List of Tables.....	124
9.2. List of Figures.....	125

RESUMEN

Las estructuras gramaticales que los investigadores usan para expresar sus afirmaciones intentan transmitir diversos grados de certeza o especulación. Estudios anteriores han sugerido una variedad de sistemas de categorización para la certeza académica; sin embargo, estos no han sido validados objetivamente, particularmente con respecto a representar la interpretación del lector, en lugar de la intención del autor.

En esta tesis intentamos un enfoque de categorización de certeza basado en un modelo de datos. Ejecutamos una serie de estudios basados en cuestionarios utilizando frases académicas seleccionadas manualmente, en inglés, para determinar cómo los investigadores clasifican varias afirmaciones académicas, utilizando tres sistemas distintos de clasificación de certeza. Luego intentamos definir objetivamente las categorías de certeza percibida entre los lectores de textos biomédicos mediante el examen del grado de acuerdo/consistencia en la selección de categorías de certeza, por parte de los mismos lectores. Posteriormente se aplican pruebas estadísticas para evaluar el grado en que el sistema de categorización proporcionado en cada encuesta refleja la percepción de aquellos a quienes se les pidió que usaran esas categorías. El sistema de categorización con la puntuación más alta, es decir, el que proporcionó el nivel más alto de acuerdo, se usó para crear manualmente un gran corpus de declaraciones anotadas con su respectivo grado de certeza. Esto, a su vez, se utilizó para generar un modelo de "machine-learning" capaz de clasificar automáticamente nuevas declaraciones entre estas categorías, con alta precisión. Proponemos que este modelo podría usarse dentro de los algoritmos existentes de minería de texto para capturar metadatos adicionales que reflejen la expresión de certeza en el texto original. Además, proporcionamos un ejemplo de una publicación académica "machine-accessible", una Nanopublicación, en la que hemos incorporado estos nuevos metadatos de certeza contextual.

Descubrimos que los lectores perciben tres categorías de certeza: un nivel de certeza alta y dos niveles de certeza más baja que están menos diferenciados, pero que muestran un grado significativo de acuerdo entre lectores. Mostramos que estas categorías se pueden detectar de manera automatizada, utilizando un modelo de "machine-learning", con una precisión de "Cross-Validation" (CV) del 89,0% en relación a un "gold-standard" generado manualmente, y una precisión del 82,2% contra un corpus formado por las respuestas de los lectores. Este

hallazgo brinda la oportunidad de capturar metadatos contextuales relacionados con la certeza como parte de proyectos de minería de texto, que actualmente omiten estas sutiles claves lingüísticas. Proporcionamos como ejemplo un conjunto de nanopublicaciones “machine-accesible” que representan todas las declaraciones analizadas en esta tesis, donde la categoría de certeza asignada por nuestro modelo de aprendizaje automático está integrada como metadatos de manera formal con base ontológica como prueba de concepto.

SUMMARY

The grammatical structures scholars use to express their assertions are intended to convey various degrees of certainty or speculation. Prior studies have suggested a variety of categorization systems for scholarly certainty; however, these have not been objectively tested for their validity, particularly with respect to representing the interpretation of the reader, rather than the intention of the author.

In this thesis we attempt a data-driven certainty categorization approach. We execute a series of questionnaire-based studies using manually-curated scholarly assertions, in English, to determine how researchers classify various scholarly assertions, using three distinct certainty classification systems. We then attempt to objectively define categories of perceived certainty that are shared among readers of biomedical text by examining the degree of consistency in certainty category selection by these readers. Statistical tests are then applied to evaluate the degree to which the categorization system provided in each Survey reflects the perception of those asked to use those categories. The categorization system with the highest score - that is, the one that provided the highest level of agreement - was then used to manually create a large corpus of certainty-annotated statements. This, in turn, was used to generate a machine-learning model capable of automatically classifying new statements into these categories with high accuracy. We propose that this model could be used within existing text-mining algorithms to capture additional metadata reflecting the nuanced expression of certainty in the original text. Additionally, we provide an example of a machine-accessible scholarly publication - a NanoPublication - within which we have embedded this novel contextual certainty metadata.

We found that there are three categories of certainty perceived by readers: one level of high certainty, and two levels of lower certainty that are somewhat less distinct, but nevertheless show a significant degree of inter-annotator agreement. We show that these categories can be detected in an automated manner, using a machine learning model, with a cross-validation (CV) accuracy of 89,0% relative to a manually-generated “gold standard”, and 82,2% accuracy against a publicly-annotated corpus. This finding provides an opportunity for contextual metadata related to certainty to be captured as a part of text-mining pipelines, which currently miss these subtle linguistic cues. We provide a set of exemplar machine-accessible Nanopublications representing all statements analyzed in this thesis,

where the certainty category assigned by our machine learning model is embedded as metadata in a formal, ontology-based manner as proof-of-concept.

1. INTRODUCTION

GENERAL INTRODUCTION

Scientific progress requires the ability to build on the knowledge of others, and extend this to new discoveries. This, therefore, requires a high fidelity in two processes - the accurate representation of the certainty of a scholarly assertion by an author, and the accurate citation of that certainty by a downstream user. If either of these fails, subsequent work will potentially be biased and/or non-reproducible. The volume of articles that are demonstrably non-reproducible has recently been shown to be strikingly large [1], and a foundational study indicated that this number might exceed 50% of the published literature [2]. Beyond the inability to validate the results reported in the study itself, there are additional negative repercussions when the results of (potentially) flawed studies become part of the scientific literature. Defective data and axioms from this literature become incorporated into databases - either by direct deposit, manual curation, or machine-extraction - thus contaminating the knowledge and data used in follow-up research. This feedback-loop is resulting in a widely-acknowledged “crisis” in science (e.g. <https://phys.org/news/2017-03-science-crisis.html>). Of particular interest to us in this thesis is that the mere existence of this crisis reveals that another cornerstone of the scientific process - peer-review - must *itself* be experiencing a crisis of insufficiency, since this is intended to be the “gatekeeper” against such contamination in the scholarly archive. We suggest that this may be due to an insufficiency in the tooling that supports rigorous peer-review in an era with such a large volume of highly domain-specific literature.

We will first provide a high-level overview of the topics and concepts that will be covered in this thesis, in order to better contextualize the problems and open questions that are addressed by this work. We will then go deeper into the existing literature about these same topics, comparing and contrasting prior art in order to justify the questions we pose, the structure of our experiments, and the methodologies used to interpret the results, and how we then applied them in a practical way to address clear gaps in the existing scholarly knowledge infrastructure.

1.1. Overview

Citations are the means by which scholars refer to the knowledge presented in previous articles. A recent study revealed an alarming anomaly in this pillar of the scholarly process [3], where the re-use of scholarly assertions in citations, which are intended to support novel hypotheses or interpretations of results, were shown to “drift” in their intensity or

their meaning compared to what was originally stated in the cited material. Subjective interpretations when authors read an article become infused into their citations, changing the qualitative strength compared to the original statement (generally towards higher certainty). Within citation chains, this phenomenon may amplify, sometimes resulting in near-factual statements which, at the origin, were far more speculative. All of this happens without necessitating any additional evidence supporting the original claim. The outcome is a cycle of dubious scholarly assertions providing the basis for new experiments, where this lack of transparency is not detected by peer-review and thus leads to further contamination of the scholarly literature. This happens, at least in part, due to a lack of resources - at all steps along the path - capable of detecting these problems.

Researchers (by and large) use natural language as the means by which they share their knowledge, generally in the form of a narrative publication. Natural Language is an intricate, but effectual system for human communications, allowing a wide range of subtle nuances to be expressed through the words chosen, or the grammatical structures employed. Computers, on the other hand, do not succeed as well in their comprehension of human language. These nuances are an obstacle for computers' accurate capture of the subtler meanings being conveyed in a sentence, and continue to be a problem in the field of Natural Language Processing (NLP) [4].

1.2. Natural Language Processing

Scientific communication exhibits particular characteristics that differentiate it from normal communication, such as the syntactic structure, passive, extraposition or the use of hedging language[5] (discussed in more detail later). Moreover, increasingly specialized knowledge (and its associated language) is required to understand new specialty domains that arise on a regular basis. Biomedicine is a branch of science that investigates the emergence of, and perturbation of, life processes, and the consequences of these. This domain has its own linguistic characteristics [6] that require the specialization of tools or workflows in order to automatically detect qualities and knowledge contained in its associated texts. Importantly, the number of publications in Biomedicine grows at an ever-increasing rate; for a researcher to remain up-to-date with this literature is largely impossible, even in a specific sub-domain. For example, the volume of scholarly articles published every year, just in biomedicine, has doubled in the last two decades, from ~

300,000 articles in 1996 to more than 800,000 in 2016¹. As a result, it is increasingly necessary to provide effective, efficient, accurate and automatic means of capturing and distributing the contained knowledge, in its variety of forms (e.g. textual assertions, graphs/figures, and the text contained in them). With this in-mind, the observation that scientists' personal biases are increasingly being reflected in scholarly publications [7],[8] is troubling, as there is a breakdown in fidelity between the creation of a potentially biased scholarly assertion, the computational extraction of that assertion including its subtle bias, and the final consumption and interpretation of that extracted knowledge by another human. In response, a wide range of initiatives related to NLP have originated in recent years that attempt to improve the extraction of knowledge from scholarly text, such as detecting textual information in images [9], recognition of reference scope [10] or automatic assignment of perceived certainty [11]. These latter two are of particular relevance to this thesis. Recognition of the scope of reference article² - that is, detection of the original span of text/media that is referred-to by a downstream citation - and sentiment analysis and opinion mining³ - that is, detection of the "mood", sentiment, or strength with which a statement is made. The former is important in order to examine the fidelity with which the subtle linguistic cues are being maintained throughout a citation chain. The latter is important for determining the relevance of these subtle cues, and how they should affect our interpretation of the statement. Among the most common confounders of accurate capture of the certainty with which a scholarly assertion is made is the widespread use of "*hedging*" in scholarly writing.

1.3. Hedging

Trainee researchers learn the rules of the expression, argumentation and scholarly writing through their careers. These are enforced by their mentors, as well as journals and funding agencies. One of the fundamental characteristics of scholarly discourse is the use of *hedging* language.

"Hedging may present the true state of the writers' understanding, namely, the strongest claim a careful researcher can make" [12].

¹ https://www.nlm.nih.gov/bsd/stats/cit_added.html

² <http://wing.comp.nus.edu.sg/~cl-scisumm2018/>

³ <http://www.conll.org/2019>

“Hedging is the expression of tentativeness and possibility and it is central to academic writing where the need to present unproven propositions with caution and precision is essential.” [13].

Since the 1970s, hedging has been studied in the domain of formal semantics [14]. Because we do not deeply address the formal semantics of hedging grammar in this thesis, we will only provide the shallowest summary of previous studies on hedging and its perspective in scholarly texts.

A review of the history of hedging was undertaken by our collaborator Anita De Waard (De Waard, unpublished) and we briefly summarize that history here. The first author to describe 'hedge' was George Lakoff [14], hedge was described as "words whose job is to make things fuzzier or less fuzzy". In his work, Lakoff analyzed how words contained in a sentence, such as *“sort of”* or *“mostly”*, may modify the semantic meaning of propositions and/or the expression of an event. Skelton was one of the first authors to suggest the existence of different levels of certainty, namely “information that is taken for granted” (i.e. facts), the “merely hypothetical”, and “logical deduction” [15]. For this, he analyzed the content of 40 articles in the “hard” sciences and humanities, examining text related to hypotheses, probabilities and evaluations, since these are the portions of a scholarly narrative that are more likely to contain the reasoning of the claim. Myers [16] performed a study on 50 articles in the domain of molecular genetics, where he identified several distinct hedging patterns being used in scholarly writing. Salager-Meyer [12] analyzed various structural components of 15 medical articles (e.g. Introduction versus Results versus Discussion), to investigate if the communicative purpose influenced the frequency and category distribution of hedging language. He concluded that the *Discussion/Comment* sections included most of the hedging phrases. Subsequently, Hyland [13] analyzed the reasons for using hedging, concluding that in the absence of the “social context” of the scholarly community, and if we were simply communicating factual statements, there would be no need to utilize this (odd) linguistic tool; however, in scholarly writing, it is used as a tool of persuasion, in the absence of fact. More recent studies, Hashemi [17] compared the use of hedging in the discussion section for 150 applied linguistics articles (50 qualitative, 50 quantitative, and 50 mixed methods studies). The result showed that quantitative applied linguistics articles more frequently employed the use of hedging and the most frequent categories of hedging were full verbs, auxiliaries, and adverbs.

The use of hedging language should not be perceived in any negative sense. Good reasons to use hedging are: Inability to reach assurance that a hypothesis is true; to provide ambiguity in order to avoid possible future contradictions; and to be objective, avoiding the injection of personal opinions [12]. Using hedging, the researcher presents their interpretation of the significance of the main outcomes from the study, helping the reader understand the fundamental elements of the information offered, while protecting themselves (and the reader) from over-extension of the interpretation or application of that discovered knowledge [13].

Most research is grounded in the data, evidence, interpretation and knowledge of prior studies. New investigations utilize citations of older investigations to show that their hypotheses are properly grounded and well-researched⁴; to give credibility to their research; and to show the motivation and certainty about the claims under which it was consolidated. This is where the use of hedging language may lead to a loss of fidelity in the transmission of scholarly knowledge. Subjectivity in the interpretation of the words used to express the knowledge in previous studies may, consciously or unconsciously, bias our interpretation of prior statements (generally “in our favour”) even to the point of modifying the idea expressed in the original statement [18][19]. The question then arises: *can this phenomenon be detected - preferably in an automated way?* To do so would require the ability to assess the “truth value” of a proposition. Such evaluations would utilize *epistemic modifiers* - words that modify the judgment/assessment of a statement, giving value to, for example, the degree of confidence being expressed. Hedging is the linguistic pattern that specifically utilizes such epistemic modifiers to qualify the certainty of the claim being made. A simple example would be:

*Gene X **could be** induced by gene Y.*

This hedging statement uses epistemic modifiers to modulate the degree of certainty about the assertion of regulation between X and Y. Other examples of commonly used epistemic modifiers are: *suggest, show, implicate, hypothesize* or *demonstrate*. For example:

⁴ <https://libguides.mit.edu/citing>

*“These results **suggest that** the APC is constitutively associated with the cyclin D1/CDK4 complex and **are consistent with** a model in which the APC is responsible for cyclin D1 proteolysis in response to IR...” [20].*

The variety of ways of expressing approximately the same idea lead to a high variability in grammatical structures that relay the confidence or certainty of the statement. This leads to key questions: *Does the reader perceive these differences, and if so, with what granularity? Do all readers perceive these statements in the same way, or does hedging result in a breakdown in the fidelity of knowledge-transmission?*

1.4. Certainty

Certainty could be defined in many ways. At its most extreme, it would be expressed as “*is in no doubt*”⁵. Certainty, in relation with scholarly articles and research fields, in this thesis is described as: “*the perception of the reader regarding the certainty intended to be conveyed by the author*”. Note that this definition does not attempt to measure the intent of the author; rather, only the perception of the reader. The reasons for this are multifold, including: The relative scarcity of authors relative to readers is relevant to the statistical significance that can be derived from a study; the relative difficulty in finding/recruiting authors vs. readers; their likely inability to remember their intent at the time-of-writing; and the inability to ask an author questions about their intent in the absence of a certainty-categorization system (i.e. in the absence of a standard metric, the only possible answer to a question about what “level of certainty” they meant to convey, would be “I intended to say exactly what I said”).

Without doubt, the phraseology used by an author implicitly indicates the author’s intent; i.e. “*[e]ach [sentence] fragment conveys a degree of certainty about the validity of the assertion it makes*” [21].

⁵ <https://dictionary.cambridge.org/es/diccionario/ingles/certainty>

1.5. Related Work and Prior Art

A number of prior studies have attempted to categorize and capture the expression of scholarly certainty. These, and other certainty categorization studies, are summarized, compared and contrasted in Table 1, where the columns represent relevant study features that distinguish these various investigations, and affect the interpretation of their outcomes. For example, the use of linguistic experts versus biomedical domain experts will likely affect the quality of the annotations, while using explicit rule-matching/guidelines will result in strict, predetermined categorizations. Similarly, the use of abstracts consisting of concise reporting language, versus full text which contains more exploratory narratives, will affect the kinds of statements in the corpus [22], and their degree of certainty.

Table 1: Comparison of corpora and approaches used in prior investigations into scholarly certainty

Citation	Nº of annotators	Annotator expertise	Text provenance	Discourse segment source	Approach to automated detection	Number of certainty classification classes	Corpus Size	Meta knowledge examined
[24]	4	following annotation guidelines	Medline	Abstract	SVM	3	2,093 statements	certainty
[25]	3	following annotation guidelines	Medline	Abstract	Maximum Entropy	4	350 abstracts	certainty
[26]	7+2	biomedical	GENIA-MK, BioNLP-ST	Abstract, Text Event	Random Forest classifier + Rule Induction	2/5	652 passages	certainty
[3]	2	publishing	2 articles [27], [28]	Full text	N/A	4	812 clauses	certainty, basis, source
[29]	3	physics	Columbia Presbyterian Medical Database	Free text	Natural Language Processor	4	230 reports	certainty, degree, change, status, quantity, descriptor
[21]	3+9	Expertise ,following annotation guidelines	Ten research articles published in 2005	Full text	N/A	4	101 sentences	focus, polarity, certainty, evidence, and directionality
[30]	3	linguistics	Clinical, FlyBase, BMC Bioinfo.	Free Text, Full Text, Abstract	N/A	2	20,924 statements	certainty, negation
[31]	2	following annotation guidelines	Medline	Abstract	N/A	3	36,858 events	manner, source, polarity, certainty, knowledge type
This Thesis	375	Biomedicine	TAC 2014	Full Text	Neural Network	3	45 statements	Certainty, basis

According to Wilbur et al., and as noted above “each [statement] fragment conveys a degree of certainty about the validity of the assertion it makes” [21]. While intuitively correct, it is not clear if certainty can be measured/quantified, if these quantities can be categorized or if they are more continuous, and moreover, if the perception of the degree of certainty is shared between readers. Most studies in this domain assume that certainty can be measured and categorized, though they differ in the number of degrees or categories that are believed to exist, and thus there is no generally-accepted standard for certainty/confidence levels in biomedical text [23].

The study of “certainty” - also referred-to as epistemic modality - is a specific domain within the larger study of linguistics, focusing on the confidence of a phrase’s commitment to its topic. [32] [33]. One of the first authors to describe different levels of certainty was J. Holmes [34]. Who explains the epistemic modality as a judgment of the author about the veracity of an action, which he classified into three levels: Certain, probable and possible, in order of decreasing certainty respectively.

Additionally, Skelton [15] described three types of certainty that exist in scholarly inquiry: *facts* (absolute certainty), the *merely hypothetical* (used in an exploratory discussion) and *logical deduction* (*a posteriori* certainty associated with experimental results). More recent authors suggest a more granular partition. Wilbur et al suggested a four-category classification: complete uncertainty, low certainty, high likelihood, and complete certainty/proven fact. Similarly, Friedman et al. [29] suggests that there are four categories of certainty: no certainty, low, moderate, and high certainty, with an additional “cannot evaluate” category. Thompson [35] also selected four levels of certainty, introducing the idea of predicting levels of certainty based on the presence of some lexical items and contextual information. De Waard and Schneider [36], aligning with three of these latter studies, also encoded four categories of certainty into their Ontology of Reasoning, Certainty, and Attribution (ORCA) as follows: Lack of knowledge, Hypothetical (low certainty), Dubitative (higher, but short of full certainty), Doxastic (complete certainty, accepted knowledge or fact). Specific examples of these four categories were proposed, as follows:

Doxastic: “We have previously **demonstrated** that accumulation of A β PP epitopes precedes other abnormalities in IBM muscle fibers” [37]

Dubitative: “Overexpression of A β precursor protein (A β PP) into mature cultured human muscle fibers induces in them several aspects of the IBM

phenotype (6-8), **strongly suggesting** that A β PP, and its toxic proteolytic product A β , play an upstream role in the s-IBM pathogenic cascade.” [38]

Hypothetical: “These miRNAs neutralize p53- mediated CDK inhibition, **possibly** through direct inhibition of the expression of the tumor suppressor LATS2.” [39]

Lack of Knowledge: “The stimulus for excessive A β production in IBM is **unknown**, and whether this precedes inflammation, or vice versa, remains to be determined [10].” [40]

These four statements exemplify the “gradient” - from highest to lowest certainty - proposed by De Waard and Schneider. The first, **Doxastic**, presents a fact with absolute certainty. A **Dubitative** statement might use the adverb “*strongly*” to strengthen the speculation “*suggests*”. A **Hypothetical** statement theorizes about the possibility of an event, lacking assuredness. Finally, when information is unknown, it may be clearly stated as **Lack of Knowledge**.

Other studies have suggested fewer or more certainty categories, and differ in the manner in which these categories are applied to statements:

BioScope [30] is a manually-curated corpus, containing 20,924 speculative and negative statements from three sources (clinical free-texts, five articles from FlyBase and four articles from BMC Bioinformatics) and three different types of text (Clinical reports, Full text articles and abstracts). Two independent annotators and a chief linguistic annotator classified text spans as being ‘speculative’ or ‘negative’; other kinds of assertions were disregarded. Thus, the study splits certainty into two categories - speculative, or not

Thompson et al. [31] apply five meta-knowledge features - manner, source, polarity, certainty, and knowledge type - to the GENIA event corpus [41]. This corpus is composed of Medline abstracts split into individual sentences. For certainty annotations, the corpus utilizes a classification system of three certainty levels - certain, probable (some degree of speculation), and doubtful (currently under investigation). Annotation was carried out by two linguistic specialists specifically trained in the meta-knowledge scheme.

Light et al. [24] investigate speculative language in biomedical abstracts. Using Medline abstracts they attempt to distinguish high and low degrees of speculation. Four annotators

used rule-matching to classify statements. Using this annotated corpus, they trained a model based on Support Vector Machines (SVM) to generate an automatic classifier. This automatic classifier, therefore, is specifically tasked for speculative statements, and categorizes them in a manner resembling their predefined rule-sets.

Malhotra et al. [25] classify hypotheses (speculative statements) in scholarly text. Three annotators classified speculative statements in Medline abstracts related to Alzheimer's disease using a four-class categorization, with predefined pattern-matching rules for sorting statements into three speculative patterns (strong, moderate, weak) and a fourth category representing definitive statements. Additionally, they explore several automated methods to distinguish speculative from non-speculative statements.

Zerva et al. [26] use a combination of the BioNLP-ST and GENIA-MK corpora - both of which consist of statements manually-annotated with respect to their certain/uncertain classification (degrees of uncertainty, when available, were merged resulting in a two-category corpus). They applied rule induction combined with a Random Forest classifier to create an automated binary classification model. This model was run on 260 novel statements, and the output classification was provided to seven annotators who were asked for simple agree/disagree validation of each automated classification. The degree of disagreement between annotators was in some cases surprisingly high, leading the authors to note that “the perception of (un)certainty can vary among users”. In a separate experiment, two annotators ranked the certainty of 100 statements on a scale of 1-5. They noted low absolute annotator agreement (only 43% at the statement-level), but high relative agreement (only 8% of statements were separated by more than one point on the five-point scale). Comparing again to the automated annotations, they found high correlation at the extremes (scored by the annotators as one or five) but much less correlation for statements rated at an intermediate level, leading them to conclude “...looking into finer-grained quantification of (un)certainty would be a worthwhile goal for future work”.

These previous works share important distinctions relevant to the current investigation. First, in every case, the number of certainty categories were predetermined, and in many cases, categorization rules were manually created. Second, in most cases, the work involved a small number of annotators with a knowledge of linguistics, or specifically trained on the annotation system, rather than experts in the knowledge domain being represented by the statements, but untrained as annotators. Third, in all cases where automated approaches were introduced, the automated task was to distinguish “speculation” from “non-

speculation”, rather than categorize degrees of certainty. Notably, there was little agreement on the number of categories, nor the labels for these categories, among these studies. Moreover, the categories themselves were generally not validated against the interpretation of an (untrained) domain-expert reader. As such, it is difficult to know which, if any, of these approaches could be generalized to annotation of certainty within the broader scholarly literature, in a manner that reflects how domain experts interpret these texts.

To achieve this would require several steps: 1) determine if there are clearly delimited categories of certainty that are perceived by readers of scholarly assertions; 2) if so, determine how many such categories exist; and 3) determine the fidelity of the transmission of certainty among independent readers (i.e. agreement). If these are determined robustly, it should then be possible to apply machine-learning to the problem of automatically assigning certainty annotations to scholarly statements that would match the perceptions of human readers.

The linguistic expression of certainty may have a wide range of detectable features which, when combined with other contextual information (e.g., type of speech segment, content summary, embedded words, etc.) might help in developing a classification tool. In fact, in addition to the value of epistemic statements, other authors attribute the foundation or basis of the statement (e.g., based on reasoning, based on data or direct evidence), as an important, quantifiable trait (De Waard et al., 2012).

1.6. Basis

We use the term “basis” to describe the foundation (implicit or explicit) upon which an author made a scholarly statement. Although NLP tools have achieved a breakthrough in data processing, they still do not capture or classify nuances only perceptible to expert reviewers. These nuances (certainty or basis) help to understand the text; as such, so long as we cannot correctly identify them and the relationships between them, we will not be able to produce accurate NLP tools. The evidence or basis on which a statement is founded might be indicated by the modality [42] [35] or in our case epistemic modality (author's degree of certainty about a statement). According to Wiess [43] the degree of certainty associated with individual objects depends on the type of basis and approach employed. Therefore, in order to correctly interpret the statements, the accurate identification of modal information is fundamental [35].

Teufel built an annotation scheme for scientific articles, with one category as “basis”. In basis, a primary distinction of the evidence "statement from own work" was made: YES (there is evidence) or NO (no evidence) [44]. Palmer classified the evidence of the statements in four categories: reported (citation), sensory (about sensation or speculation), direct and indirect evidence (reasoning) [45]. Wilbur performed a four-way progressive classification scheme: "no evidence", "no verifying information", "citation" and direct "evidence provided" in the paper [21]. Thompson’s modality scheme encompasses other components in addition to the evidence. "Knowledge type", encompasses whether a statement is speculative or based on evidence. "Point of view", encompasses whether the statement is based on the author's own work or refers to an external author (citation) and "level of certainty" that does not include any type of evidence [35]. Finally, De Waard and Pander proposed a model with three concepts. "Value" (referencing the level of certainty), "source" of the knowledge and "basis" where a clear distinction is made between "reasoning" (argumentation or speculation), "data" (direct reference to the results) and "implicit" or absent (unclear attribution) [3].

Our classification model for basis / evidence contains four differentiated classes:

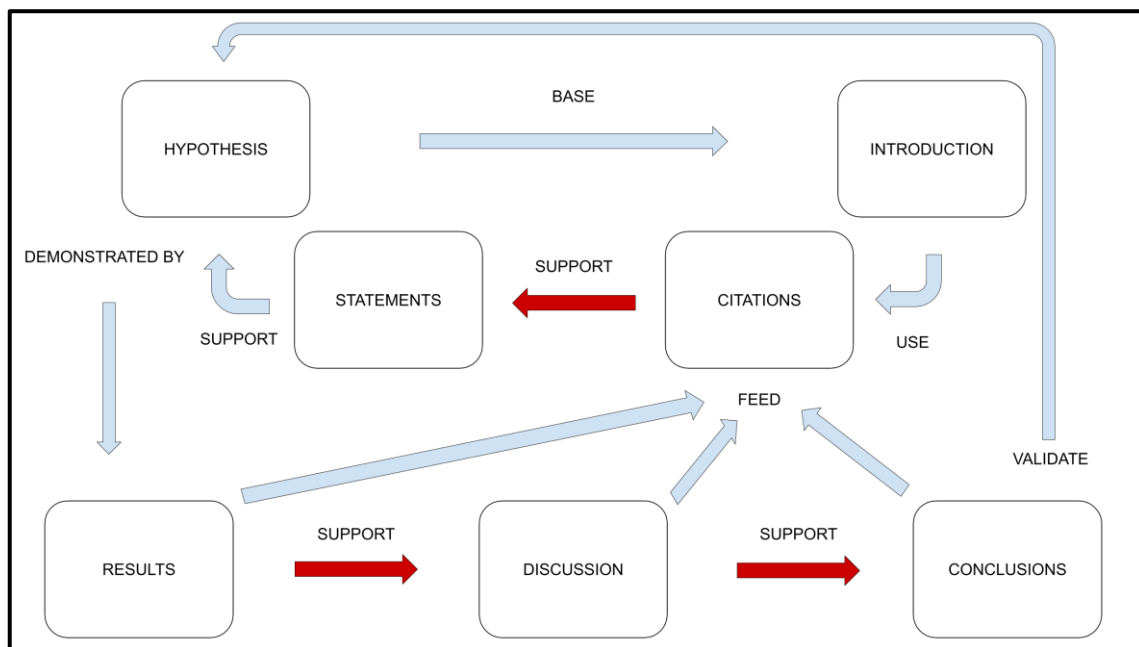
- Direct Evidence: Empirical evidence from an experiment, without necessary reasoning. *e.g., The screen of my computer is black when switched off - I know this because I switched it off and it became black.*
- Indirect Evidence / Reasoning: When the result comes from secondary experiments or is reasoning from a direct result or prior knowledge. *e.g., I measured the temperature of the screen, and it was cold. Screens are warm when they are switched on, so my screen must be off. Therefore it must also be black.*
- Speculation: There is no substantiated evidence or it is not based on the data. *e.g., I propose that the screen may be black.*
- Citation: Attribution to other authors, can be direct, absent or self-citation. *e.g., Mario's computer screen is black [Prieto et al. 2014].*

Recognizing the basis/evidence of a statement gives us a greater understanding of the text we are reading. In addition, a recognition of the basis most likely influences in subsequent citations and references by other authors.

1.7. Citations & References

“Rhetoric”, in the context of scientific argumentation, relates to the varying ways of expressing acquired or discovered ideas and knowledge that influence the reader to believe that what they read is the truth. It is intended to convince the scientific community, with hypotheses, arguments, and the results obtained through good scientific method, that the work presented follows the canon of best practice⁶. This leads to the reasoning and argumentation structure that appears almost ubiquitously in the typical discourse of scholarly publications: Introduction, Materials and Methods, Results, Discussion, Conclusion.

Indeed, convincing others of a work’s conclusion is the main purpose of the scientific article [46]. Beyond basic academic disclosure, the rhetorical structure of the scientific publication - a model that aims to convince - has a primary role in the dissemination of knowledge. This structure has remained largely unchanged for hundreds of years, despite the more contemporary Machine Learning and Linked Data approaches to complement the capture and publication of scientific knowledge in a manner amenable to direct utilization by a machine [47]. The rhetorical structure of a scholarly paper mirrors, and supports, the flow of scholarly discourse shown in Figure 1.



⁶ http://www.rhetinfo.com/uploads/7/0/4/3/7043482/philosophy_and_rhetoric_of_science.pdf#page=11

Figure 1: The Flow of Scholarly Discourse. A hypothesis is the base of the introduction section, which use citations to support the validity of the hypothesis. The hypothesis is demonstrated by results. Results are used to support discussion and the conclusions, which validate the primary hypothesis. Results, discussion and conclusions feed citations in new articles.

Within this argumentative structure there are two components that are particularly interesting for this thesis. Those are: statements supported via citations (to prior literature); and statements (conclusions) that are supported by new data [48]. We highlight these (red arrows) because they represent the two cases where the level of expressed certainty about a given claim, over a citation chain, might be expected to change. First, the change in certainty may indicate the existence of new data supporting (or refuting) the claim. Thus, for a researcher looking for the core evidence backing a claim, or for an automated data collection/mining workflow, the ability to detect these kinds of “certainty inflection-points” would be of high utility, as they would be indicative of the presence of associated relevant datasets. This is an entirely valid reason for the level of certainty to change through a citation chain. An invalid reason, however, is termed “hedging erosion”.

1.8. Hedging Erosion

With hedging erosion, the change in certainty is caused by the misinterpretation or misrepresentation of a prior statement (knowingly or otherwise). De Waard [3] has recently shown, through manual reconstruction of a citation chain (Figure 2), that this phenomenon does in fact occur, and most importantly, occurs *in the absence of any additional evidence* [49]. Greenberg [19] also demonstrated a process called “Citation transmutation” where, through references, a claim is transformed from hypothesis into a fact.

Looking closely at Figure 2, it is clear that the transformation from low-certainty to higher certainty, or even to fact, may be incremental and difficult to perceive from one step to another. This leads to recognition of the importance of another problem, identified by Gilbert [54], which is the tendency not to cite the original source of a claim, but rather to cite only the most recent citation of that claim. This, ultimately, generates unfounded authority [19]. The more recent statements, generally harboring greater certainty, will often be used to establish new hypotheses, or experiments. This would be a useful scenario to be able to detect during, for example, the process of peer review or hypothesis generation as a means of validating an experimental design. Peer-review is intended to protect against

"hedging-erosion", however, peer-review is tedious and increasingly specialized work, and there is little tooling to support peer-reviewers. Moreover, over-specialization (of both researchers and their peer-reviewers) may result in the reviewer lacking the specific domain knowledge related to the legacy of a certain scholarly claim. Even if they were to check the citation, if that citation is not to the original source, the "erosion" in the claim may not be perceptible to them. This same problem is worsened in the context of text mining algorithms that will generally be unable to capture the nuances of scholarly assertions when extracting the entity-relationships that make up the claims.

How a claim becomes a fact
<i>"These miRNAs neutralize p53- mediated CDK inhibition, possibly through direct inhibition of the expression of the tumor suppressor LATS2."</i> [39]
<i>"In a genetic screen, miR-372 and miR-373 were found to allow proliferation of primary human cells that express oncogenic RAS and active p53, possibly by inhibiting the tumor suppressor LATS2 (Voorhoeve et al., 2006)."</i> [50]
<i>"[On the other hand,] two miRNAs, miRNA-372 and -373, function as potential novel oncogenes in testicular germ cell tumors by inhibition of LATS2 expression, which suggests that Lats2 is an important tumor suppressor (Voorhoeve et al., 2006)."</i> [51]
<i>"Two oncogenic miRNAs, miR-372 and miR-373, directly inhibit the expression of Lats2, thereby allowing tumorigenic growth in the presence of p53 (Voorhoeve et al., 2006)."</i> [52]

Figure 2: How a claim becomes a fact. These sentences represent a series of scholarly assertions about the same biological phenomenon, revealing that the core assertion transforms from a hedging sentence into statements resembling fact through several steps, but without additional evidence [53].

1.9. Text Mining

As mentioned above, given that the volume of literature published grows by approximately a 3.5% per year [55], text mining is becoming an increasingly important way to capture new knowledge in a searchable and machine-accessible way.

Text mining is a productive task to extract and analyze texts, in order to gain information [56].

"The goal of biomedical text mining is... to allow researchers to identify needed information more efficiently, uncover relationships obscured by the sheer

volume of available information, and in general shift the burden of information overload from the researcher to the computer by applying algorithmic, statistical and data management methods to the vast amount of biomedical knowledge that exists in the literature as well as the free text fields of biomedical databases.”[57].

Clearly, given the discussion of certainty, hedging, and hedging erosion discussed above, accurate, automated knowledge capture necessarily requires accurate capture of the certainty with which the claim is being expressed, and the ability to qualitatively or quantitatively compare it to similar statements, possibly over a citation chain. Equally importantly, there is increasing pressure to publish knowledge, *ab initio*, explicitly for machines, in particular with the widespread adoption of the FAIR Data Principles for scholarly publishing [58], and with several machine-accessible knowledge publication formats having recently been suggested, including NanoPublications[59], and Micropublications[60]. In order to capture the intent of the author in these *ab initio* machine-readable publications, it will be necessary for them to include formal machine-readable annotations of the degree of certainty behind that claim.

1.10. Micropublications and Nanopublications

Micropublication: a framework for publishing scholarly discourse in a manner that is accessible to mechanized exploration. A micropublication may contain data, or a narrative statement of any kind [60].

The minimal components of a micropublication are the statement and its attribution. In addition, metadata (in the form of narrative text) could/may be included to support the argument, and its association-type specified (e.g. “supports” or “refutes”). Micropublications are generally represented using the World Wide Web Consortium’s Resource Description Framework (RDF) - a formal framework for data and knowledge representation.

Nanopublication: “[a] Nanopublication is the smallest unit of publishable information: an assertion about anything that can be uniquely identified and attributed to its author.” [61]. Nanopublications have three essential elements:

- **Assertion:** The smallest unit of information that is sufficient to convey a piece of knowledge, encoded in RDF (as opposed to natural language, as in

Micropublications). This may be a single subject/predicate/object “triple” (Mosquitoes Transmit Malaria), or it may need to be further contextualized in order to be complete (Mosquitoes Harbor Plasmodium; Plasmodium Causes Malaria).

- **Provenance:** Any kind of metadata about the assertion. This includes, for example, its supporting DOIs, its authorship, the date the assertion was made, etc.
- **Publication Info:** Any kind of metadata about the nanopublication itself.

As with micropublications, Nanopublications are represented using RDF and ontologies, allowing universal interoperability [61]. Nanopublications are one of the formats commonly selected for representing data according to the FAIR Data Principles.

1.11. The emerging requirement for FAIR Data

Open science is becoming a core philosophy of agencies and publishers globally. Both the EC and the G7 Nations [62] explicitly prioritized the reusability of scholarly data in 2016, and in the same year, the FAIR Data Principles were published [58]. These Principles - speaking to the Findability, Accessibility, Interoperability, and Reusability of scholarly digital objects - have rapidly become the hallmark for open science worldwide, being adopted by the USA-based Big Data to Knowledge (BD2K) of the National Institutes of Health[63], Science Europe[64], and the G20[65].

“FAIRness” is specifically aimed at improving the automated discovery and appropriate reuse of data by machines. As such, machine-first formats such as NanoPublications, and the rich structured metadata they are capable of carrying related to scholarly claims, will increasingly become *de facto* requirements. In parallel, therefore, the requirement to capture scholarly hedging in a formal, machine-readable way, will also become a requirement. *There is, therefore, an urgent need to formally determine what scholarly certainty “is”, how it is perceived, if it can be detected, if it can be categorized, and how it can be represented in a machine-accessible way.*

Machine-accessible data is the focus of FAIR because it is only through enabling machines to find and accurately process data autonomously that we will achieve the deeper goals of a world increasingly turning to Artificial Intelligence and machine-learning to solve critical problems in all fields. Beyond its use in the life and health sciences, there have been calls

for *ab initio* machine-readable data even from the domain of Securities where AI/Machine Learning approaches are being used, for example, for rapid detection of fraud:

“The success of today’s new technology depends on the machine readability of decision-relevant information. And I don’t mean just for numerical data, but for all types of information. This includes narrative disclosures and analyses found in the written word. It also includes contextual information about the information, or data about the data, often referred to as “metadata.” Today’s advanced machine learning methods are able to draw incredibly valuable insights from these types of information, but only when it is made available in formats that allow for large-scale ingestion in a timely and efficient manner.”[66]

1.12. Machine Learning (ML)

“Machine learning (ML) is the study of computer algorithms capable of learning to improve their performance of a task on the basis of their own previous experience” [67].

Automated recognition of certainty in scholarly articles has gained importance in recent years [25] and together with other ML tasks is becoming an essential component in science. Process automation is one of the basic elements of ML, it reduces the time used to perform difficult tasks and minimizes the margin of error, removing "human bias". In addition, through statistical inference it helps to discover and address problems that had passed unnoticed in existing data. [68]. Artificial intelligence algorithms can be classified according to their purpose.

Supervised: Algorithms that need a training set and a label manually curated by humans. Subsequently, the model created finds patterns in the data, makes a prediction and adds that label to new data. There are two types of supervised learning.

- **Regression:** The predicted result is a continuous value. It is usually used to identify a linear relationship between variables. Such as, Linear Regression (e.g. level of certainty for a statement and the level of similarity between the responses of two participants.)

- **Classification:** The predicted result is a category, a discrete label. Such as, Long Short-Term Memory, an artificial neural network architecture (e.g. assign a level of certainty to new statements)⁷.

Unsupervised: Algorithms that do not require labelled data. An unsupervised algorithm searches for "features" or patterns in the data, then clusters them according to those characteristics. Since the data does not need to be pre-classified, this is particularly appropriate for cases where the starting hypothesis is not known⁸.

Numerous machine learning algorithms have been described in the literature. Thus, we will only present those that were employed in this work for the detection of certainty and its basis. These are: Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), and Neural Networks (NN).

Principal Component Analysis

Many datasets have high dimensionality - that is, they contain many distinct observations or variables. It is common, however, to find that multiple variables are distinct reflections of the same underlying cause, simply reflecting the same parameter in a different way. It is therefore useful to reduce the number of variables, but to do so without losing the genuine variability in the data. PCA is an unsupervised machine learning algorithm that reduces the number of initial variables (d) by creating uncorrelated and linear combinations between them. These are ordered from higher to lower variability, resulting in final variables $p < d$ [69].

It can be simplified to six steps:

1. Compute the mean for every variable (d).
2. Compute the covariance, the result would be a $cov = (d \times d)$ dimensional matrix.
3. Compute the eigenvectors and eigenvalues of the covariance matrix.
4. Sort the eigenvectors from the highest to lowest eigenvalues.
5. Adjusting to new data. Here we apply the formula:

$$Finaldata = RowFeatureVector \times RowDataAdjust.$$

*"where **Row Feature Vector** is the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in the rows, with the most significant*

⁷ <https://www.analyticslane.com/2018/11/26/diferencias-entre-regresion-y-clasificacion-en-aprendizaje-automatico/>
⁸ <https://medium.com/@juanzambrano/aprendizaje-supervisado-o-no-supervisado-39ccf1fd6e7b>

*eigenvector at the top, and **Row Data Adjust** is the mean-adjusted data transposed” [70].*

Hierarchical Clustering Analysis (HCA)

Sometimes it is necessary to know how different variables of independent groups or populations are related and the distance or hierarchy between them. HCA (agglomerative clustering) is an unsupervised machine learning algorithm that attempt to build a cluster hierarchy. Each sample starts in its own cluster and consecutively they are combined with other samples creating larger clusters. It is normally represented by a tree where leaves are the samples and the root is a single cluster gathering all samples inside⁹. In this thesis, to measure the distance between clusters, we employed the method ‘average’, where distance is calculated as the average of all the points in a cluster to all points in the other cluster [71].

$$d(u, v) = \sum_{ij} \frac{d(u[i], v[j])}{(|u| * |v|)}$$

10

Neural Networks (NNs)

Neural networks are supervised machine learning algorithms that try to analyze, model and predict complex data. The classificatory capacity of neural networks is one of the greatest utilities, being superior to classical statistical techniques and eliminating the need for compliance with theoretical assumptions [72].

NNs have high versatility. Simple NNs follow a model of a single layer with interconnected nodes. Additionally we can find different numbers of layers, each one with diverse connection characteristics and number of nodes that iteratively extract complex features of deeper levels and perform multiple non-linear transformations to the raw data [73] [74].

⁹ <https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering>

¹⁰ <https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html>

This thesis employed Long Short-Term Memory (LSTM). LSTM is a Recurrent Neural Network (RNN) - a kind of NN which uses feed-forward loops to reuse the information, producing temporality in the data sequence [75].

"The LSTM contains special units called memory blocks in the recurrent hidden layer. The memory blocks contain memory cells with self-connections storing the temporal state of the network in addition to special multiplicative units called gates to control the flow of information. Each memory block in the original architecture contained an input gate and an output gate. The input gate controls the flow of input activations into the memory cell. The output gate controls the output flow of cell activations into the rest of the network. Later, the forget gate was added to the memory block" [75]

All of these clustering and Machine Learning technologies were employed for the purpose of optimally identifying, and then intelligently reusing classes of statements that have been formally represented in the form of an Ontology.

1.13. Ontologies

Until now, most of the knowledge generated by scholarly activity is contained in the form of narrative text or in databases. Text is largely unstructured and non-machine-accessible. Databases, although they are ordered, do not (usually) have a formal definition of types, properties and relationships between the elements that compose them. Recently, there has been a movement in Data Science to utilize computable structures called Ontologies as a means to provide highly structured data and knowledge in a form that can be accessed and "interpreted" (i.e. correctly processed) by machines.

"In computer science and information science, an ontology formally represents knowledge as a set of concepts within a domain, and the relationships among those concepts... Ontologies are the structural frameworks for organizing information and are used in artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics ... "[76]

Formalizing certainty as an ontology would help to better represent the knowledge claim when isolated from its context, as is often the case with text-mining outputs. Previous works have begun to address this objective. The lightweight ontology ORCA (Ontology of Reasoning, Certainty and Attribution) created by De Waard and Schneider - to our knowledge, the only attempt to formalize certainty categories into an ontological framework - model their four proposed levels of certainty (doxastic, dubitative, hypothetical and lack of knowledge); however, these four levels are only speculative, and have not been validated against reader-perceptions, nor are they associated with formal definitions that would enable automated classification. Our work, therefore, will need to determine if a modification or extension of the ORCA ontology is warranted, or if our conclusions warrant synthesis of an entirely novel ontology to represent certainty.

1.14. Identifying Reference Spans

In the scientific literature, it's a common practice to use citations to other articles and authors to ground your ideas, projects, experiments and to highlight certain contributions of the referenced paper. It is a common method of demonstrating your knowledge in a specialist domain, or the innovation of your research with respect to what has already been published.

Unfortunately, citing texts are often inaccurate or not sufficiently informative [77]. Moreover, they almost invariably cite an entire article, making it difficult to clarify which portion of the article (sentence, image, table, etc.) is being referred to [78] [79]. Therefore, being able to identify these article fragments - called 'reference spans' - would help clarify the purpose and accuracy of a citation. Many projects are involved in the creation of text mining tools that help recover reference spans, with approaches such as:

TF - IDF: Term Frequency - Inverse Document Frequency. TF-IDF is a weight measure commonly used in knowledge retrieval and text mining. This measurement highlights the importance of a word in a document taking into account its context within a collection of documents or other type of corpus. Words become proportionally more important the more times they appear in a document but are offset by their prevalence (total number of documents where that word appears) in the corpus.

Word2Vec: Word2Vec is a machine learning algorithm that creates a vector representation of words. These words can be represented in a multidimensional space, where similar or related words are closer. Therefore, captures a large quantity of precise syntactic and

semantic information, since similar words (e.g. words like “lion”, “tiger” and “panther”) fall into the same multidimensional area. Normally, it must be trained on a large corpus in order to generate high-quality word vectors, preserving internal similarity on the meaning(s) of words. A word may also have various “types” of similarity, termed its “context”.

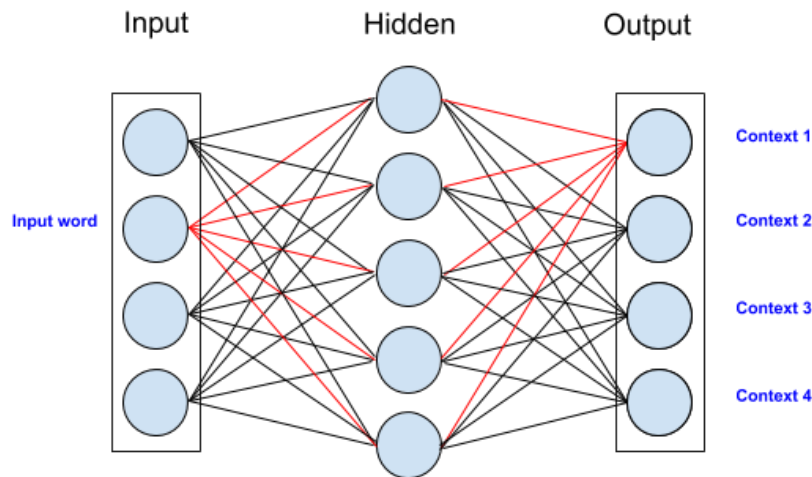


Figure 3: Neural Network structure.

The algorithm is trained as a neural network with a single hidden layer. Two different models are available - CBOW and Skip-gram. The representation of the vector will be formed by the weights from the active node as input to the hidden nodes. The network, although it begins with random weights, is trained to adjust them to the word input, to predict how well that word will fit in a particular context. The ultimate goal of this model is, given an input word, predict which words fall within the “window” of the number of words examined around the input word.

Recently, it has been shown that the use of unsupervised Neural Networks encoded as word-embedding, such as Word2Vec (vector representations of words) is capable of recognizing latent knowledge and capturing complex concepts in the material science domain [80].

TF - IDF Embedding Vectorizer: A mixture of Word2Vec and TF - IDF. *“if a word was never seen - it must be at least as infrequent as any of the known words - so the default idf is the max of known idf’s”¹¹.*

Doc2Vec: A Machine Learning algorithm based on the idea of Word2Vec. Rather than creating a vector representation of a word, like Word2Vec, it creates a vector representation

¹¹ <http://nadbordrozdz.github.io/blog/2016/05/20/text-classification-with-word2vec/>

of a document (regardless of its length). As such, “document” may mean a sentence, paragraph or a whole document. Scientific articles are separable into smaller components, such as sections, paragraphs or sentences. This, therefore, is a useful way to approach the analysis of a document, since these fragments may address different topics, or represent them in varying linguistic ways, thus allowing more targeted inspection.

2. DESCRIPTION OF THE RESEARCH PROBLEM

Description of the Research Problem

As discussed previously in the “Related Work” section, several metrics have been developed to analyze scholarly certainty. These metrics seek to explore, classify and qualify the characteristics of natural language, such as axiomatic facts versus hypothetical statements. The language used in narrative text is strongly associated with the specialist domain represented in that text. For example, in the sciences the structure and format of the language, despite being free-form, follows a large but common set of general patterns and, in fact, these patterns are often taught to young researchers by their peers or mentors. One common example is the use of the “passive voice” (e.g. “this was seen”) versus the “active voice” (e.g. “we saw”); however, there are many such examples within the specialist domains of observational and experimental sciences, some of which are shown in Table 2.

Table 2: Passive voice versus active voice examples in experimental sciences.

<i>“We know from the <i>rb</i> analysis above that <i>cis-eQTLs</i> are almost perfectly correlated in different brain regions” [81]</i>
<i>“These data suggest that there might be an interaction between the high-salt intake and the mineralocorticoid-induced loss of potassium” [82]</i>
<i>“We speculate that these spectral changes are associated with a complicated, and as yet unknown, interaction between the surface luminescence centers and hydrogen(<i>H</i>) content.” [83]</i>

Prior studies have assumed that, as a result of these linguistic patterns (or perhaps, in spite of them) communication related to the certainty of a given scholarly assertion is happening in an effective manner. However, in none of these prior studies has that assumption been tested; moreover, the details of this assumption, such as the number of certainty categories, was not shared in-common between these prior works. Several questions, therefore, are largely open: To what extent are we successfully communicating “certainty” with respect to a shared interpretation of the readers; and to what resolution are we capable of communicating certainty, given the vast array of grammatical constructs used?

To begin such a study, we must start with the information receiver, and determine if all recipients of information are perceiving it the same way. If not, then we would be forced to

conclude that the scholarly narratives are extremely inefficient at communicating scientific results! However, the progress of science indicates otherwise. Therefore, we must first establish two core facets of knowledge: 1) How many ‘categories’ of certainty can be transmitted with high-fidelity (that is, where the perception of most readers agree), and 2) what are the defining features of those categories. Once this is determined, we propose that this meta-knowledge could be formalized, and thereby become amenable to automated knowledge capture/sharing/exploration.

The use of “perception” with respect to the first facet immediately indicates that this issue will have to address the difficult problem of *the subjectivity of each reader*. Subjective studies are commonly accomplished through the application of a questionnaire, asking the participant’s opinion about some issue. We will apply this approach over a set of participants who are not trained in certainty/linguistics, but are trained in the knowledge domain being expressed in the scholarly assertions. Questionnaires will use a variety of certainty scales as we attempt to dissect the number and types of certainty categories that our participants are reproducibly detecting, using inter-annotator agreement to assess the quality of the proposed categorization system represented by each questionnaire.

Once a categorization system has been identified, we will formalize it into a Description Logic, with formal classes being backed by a machine learning model, which will then enable reliable and automated sharing and reuse by machines. This formalization will be executed either by publishing a Certainty Ontology *de novo*, or by extending/expanding an existing Certainty Ontology, such as the ORCA ontology of De Waard (De Waard, et al., 2012). This also accomplishes the goal of bringing our research outputs in-line with many of the requirements of the FAIR Data Principles, and moreover, enables the use of these categories within other FAIR data structures such as Nanopublications.

3. RESEARCH OBJECTIVES

Research Objectives

The main objective of this thesis is to move towards a deeper understanding of the concept of scholarly certainty, and explore ways that the creation of certainty metrics might be applied to addressing aspects of the ‘crisis’ in science. For example, we wish to clarify the degree to which scholars experience certainty around scientific statements, and how do they express this understanding linguistically. Further, we wish to establish means to recapture these impressions from the narrative literature, and create a framework for transparency and reusability of these impressions, improving the contextual metadata related to citations/references, and the automated capture of scholarly attributions through, for example, natural language processing. This translates into three specific lines of research activity.

Research line one: **certainty/confidence determination**. This relates to the ability of researchers to discern the different levels of certainty that exist in the scientific literature. Certainty level determination will be the first step toward creating a transparent, unambiguous, and harmonized approach to simplifying the complexity and subjectivity of scientific discourse.

Research line two: **certainty assignment**. This refers to the creation of machine learning models to automatically undertake the categorization of statements they encounter, with regard to the levels of certainty identified in research line one. The need for automated certainty assignment follows directly from, and depends on, our success in Research line one, where that work provides the categories used to train the classifier, which in turn is able to assign metadata related to a new statement’s confidence. This will enable the creation of automated tools capable of, for example, comparing citations to the text they reference and determining how accurately the citation matches the confidence expressed in the reference, which until now has not been possible other than through tedious and subjective manual curation. In addition, this can then be used as a filter or a flag to capture this important qualitative metadata when scholarly assertions are ingested into databases through text-mining. Certainty assignment, therefore, is crucial to automate the process of capturing and disseminating this key piece of metadata about a scholarly assertion.

Research line three: **certainty logics and its evaluation**. This follows directly from the previous two research lines, and is a requirement for the dissemination and reuse of the knowledge gained from them. After defining categories, these must then be formalized in a manner that allows them to be assigned to statements in a machine-readable way, and shared between computational agents on the Internet in a manner that preserves their

interpretation, and where possible, follows the emergent requirements for adherence to the FAIR Principles. This will involve several steps: formally capturing the certainty categories in a logical format; defining how to utilize those categories in a formal, machine-accessible syntax; and objectively demonstrating that these formal representations are, in fact, machine-accessible according to FAIR.

Certainty logics will help all stakeholders – in particular, computational agents - share a common understanding of the degree of “hedging” associated with scholarly statements, and moreover, compare these expressions of certainty along a citation chain.

The next section introduces our main hypotheses, breakdown of the distinct steps taken.

3.1. Research Hypotheses

Overarching Hypothesis: *Certainty/Confidence is a measurable quality of a scholarly statement, communicated (expressed and perceived) by scientists in a relatively consistent manner, that can be automatically detected and assigned.*

Our hypothesis can be split into three sub-parts, which focus on the pillars of: certainty determination, certainty assignment, and formal expression of certainty using description logics. We formally state these Hypotheses as follows:

H1: *Scholarly certainty can be quantified, and detected in a reproducible manner, across linguistic groups.*

H1.1: *Certainty is not a continuum, but rather, can be represented as discrete categories, perceived uniformly by the readers of scholarly literature.*

H2: *Artificial intelligence models can be designed that accurately assign certainty levels to novel scholarly statements.*

H2.1: *Application of certainty categories to citation chains will reveal “inflexion points”, where the certainty level changes (increases or decreases).*

H3: *We can formalize the capture and exchange of these certainty assignments through using semantic technologies such as Resource Description Framework and Web Ontology Language, and formal*

knowledge-publication frameworks such as Nanopublications, in a manner that follows the FAIR Principles.

A summary representation of these hypotheses and how they are connected with certainty determination, assignment and logics is presented in Table 3.

Table 3: Hypothesis and their corresponding certainty research areas.

Hypothesis	Certainty Research Area
H1: Certainty/Confidence is a measurable facet of a scholarly statement, perceived by scientists in a uniform way	Determination, Assignment, Logics
H1.1: It is possible to understand and discern different levels of certainty, observe their distribution, create a quantitative metric that encompasses its entire dimension and catalogue its range of use	Determination
H2: The creation of artificial intelligence models that assign the different levels previously extracted to new statements help automate a comparative process.	Assignment, Logics
H2.1: Using these levels we can detect the historical certainty of assertions along a citation chain and identify where inflection points could arise.	Determination, Assignment
H3: Designing and testing a machine-interpretable representation, such as nanopublications, which utilizes a formal ontological framework for representation of the certainty categories, grounded in (and output from) the machine-learned models.	Logics

4. MATERIALS AND METHODS

4.1. Broad overview

Using the TAC Biomedical Summarization Corpus [84], we extracted 45 manually curated scholarly assertions (selection process described below). Using this corpus, a total of 375 researchers in the biomedical domain, in comparable research institutes and organizations, were presented with a series of assertions and asked to categorize the strength of those assertions into four, three, or two certainty categories over three independently-executed questionnaires. To determine the degree of agreement between annotators, G Index [85] coefficient analysis was applied. This allowed us to evaluate the power of each categorization system - that is, to test the discriminatory effectiveness of the categories themselves, versus the quality of the annotations or annotators. We extracted the essential features of inter-rater agreement from the questionnaire data using Principal Component Analysis (PCA) to guide our interpretation of the way annotators were responding to the categories presented. The essential number of components identified by PCA were extracted using Horn's parallel analysis, with three categories appearing to be the optimal. We then clustered our collection of statements into these three categories using a k-means algorithm [86] [87]. Based on our examination of the contents of these categories, we then manually generated a corpus of statements annotated using these three categories, and applied deep-learning techniques over this corpus to generate an automated classifier model. Cross-Validation (CV) was used to evaluate accuracy. We then used this classifier to reveal changes in expressed certainty over statements in several sequential papers and/or over citation chains in the biomedical literature, and demonstrating that it was successful in detecting known trends towards higher levels of certainty over time. To aid in this activity, we made preliminary attempts to automate the identification of citation chains via the detection of reference spans. Finally, we show, and formally evaluate, how certainty metadata could be represented in a popular machine-readable publication format called a NanoPublication, showing how it could be useful to enhance text-mining platforms.

4.2. Survey statement selection

The first step in designing the certainty Survey was to select a corpus of scholarly statements that had been richly curated by experts to ensure that they represented real scholarly assertions, of varying types, rather than simple narrative or conversational statements. For this we used the Text Analysis Conference (TAC) scientific summarization Corpus 2014 [84]. This workshop is a regular event that focuses on evaluating the ability of computational tools to automatically recognize the "reference spans" for different

citations that contain the citing articles and the identification of the discourse segment type to which it belongs. All of these “gold standard” annotations were generated by a group of four biomedical domain experts, thus, this represents a high-quality manual curated corpus.

The 45 text blocks used in the three Surveys were extracted from published articles related to genetic and molecular issues, and were selected from the “Citation Text” and “Reference text” portions of the TAC 2014 Biomedical Summarization Track. Each text block contained a sentence or sentence fragment representing a single scholarly assertion that we highlighted and asked the respondents to evaluate, with the remainder of the text being provided for additional context. The 45 assertions were selected using different epistemic modifiers, such as modal verbs (can, could, may, might, shall, should, will, would), qualifying adjectives and adverbs (interestingly, strongly, clearly, possibly, likely, slightly, unknown), reporting verbs (consider, imply, suppose, demonstrate, agree, confirm, suggest), which are believed to be grammatical indicators of “value of truth” statements [3]. Given that they are intended to be used for a human Survey, with the aim of avoiding annotator fatigue, these were further filtered based on the length of the statement to give preference to shorter ones.

De Waard [3] identified epistemic modifiers that she believed were indicative of various degrees of certainty (Lack of knowledge; Hypothetical or low certainty; Dubitative or higher likelihood; Doxastic or complete certainty). In order to ensure that our statement-set was representative of a wide range of certainty levels, and in the absence of other proposed linguistic indicators of certainty, we utilized these epistemic modifiers as a filter in our statement selection in an attempt to capture the full range of certainty possibilities. A subset of 45 statements were then selected arbitrarily ensuring representation of all “levels” suggested by De Waard. This provided our final corpus [88] that was used for the remainder of the studies.

4.2.1. Statements used

The 45 statements selected from the TAC 2014 Biomedical Summarization Track¹² are as follows:

1- Today, although aerobic metabolism and the corresponding generation of ROS remain the most widely accepted cause of aging, substantial gaps and unknowns persist. Although it is commonly assumed that an increase in oxygen consumption produces an increase in

¹² TAC 2014 Biomedical Summarization Track. de Waard, A., Vanderwende, L. <https://tac.nist.gov/2014/BiomedSumm/>

ROS production, we would argue that this positive correlation is only true if the increase in oxygen consumption was secondary to a higher tissue pO₂ or an increase in the number of sites.

2- There is now compelling evidence that particular components of this regulatory machinery act as tumor suppressors or protooncogenes, whose mutations occur so frequently as to prompt speculation that disabling “the RB pathway” may be essential for the formation of cancer cells (Sherr, 1996, Sellers and Kaelin, 1997, Nevins, 2001, Hahn and Weinberg, 2002 and Ortega et al., 2002).

3- Interestingly, the inhibition of p53, alone or in combination with pRb, was sufficient to overcome the RASV12-induced arrest to an extent similar as LT. In conclusion, our results indicate that loss of p53 is sufficient to overcome RAS-induced oncogenic stress in primary cells. However, this was not sufficient for full blown transformation of primary human cells, which also required the collaborative inhibition of pRb, together with the expression of hTERT, RASV12.

4- Hence, the extent to which miRNAs were capable of specifically regulating metastasis has remained unresolved.

5- Therefore, it is unlikely that the localization of Aurora-A in the nucleus as determined by these two antibodies is a result of cross-reaction of anti-Aurora-A antibodies.

6- However, the possibility seems unlikely because we have also observed that the human lung cancer cell line, H460, displays similar growth inhibition when it is transduced by MIGR1-Lats2 retrovirus (data not shown), while expression of Lats1 in H460 resulted in a G2/M block and induced apoptosis (Xia et al., 2002).

7- The expression of Oct4 in various forms of human cancer [8], [9] and a recently described role for Oct4 in adult stem cells [10] and the expansion of epithelial progenitor cells [7] supports the theory that cancer is a disease of stem cells. This theory postulates that cancers arise in stem cells or early committed progenitors [58] due to their inability to differentiate in a regulated fashion. Oct4 directly regulates the transcription of genes such as Trp53, Brca1, Parp1, and Bmi1 which play a central role in a cell's proclivity to undergo transformation, apoptosis, senescence, and now differentiation.

8- When these 56 tumors were stratified based on clinical progression, we found that miR-31 expression was diminished in primary tumors that subsequently metastasized, when compared to normal breast tissue and primary tumors that did not recur; moreover, low

miR-31 levels correlated strongly with reduced distant disease-free survival relative to tumors with high miR-31. Thus, miR-31 may represent a marker for metastasis in a variety of breast cancer subtypes; however, its utility as a prognostic indicator will depend on extension of these initial observations.

9- These worms also were shown recently to exhibit increased levels of nuclear DNA damage (Hartman et al., 2004), arguing that mitochondrial oxidants might be an important source of overall genomic instability. This conjecture is also supported by observations in mice that are heterozygous for mitochondrial superoxide dismutase (Sod2+/-).

10- Since these 32 genes are also required for zVAD.fmk-induced necroptosis, we hypothesize that these 32 genes represent potential core components of the necroptotic pathway.

11- Although the exact mechanism responsible for this effect is still unclear, we suggest that suppression of LATS2 is an important factor.

12- Finally, we have identified 26 direct Oct4 transcriptional targets which may represent candidate regulatory nodes by which cell fate decisions could be directed to facilitate the use of hESCs in therapeutic and regenerative medicine.

13- Each of these physiologic changes—novel capabilities acquired during tumor development—represents the successful breaching of an anticancer defense mechanism hardwired into cells and tissues. We propose that these six capabilities are shared in common by most and perhaps all types of human tumors.

14- We therefore hypothesized that conserved regions in mRNAs may serve as docking platforms for modulators of miRNA activity.

15- While the remainder have invented a way of activating a mechanism, termed ALT, which appears to maintain telomeres through recombination-based interchromosomal exchanges of sequence information (Bryan et al. 1995). The role of telomerase in immortalizing cells can be demonstrated directly by ectopically expressing the enzyme in cells, where it can convey unlimited replicative potential onto a variety of normal early passage, presenescent cells in vitro.

16- Whether miRNAs act mainly as tumor suppressors (suppressor-miRs), promoters of tumorigenesis (onco-miRs) or both is still widely elusive, but the global decrease in miRNA expression in human cancers suggests that most miRNAs may act as direct suppressor-miRs or post-transcriptional repressors of known oncogenes.

17- The two major sites for ROS generation are believed to be at sites I and III where large changes in the potential energy of the electrons, relative to the reduction of oxygen, occur. Experimental manipulations that increase the redox potential of site I (Kushnareva et al., 2002) or site III (Chen et al., 2003) generally increase the rate of ROS generation, supporting the notion that the redox potential of these reactive sites is important in free radical formation.

18- In light of the recent report that Drosha is responsible for the nuclear processing of miRNA primary transcripts (23), our results can be explained by the idea that elements needed for Drosha recognition reside within the sequences that flank the miR-223 predicted hairpin.

19- The observation that the differentiation of myeloid and other lymphoid cell types was not totally blocked when the B-lymphoid lineage increased suggests that miR-181, at least when considered singly rather than in combination with other miRNAs, appears to function more as a lineage modulator than as a switch.

20- Indeed, 22.6% of breast cancers exhibit evidence of deletion (one homozygous deletion of 15.0 Mb) at the PTPN12 locus, though the deletions exhibit a median size of 22.9 Mb, suggesting that multiple driver mutations may exist in this region. These data suggest that PTPN12 is inactivated, in part, via deletion in a wide range of cancers and support the hypothesis that PTPN12 is a frequently inactivated tumor suppressor.

21- The conclusion is that tumor cells generate many of their own growth signals, thereby reducing their dependence on stimulation from their normal tissue microenvironment. This liberation from dependence on exogenously derived signals disrupts a critically important homeostatic mechanism that normally operates to ensure a proper behavior of the various cell types within a tissue. We suspect that growth signaling pathways suffer deregulation in all human tumors. Although this point is hard to prove rigorously at present, the clues are abundant (Hunter 1997).

22- By using viral and mutant human genes, as well as homologous recombination, it became apparent that the pRb and p53 pathways are involved in this process. [Hahn et al. 1999 and Wei et al. 2003].

23- It is probable that mouse Lats2 also regulates the G1/S transition through down-regulation of Cyclin E/Cdk2 kinase activity in NIH3T3 cells (Li et al. 2003).

24- Thus, although we agree that increased levels of p53 cause activation of a rapid and massive apoptotic response in EC cells, [we believe that the trigger for this response is the persistence of DNA damage in cells that can poorly repair it](#). In this context, while the development of TGCTs would be allowed by a partial functional inactivation of p53, such mechanism would be insufficient to counteract the proapoptotic function of p53 induced by a persistent damage, causing a rapid cell death.

25- [Here we present evidence that INPP4B functions as a tumor suppressor](#). Knocking down the expression of this enzyme in HMEC cells results in anchorage-independent growth and enhanced motility, similar to changes induced in response to knockdown of the PTEN tumor suppressor gene.

26- miRNAs are regulatory, non-coding RNAs about 21-23 nucleotides in length and are expressed at specific stages of tissue development or cell differentiation, and have large-scale effects on the expression of a variety of genes at the post-transcriptional level. Although their biological functions remain largely unknown, recent studies suggest that miRNAs contribute to the development of various cancers [19] and might function as an important component of the cell's natural defense against viral infection [20] [22]. [It has been proposed that an unique miRNA expression profile for a particular cancer would be a useful biomarker for cancer diagnosis \[23\] and prognosis \[24\]](#).

27- [We confirmed the consistency of miRNA expression driven by miR-Vec by cloning eight miR-Vec plasmids expressing randomly chosen miRNAs](#). With one exception, all constructs yielded high expression levels of mature miRNAs.

28- We show that high level expression of oncogenic RAS induces deregulated growth and loss of differentiation in cultured rat thyroid cells, whereas low levels of the same protein elicit only deregulated growth without interference with differentiation. [In accordance with these observations, our data show the requirement of high RAS oncoproteins overexpression to achieve the loss of the differentiated phenotype, at least 10-fold over the endogenous wild-type RAS](#).

29- A strong link between miRNA dysregulation and human cancer has been established [14]. [Consequently miRNAs have been demonstrated to act either as oncogenes \(e.g., miR-155, miR-17-5p and miR-21\) \[15,16\] or tumor suppressors \(e.g., miR-34, miR-15a, miR-16-1 and let-7\) \[17-20\]](#). However, the precise regulatory features that tip the balance towards a cancer phenotype with respect to tumor suppressor versus oncogenic miRNA expression are poorly understood.

30- Recently, it has been recognized that animal and plant genomes contain an abundance of small regulatory RNAs of approximately 22 nucleotides (nts) in length [11], [12], [13], [14], [15], [16] and [17]. One class of small RNAs called microRNAs (miRNAs) has a variety of functions and is involved in the regulation of gene expression [18], [19], [20], [21], [22], [23], [24], [25] and [26].

31- Since activating mutations are found in several members of the MAPK and PI3K pathways in many human cancers, we reasoned that the coactivation of these two pathways might replace H-RASV12 in transformation. To test this hypothesis, we manipulated the MAPK and PI3K pathways in HA1E cells.

32- Our observations raise the important prediction that many malignancies considered to be non-TK driven because of the absence of a dominant TK mutation may indeed be dependent on TK signaling. It is likely that in different cell types, different PTPs may play roles similar to PTPN12 in suppressing tumorigenesis, possibly by antagonizing different combinations of TKs.

33- We were able to confirm that the cancer tissues had reduced expression of miR-126 and miR-424, and increased expression of miR-15b, miR-16, miR-146a, miR-155, and miR-223, after individual miRNA level in each sample was quantified and normalized to U6 expression. Another study suggested that overexpression of miR-17-5p, miR-20a, miR-21, miR-92, miR-106a, and miR-155 could be considered an miRNA signature of solid cancer [59].

34- To address this question, synthetic miR-143 and miR-145 precursors were transiently transfected into HeLa cells and the effect of overexpression of their mature miRNAs on HeLa cell growth was evaluated by cell counting. Both miR-143 and miR-145 were found suppressive ($p < 0.005$ for miR-143 and $p < 0.008$ of miR-145, t-test) to HeLa cell growth. Data suggest that both miR-143 and miR-145 probably need to be downregulated in cervical cells for tumor progression.

35- In a third experiment, we tested the ability of cells to form tumors in athymic nude mice. We injected 2 million cells subcutaneously and monitored tumor growth after 4 weeks. Both the LT overexpression and the concerted inhibition of pRb and p53 in BJ-ET/st/RAS allowed for highly efficient tumor formation in mice. Inactivation of only p53 gave rise to inefficient tumor growth in some of the mice (2 out of 6), most likely due to increased genomic instability in the absence of p53 while propagating the cells before the injections. Lastly, mice injected with BJ-ET/st/RAS cells, in which only pRb expression was inhibited,

developed no tumors. In conclusion, our results indicate that loss of p53 is sufficient to overcome RAS-induced oncogenic stress in primary cells.

36- Interestingly, knockdown of Bmf blocked necroptosis induced by zVAD.fmk and TNF, but not apoptosis of TNF/CHX-treated NIH 3T3 cells. Although Bmf has been implicated as a proapoptotic molecule, this result suggests that, at least in the death receptor signaling pathway, Bmf is primarily involved in mediating necroptosis, but not apoptosis. [It is possible that the activation of Bmf may induce either apoptosis or necroptosis in a stimulus and cellular context-dependent manner.](#)

37- Although these observations do not eliminate the possibility that other genes within this 1q32 amplicon may cooperate with IKBKE to induce transformation, [these findings suggest that IKBKE is a key target of the 1q32 amplification in breast cancer cell lines and tumors.](#)

38- 1299 probesets (1155 unique transcripts) were found to be correlated to Oct4. Seventy-five probesets (69 transcripts) were negatively correlated, while 1224 probesets (1086 transcripts) were positively correlated. [The validity of this method for the identification of genes related to stem cell identity is assured by the presence of genes which have previously defined roles in ESCs](#) such as Utf1, Fgf4, Nanog, and Sox2 which were correlated to Oct4 in 100%, 99%, 97% and 49% of the trials respectively.

39- Our work adds to this list a set of hematopoietic-specific miRNAs that [presumably act by pairing to the mRNAs of their target genes to direct gene silencing processes critical for hematopoiesis.](#)

40- The two major sites for ROS generation are believed to be at sites I and III where large changes in the potential energy of the electrons, relative to the reduction of oxygen, occur. [Experimental manipulations that increase the redox potential of site I \(Kushnareva et al., 2002\) or site III \(Chen et al., 2003\) generally increase the rate of ROS generation,](#) supporting the notion that the redox potential of these reactive sites is important in free radical formation.

41- Here, we identify IKBKE as a human breast cancer oncogene. IKBKE is amplified and overexpressed in a significant percentage of human breast tumors. In addition, IKBKE expression leads to cell transformation when overexpressed in immortalized human cells at levels found in breast cancer specimens, and cancer cell lines that exhibit IKBKE copy-number gain or amplification require IKK expression for viability. [Together, these observations strongly support the conclusion that amplifications of IKBKE play an](#)

important role in the pathogenesis of a subset of breast tumors and identify IKK as promising cancer target.

42- Although some studies have shown nuclear localization of LATS2,15,23 most studies clearly demonstrate centrosomal localization of both LATS1 and LATS2 during interphase 14,22,27,29-33 and association with the mitotic apparatus in mitotic cells.

43- The emerging role of miRNAs in the regulation of fundamental set of cellular mechanisms such as proliferation, apoptosis, development, differentiation and metabolism [9]–[16] clearly suggests that any aberration in miRNA biogenesis pathway or its activity contributes to the human disease pathogenesis including cancer [17].

44- We suggest that the critical importance of pairing to segment 2-8 for target identification in silico reflects its importance for target recognition in vivo and speculate that this segment nucleates pairing between miRNAs and mRNAs.

45- Therefore, in our human model system, it is unlikely that p14ARF has a essential protective role against RAS-induced oncogenic transformation, in contrast to what has been found in mice.


Each fragment contains one statement highlighted in blue with the remainder of the text being provided for additional context. The blue statement is the portion we asked our Survey participants to evaluate for the level of certainty perceived.

4.3. Survey design

We executed three Surveys - S1, S2 and S3 - where respondents were asked to assign certainty based on a number of certainty categories - four, two, and three respectively. All Surveys used the same corpus of 45 scholarly assertions. To minimize the bias of prior exposure to the corpus, the Surveys were deployed over three comparable but distinct groups of researchers, all of whom have sufficient biomedical expertise to understand the statements in the corpus. Recruitment for the Surveys was primarily achieved through personal contact with department leads or heads of five institutions with a focus on biomedical and biotechnology research. Participation was anonymous. Structural and statistical information about each Survey is as follows:

<p>SURVEY #1 (S1)</p>	<p>Date: 12/16</p> <p>Target audience: Centro de Biotecnología y Genómica de Plantas.</p> <p>Method of recruitment: Participation in the Surveys was primarily achieved through personal contact with department leads/heads and internal email distribution.</p> <p>Platform: Survey Gizmo</p> <p>Participants: 101</p> <p>% To Completion: 74.2%</p> <p>Certainty Categories Tested: <i>High, Medium High, Medium Low, Low.</i></p>
<p>SURVEY #2 (S2)</p>	<p>Date: 11/17</p> <p>Target audience: Leiden University Medical Center.</p> <p>Method of recruitment: Participation in the Surveys was primarily achieved through personal contact with department leads/heads and internal email distribution.</p> <p>Platform: Qualtrics</p> <p>Participants: 215</p> <p>% To Completion: 69.8%</p> <p>Certainty Categories Tested: <i>Relatively High, Relatively Low.</i></p>
<p>SURVEY #3 (S3)</p>	<p>Date: 11/18</p> <p>Target audience: University Medical Center Utrecht, Cell Press, and the Agronomical Faculty of Universidad Politécnica of Madrid.</p> <p>Method of recruitment: Participation in the Surveys was primarily achieved through personal contact with department leads/heads and internal email distribution.</p> <p>Platform: Qualtrics</p> <p>Participants: 57</p> <p>% To Completion: 84.2%</p> <p>Certainty Categories Tested: <i>1, 2 and 3.</i></p>

The first Survey (S1) was performed in December 2016, using an online platform specializing in delivery of questionnaires, called "Survey Gizmo". The majority of participants came from the Centro de Biotecnología y Genómica de Plantas (UPM-INIA), Spain. Each participant was presented with an introductory page about the main goal of the Survey, what 'certainty' means within the context of our study, and a concise explanation regarding the purpose of the questionnaire, and what they are expected to do (Figure 4).



Wilkinson Laboratory

Welcome to our questionnaire. Before we begin, a brief explanation.

What is certainty?

Certainty is the mental state of being without doubt. ...but people say "I am VERY certain"? or "I am QUITE certain"? What does that mean? Can we interpret it? Can we quantify it?

The level of certainty?

We want to understand two things: 1) How do scientists express uncertainty, and 2) do other scientists *interpret* those statements in the same way?

To answer these questions, we are asking YOU, dear colleagues, to help us quantify the level of certainty being expressed by your peers.

For Example:

Gene X could be induced by gene Y.

After reading that statement, what is your impression? Is it a strong, confident statement, or is it a weaker, uncertain statement? NOTE: **Please, forget what you know about biology**, and only evaluate the sentence in blue. We are not interested if the statement is true or not (even if you think you know!!), we ask you to only focus on **how much certainty you have** after reading that statement.

One more thing...

While we have your attention, we would like you to help us understand one more thing - what led the researchers to make that statement?

- **Direct Evidence** - The screen of my computer is black when switched off - I know this because I switched it off and it became black.
- **Indirect Evidence/reasoning**: I measured the temperature of the screen, and it was cool. Screens are warm when they are switched on, so my screen must be off. Therefore it must also be black.
- **Speculation**: I propose that the screen may be black
- **Citation**: Mario's computer screen is black [Prieto et al. 2014].

It will only take about ten minutes of your time

Are you CERTAIN you are ready to begin? ;-)

Figure 4: Example of the Survey 1 questionnaire introduction. A brief introduction to what certainty is and intent of the questionnaire.

Each participant was then presented with 15 statements, randomly selected from the corpus of 45. For each clause, they were asked to evaluate the level of certainty they believed was being expressed by the statement highlighted in blue. The possible answers were: *High, Medium High, Medium Low, Low*.

Scientific Statement:

Our observations raise the important prediction that many malignancies considered to be non-TK driven because of the absence of a dominant TK mutation may indeed be dependent on TK signaling. It is likely that in different cell types, different PTPs may play roles similar to PTPN12 in suppressing tumorigenesis, possibly by antagonizing different combinations of TKs.

Forget what you know about biology... What do you think is the certainty level expressed by the authors in the statement highlighted in blue?

High

Medium High

Medium Low

Low

Figure 5: Example of the Survey 1 questionnaire interface. Question 1.1. A scholarly assertion is highlighted in blue, in its original context. Participants were asked to characterize the blue assertion, using one of four levels of certainty (*High, Medium High, Medium Low* or *Low*).

Below that question, and for the same statement, the participant was asked to indicate their opinion regarding the basis upon which the author made the statement. For this, the possible answers are: *Direct Evidence, Indirect Evidence / Reasoning, Speculation, Citation* and *I don't know*.

What is the basis for this statement?

Direct Evidence

Indirect Evidence / Reasoning

Speculation

Citation

I don't know

Figure 6: Example of the Survey 1 questionnaire interface. Question 1.2. A scholarly assertion is highlighted in blue, in its original context (In Question 1.1). Participants were then asked about the basis of the given statement, using one of five options (*Direct, Evidence, Indirect Evidence/Reasoning, Speculation, Citation* and *I don't know*).

Finally, after all 15 statements were evaluated, participants were presented with a final challenge. Five assertions (i.e. only the blue portion) from the corpus of 45, were presented and the respondent was asked to rank those in order from highest to lowest level of certainty (from one to five, with one being the highest and five the lowest). These assertions had been previously grouped into sets of five, where one of the nine possible sets was presented to the respondent. The respondent may or may not have seen any of these statements in the 15 prior Survey Question 1 statements.

Last Question !! What are the level of certainty of these statements?

RANK each statement from High Certainty(1) to Low Certainty(5) by "dragging and dropping".

We propose that these six capabilities are shared in common by most and perhaps all types of human tumors.

It became apparent that the pRb and p53 pathways are involved in this process.

We would argue that this positive correlation is only true if the increase in oxygen consumption was secondary to a higher tissue pO₂ or an increase in the number of sites.

Hence, the extent to which miRNAs were capable of specifically regulating metastasis has remained unresolved.

Data suggest that both miR-143 and miR-145 probably need to be downregulated in cervical cells for tumor progression.

Figure 7: Example of the Survey 1 questionnaire interface’s final question. Only the blue segment of five scholarly assertions were presented. Participants were asked to drag and drop each statement from high to low certainty.

In Survey #2 (S2), the same set of 45 statements was reused to ensure cross-comparisons between questionnaires were valid. S2 was distributed among all the departments of the Leiden University Medical Center (LUMC), using Center's internal mail between November and December 2017. The Qualtrics questionnaire platform was utilized, primarily due to its lower cost; the Qualtrics platform is essentially the same as Survey Gizmo with respect to its one-at-a-time presentation of the questions and the Web buttons respondents use to answer them, with the primary difference between the platforms being aesthetic (color, font, branding). In S2, only two levels of certainty are offered: *Relatively High* and *Relatively Low*.

Participants were first presented with the introductory page, which is identical to that of S1, but for some small edits related to having only two levels of certainty. Subsequently 20 statements were presented to each participant, randomly selected from the set of 45, and they were asked about the level of certainty of the statement highlighted in blue. The increase in number (from 15 to 20) was to increase the coverage of responses to each statement. As with S1, a second question appeared below the ranking question, asking about the basis for the assertion, maintaining the same options, *Direct Evidence*, *Indirect Evidence / Reasoning*, *Speculation*, *Citation* and *I do not know*.

At the end of the Survey a final question, again, asked the respondents to rank a set of five statements. In this case, however, the statements were randomly selected from the 45 “blue” statements (versus being pre-grouped, as in S1). This question is intended to be used as an internal control (see Discussion section 5.3).

Survey 3 (S3) was executed in December 2018. A link to the online questionnaire was sent through personal contacts with principal investigators and department heads of the Utrecht University Medical Center (UMC Utrecht), Cell Press and the Agronomical Faculty of Universidad Politécnica of Madrid. After the introductory page, 20 randomly-selected statements (from the same corpus of 45 statements) were presented to each participant. In this case, they were offered to choose from three levels of certainty, labelled 1, 2 and 3. Beneath these questions, as before, participants were asked about the basis of the assertion maintaining the same options, *Direct Evidence*, *Indirect Evidence / Reasoning*, *Specification*, *Citation* and *I do not know*. Finally, as with the previous two Surveys, participants were asked to do a final ranking of five randomly selected statements.

4.4. Ranking study

To determine if raters were responding randomly or in some other way inconsistently, we normalized the statements from each questionnaire to a common scale (from -1 to +1) to allow us to compare between surveys (e.g., Question 1 of S1 vs. Question 1 of S2) and between questions (e.g. Question 1 of S1 vs. Question 2 of S1). For this we employed a variation of *Relative importance Index (RII)* [89] as follows:

$$\text{Question 2, Survey 2. Statement's rank five levels} = \frac{(L1x1)+(L2x0.5)+(L3x0)+(L4x-0.5)+(L5x-1)}{AxN}$$

$$\text{Question 1, Survey 3. Statement's score three levels} = \frac{(L1x1)+(L2x0)+(L3x-1)}{AxN}$$

Each level of certainty represented in each Survey was assigned a value, with the highest level of certainty always assigned the value 1 and the most negative level the value -1. Other levels (one or two, depending on the Survey) are distributed evenly between 1 and -1. This formula is applied to each statement, for all responses. The result is a ranked order of statements in a range 1 to -1 (from most certain to least certain). In parallel, the final

question of the Survey presented five statements to be sorted according to their perceived level of certainty. This provides a set of five possible “scores” for each statement - L1 represents the number of times a statement (e.g. statement #34) was selected in position 1, L2 is the number of times St.#34 was selected in position 2, and so on. This number is multiplied by a specific weight for each position. This weight is the result of dividing 2 (range between 1, -1) between (Number of categories - 1). *I.e.* S1 with four levels of certainty would be $(2 / (4-1)) = 0.66$. Therefore *w* for L1 is 1, for L2 = $1 - 0.66 = 0.33$; for L3 = $0.33 - 0.666 = -0.33$; and L4 = $-0.33 - 0.66 = 0.99 \sim (-1)$. Finally, it is divided by the total number of responses for each statement (N) and the highest weight (A), which in our case is always one to provide a relative score for each statement

4.5. Statistical Analysis of Survey Responses

We evaluated each Survey by quantifying the degree of agreement between participants who were presented the same assertion. This was done by comparing the level of certainty they indicated was expressed in each statement, given the categories provided in that Survey.

Agreement between participants was assessed by Holley and Guilford's G Index of agreement [85], which is a variant of Cohen's Weighted Kappa (Kw; [90]). Ideally G measures the agreement between participants. It was performed based on the following formula:

$$G = \frac{\text{Probability Observed}(Po) - \text{Probability expected by Chance}(Pc)}{1 - \text{Probability expected by Chance}(Pc)}$$

$$G = \frac{PO - Pc}{1 - Pc}$$

The key difference between Kw and G is in how chance agreement (*Pc*) is estimated. According to [91], "G appears to have the most balanced profile, leading us to endorse its use as an index of overall interrater agreement in clinical research". G is defined *a priori*, being homogeneously distributed among categories as the inverse of the number of response categories [91], thus making G=0.25 for S1; G=0.50 for S2; and G= 0.33 for S3. The accepted threshold for measuring agreement and its interpretation has been suggested by Landis & Koch, 1977 [92,93] as follows: ≤ 0.2 = Poor; 0.21 - 0.4 = Fair; 0.41 - 0.60 = Moderate; 0.61 - 0.80 = Substantial; 0.81 - 1.00 = Almost Perfect. Anything other than the

'Poor' category is considered in other studies to represent an acceptable level of agreement. [94],[95].

4.6. Clustering

We investigated the ideal number of clusters into which statements group based on the profile of the annotator's responses, as a means of identifying the correct number of certainty categories. To estimate this, Hierarchical Clustering analysis (HCA) and the Spearman correlation test were performed to determine certainty category association between questionnaires (Fig.12), using the shared classified statements in that category as the metric [96] [97], [98], [99]. Although these represent conceptually distinct analyses, we represent them in the same chart because the outputs are mutually supportive. HCA finds clusters of similar elements, while Spearman correlation coefficient considers the weight and direction of the relationship between two variables. It is worth emphasizing the importance of the rank-based nature of Spearman's correlation. Spearman's formula rank the variables in order, then measures and records the difference in rank for each statement/variable. Thus, *"...if the data are correlated, [the] sum of the square of the difference between ranks will be small"* [100], which should be considered when interpreting the results. Interpretation of Spearman correlation was as follows: Very Low ≤ 0.2 ; Low ≤ 0.5 ; Moderate ≤ 0.7 ; High ≤ 0.9 and Very High > 0.9 [87] [101]. All Spearman interactions are based on hypothesis testing. To determine the importance of the results, p-values were generated as an indicator of the existence of correlation between certainty categories. HCA and Spearman values were generated using the python libraries *seaborn* and *pandas*.

Normalization allows the comparison of two nominal variables on different scales. Prior to Principal Component Analysis (PCA) and cluster analyses, we first normalized for the different number of annotators for each statement, and centered, using the *scale* function from R and the Python package *scikit-learn*. PCA is a widely used method for attribute extraction to help interpret results. We used PCA to extract the essential features of inter-rater agreement from the questionnaire data [97] [99]. We applied PCA using *scikit-learn* to the result-sets, and utilized K-means from the same python package to identify cluster patterns within the PCA data. These cluster patterns reflect groups of similar "human behaviors" in response to individual questions under all three Survey conditions. The input to both statistical functions was the results of the questionnaire in the form of a contingency table, where each statement is represented by the profile of annotations it received from all annotators. The optimal number of components was selected using Horn's parallel analysis,

applied to certainty categories on the three different Surveys. Detailed output is provided in Figure 25, 26 and 27 of supplementary information; our decision to choose three components as the most robust number to capture relevant features of our data is justified in the Results section. To determine the optimal K (cohesion of the clusters), several indices were analyzed using the R package NbClust [102]. NbClust provides 30 different indices (e.g., Gap statistic or Silhouette coefficient) for determining the optimal number of clusters based on a “majority rules” approach [103]. Membership in these clusters was evaluated via Jaccard similarity index comparing, pairwise, all three clusters from each of the three Surveys to determine which clusters were most alike (Table 9). This provides additional information regarding the behavior of annotators between the three Surveys; i.e., the homogeneity of the three identified categories between the three distinct Surveys. The *Princomp* and *paran* functions in R were utilized to execute PCA (Table 10) and Horn’s parallel analysis, respectively. The *PCA* and *K-Means* functions from *scikit-learn* were employed to create the visualizations in Fig. 17 [104].

Additionally, to be able to compare questionnaires with different certainty categories, we applied the *Relative Importance Index* (RII) in the three Surveys (previously described in the Ranking study section), obtaining a score for each statement. Subsequently, we conducted a comparison both between questionnaires and between each question within each questionnaire, using Wilcoxon rank-sum test (two samples) and Kruskal-Wallis test (> two samples) to determine if there were significant differences between the test subject groups, given that each questionnaire was carried out using a distinct set of respondents. Finally, we also executed an internal-control validation using the RII score from Q2 from each Survey to measure the degree to which respondents were behaving inconsistently or answering randomly and an independence test between Q2 of S2 and S3 using raw data to confirm the significant similarity between the responses of both questionnaires.

4.7. Identifying Reference Spans

We executed numerous attempts and approaches to accurately capturing the reference spans of a citation, with the desire to automate the creation of citation chains. These were intended to be used in the latter portions of the thesis, as the input to the detection of “inflection points” (i.e., where certainty changes) along a citation chain. To achieve this, we used four different approaches: TF-IDF, Word2Vec, TF-IDF Embedding Vectorizer and Doc2vec.

By way of preprocessing, several methods were applied to training set and test set in order to facilitate the capture of information. All of them were applied both jointly and separately.

- Stemming: Method to reduce a word to its root. (Playing -> Play)
- Lemmanization: Process similar to stemming but it makes sure that the resulting word belongs to the language. (Ran -> Run).
- Stopwords: set of words that are omitted or removed from the sentence, since they are considered meaningless.
- Removing references.

Word2Vec, Doc2Vec dictionaries were obtained taking as input the TREC Genomics collections, 2004 and 2006 [114]. Both consist of a corpus of 1.45 billion tokens specialized to the biomedical domain [115].

4.8. Certainty Assignment of a New, Larger Corpus

To this point, the focus of the work has been in determining the number of certainty categories, and to some degree, their nature. This was achieved through a series of manually-executed Surveys. While this is informative, it is of little utility; certainty annotation on a new corpus would need to be undertaken manually by individuals who had extensively studied our experimental results, and this is clearly impractical.

This section describes the algorithms that we have used to experimentally compare several machine learning systems with the goal of automating certainty assignments to new statements. We investigate several aspects, including whether an artificial intelligence algorithm can distinguish certainty categories, and if so, which algorithm can do so most efficiently.

Training a machine learning algorithm requires considerable sample data. According to Jason Brownlee¹³, *"No one can tell you how much data you need for your predictive modeling problem"*. The amount of data required depends on several factors, such as the complexity of the underlying connection between the input variable and the variable that we want to obtain or the complexity of the learning algorithm to inferentially learn the existing links between variables from specific patterns.

¹³ <https://machinelearningmastery.com/much-training-data-required-machine-learning/>

It is clear that our sample set of 45 publicly-categorized statements is insufficiently large to be used as a training dataset. Thus, it is necessary to create a larger set of new classified sentences; however, it is impractical to use the Survey approach to annotate these. As such, it was necessary to use our own experience in studying these statements and their three-category classification to manually annotate, based on this experience, a new larger corpus, followed by a validation of our annotations compared to those generated by the public survey process.

We started with statements from the MedScan corpus [105]. MedScan is natural language processing software that manage sentences from MEDLINE abstracts oriented to biomedical domain. We obtained 23,000 statements/clauses from this corpus. We then applied two filtering steps to retain only those sentences that are similar to those in the 45-statement corpus.

- **Hedging:** is the grammatical expression that encompasses hesitation. According to [3], it usually follows a specific structure like "This may indicate that". Therefore, we use the word "that" as our first filter, to enrich for statements that include hedging. "That" is a non-informative and very common word¹⁴. It is used as a relative and demonstrative pronoun; it is also used as a determiner. Most importantly, it is used as a conjunction, connecting names, verbs or adjectives with the following clause: e.g., "This result suggests *that*..." "We report *that*..." or even referring to other articles "It is thought *that*..."
- **Paragraph length:** Since MedScan parses abstracts, we filtered for paragraphs with a length of less than 350 words, in order to enrich our new corpus with assertions that are brief, as per our original corpus.

From these filtering steps, we arrived at a corpus of 3271 novel statements. These statements were then manually annotated by us, using the certainty categories from our surveys, based on our familiarity with the majority-rules classification of the 45 statements in each survey.

We then undertook several validations to ensure that this self-executed classification was comparable to the public classification of the original corpus. In particular, we self-classified the original 45 statements using the categories obtained from majority-rule in all

¹⁴ <https://dictionary.cambridge.org/es/gramatica/gramatica-britanica/pronouns/that>

three Surveys, and compared our classification to that of the public annotations obtained in the Surveys. To measure the difference between our responses and participant responses, we employed cosine similarity (*scikit-learn*¹⁵). Cosine similarity measures the cosine angle between two vectors. When both vectors point in the same direction, the angle is 0 and result cosine 1 (Figure 8). When the two features are in opposition to one another the result is -1. The results of these control tests are shown in Figure 22.

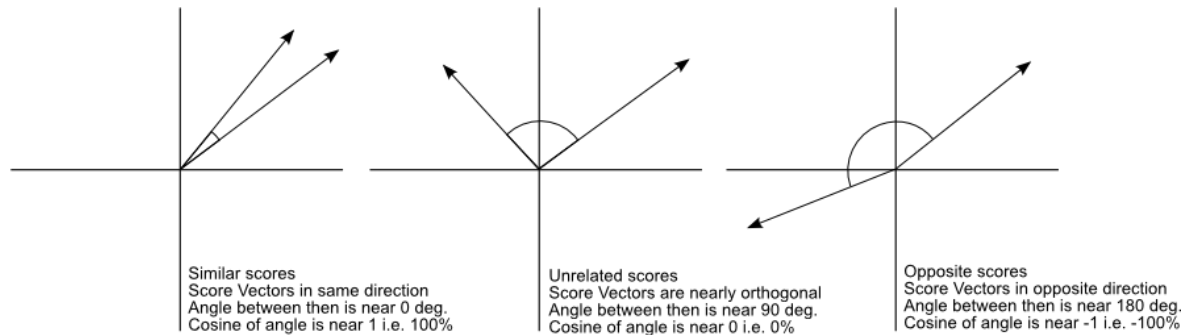


Figure 8: Cosine Similarity explanation¹⁶

We tested three approaches to machine learning, using identical sets of training and testing corpora (a subset of 250 statements from the 3271 selected above): XGBoost [106], Random Forest [107] and Neural Networks [108].

- **XGBoost:** “Extreme Gradient Boosting”, is a machine learning implementation of gradient boosted tree algorithms designed for speed and performance [109]. The parameters used to learn the classification follows the guidelines proposed by Brownlee [110] [111]. The specific parameters used were:

- **Objective:** ‘binary:logistic’ - “Specify the learning task and the corresponding learning objective”
- **Learning_rate:** 0.25 - “Boosting learning rate”
- **n_estimator:** 300 - “Number of boosted trees to fit.”
- **max_depth:** 50 - “Maximum tree depth for base learners.”

¹⁵ http://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html
¹⁶ <http://blog.christianperone.com/wp-content/uploads/2013/09/cosinesimilarityfq1.png>

- **min_child_weight:** 1 - *“Minimum sum of instance weight(hessian) needed in a child.”*
 - **subsample:** 0.5 - *“Subsample ratio of the training instance”*
 - **colsample_bytree:** 0.8 - *“Subsample ratio of columns when constructing” each tree.*
 - **scale_pos_weight:** 1 - *“Balancing of positive and negative weights.”*
 - **seed:** 27 - *“Random number seed.”*
- **Random Forest:** The parameters used to obtain the classification follow the guidelines from Koehrsen [112]. The specific parameters used were:
 - **n_estimators:** 1000 - *“Number of trees in the forest”*
 - **Criterion:** mse - Mean Squared Error. *“Function to measure the quality of a split”.*
 - **Min_samples_split:** 2 - *“The minimum number of samples required to split an internal node”*
 - **Random_state :** 42 - *“seed used by the random number generator”*
 - **Min_samples_leaf:** 1 - *“minimum number of samples required to be at a leaf node”*
 - **Neural Network:** As we described in the Machine Learning section. Neural networks is a machine learning algorithm where information is connected through nodes. It is intended to analyze and predict complex data. The specific parameters used were:
 - **Activation:** relu, sigmoid- functions that define the state of a node and its output for the next layer. “relu”: Rectified Linear Units, is the simplest non-linear function that loses the effect of backpropagation errors from other functions. “sigmoid”: a function that falls the values of the nodes between 0 and 1¹⁷.
 - **Dense:** 40, 20, 10, 2 - Number of nodes/neurons of each layer
 - **loss:** 'binary_crossentropy' - Compute the error among the output and the predicted output.

¹⁷ <http://www.machineintelligence.com/different-types-of-activation-functions-in-keras/>

- **optimizer:** 'rmsprop' - Function to change the weights of the connections¹⁸.
- **metrics:** 'accuracy' - Function to evaluate the performance of the model.

The outcome of this exploratory study showed that Neural Networks had an initial 65% accuracy, compared to 40% for Random Forest and 60% for XGBoost. As such, we decided to focus on Neural Networks, and improve on that score by tuning the hyperparameters, leading to a more accurate model. This model was used for the classification we will now describe.

4.9. Certainty Classification and Machine Learning Model

We addressed the creation of a machine-learning model by considering this task to be similar to a sentiment analysis problem, where algorithms such as Recurrent Neural Network (RNN) with Long Short Term Memory (LSTM) have been applied [113] [114] [115]. The training set data was the 3271 statements extracted from Medscan, as described earlier. A 6-layer neural network architecture was employed to train and validate model performance.

Table 4: Final machine learning model parameters.

Layer	Shape/Nodes
Word2Vec Embedding Words	200
Dense - Dropout - L1 regularizer	300 - 0.25 - 0.000001
Dense - Dropout - L1 regularizer	200 - 0.25 - 0.000001
Dense - L1 regularizer	100 - 0.000001
LSTM - Dropout - Recurrent dropout	50 - 0.1 - 0.1
Dense - Activation	3 - Sigmoid
Loss - Optimizer*	Binary crossentropy - rmsprop

*Loss and Optimizer are required parameters to compile the model.

¹⁸ <https://medium.com/data-science-group-iitr/loss-functions-and-optimization-algorithms-demystified-bb92daff331c>

The following points explain the parameters used in our model:

- **Embedding:** Word2Vec dictionary - Words are represented by single vectors of 200 length and set the weights for the embedding matrix. This dictionary was obtained taking as input the TREC Genomics collections, 2004 and 2006 [116]. Both consist of a corpus of 1.45 billion tokens specialized to the biomedical domain [117].

We used the *Gensim* python package; specifically, the class *gensim.models.Word2Vec*. The parameters used to obtain the vectorization of the words follows the guidelines from [118].

- **Size:** 200 → Word vector's dimension.
- **Alpha:** 0.05 → Initial learning rate.
- **Window:** 2 → Maximum distance between input word and context.
- **Min_count:** 5 → Minimal prevalence of words.
- **Sample:** 1e-4 → Threshold for aleatory down-sampling words with high frequency.
- **Sg:** 1 → Skip - gram model.
- **Negative:** 10 → *"negative sampling will be used, the int for negative specifies how many "noise words" should be drawn"*¹⁹.
- **Iter:** 20 → Number of Iterations.
- **Dropout:** Technique to regularize weights and reduce overfitting.
- **kernel_regularizer:** *"Regularizers allow to apply penalties on layer parameters or layer activity during optimization"*²⁰.
- **LSTM:** *"The core of the model consists of an LSTM cell that processes one word at a time and computes probabilities of the possible values for the next word in the sentence"*²¹.

Validation was executed using 20-fold CV scheme, which is considered adequate for a corpus of this size [119], [120], [121]. To design the neural network model, the Python library *Keras* [122] was utilized, with *TensorFlow*[123] as the backend. Precision, recall and overall accuracy were calculated as additional supporting evidence for classifier

¹⁹ <https://radimrehurek.com/gensim/models/word2vec.html>

²⁰ <https://keras.io/regularizers/>

²¹ <https://www.tensorflow.org/tutorials/sequences/recurrent>

performance from a confusion matrix [24], [25], [26], comprised of the following terms and formulas: True Positive (TP); True Negative (TN); False Positive (FP); False Negative (FN).

Precision measures how accurately a model predicts a value or a class.

$$Precision = \frac{TP}{TP+FP}$$

Recall measures the percentage of samples of a class that belong to that class. [124]

$$Recall = \frac{TP}{TP+FN}$$

F-score is the harmonic mean (average) of the precision and recall for a specific class.

$$F - Score = \frac{Sensitivity \times Precision \times 2}{Sensitivity + Precision}$$

Accuracy is the total percentage of the sample correctly classified. [125]

$$OverallAccuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Finally, we also employed Kappa as a commonly-used statistic to compare automated and manual adjudication [126]. Kappa was calculated using the *pym* python package.

All raw data and libraries used are available in the project GitHub, together with [*Jupyter Notebooks*](#) (both R and Python 2.7 kernels) showing the analytical code and workflows used to generate the graphs presented in this manuscript and the Supplementary Information [127].

5. RESULTS

5.1. Certainty Surveys

Survey 1: Survey 1 was started by 125 participants, of whom 75 completed 100% of the Survey (average of 13 responses per participant). 26 raters completed the Survey partially. 23 did not answer any questions and one answered Low for all statements. These latter 24 participants were removed from all subsequent analyses. All statements, except statement #13 (2.2% of total as Poor), scored at or above the minimum agreement ($G = 0.21$; “Fair” degree of agreement on the [93] scoring system). In addition, seven of 45 statements (15.5%; represented by items 7, 23, 24, 31, 32, 41, 43) obtained agreement in two certainty categories simultaneously. Table 5 and Figure 9 show the distribution of sentences among certainty categories and agreement levels.

- 11 of 45 sentences (24.4%) were classified as **High** certainty.
 - 4/11 (8.9% of total) obtained **Fair** agreement [0.21 - 0.40]
 - 3/11 (6.6% of total) showed **Moderate** agreement [0.41 - 0.60]
 - 3/11 (6.6% of total) obtained **Substantial** agreement [0.61 - 0.80]
 - Only 1 of 13 (2.2% of total) reached **Almost Perfect** agreement [0.81 - 1.00].
- 14 of the 45 (31.1%) sentences obtained agreement for the level of certainty **Medium High**.
 - 8/14 (17.7% of total) achieved **Fair** agreement.
 - 5/14 (11.11%) obtained **Moderate** agreement
 - One statement (2.2% of total) achieved **Substantial** agreement.
- **Medium Low** is represented by 12 of 45 statements (26.6%).
 - 9 of 12 statements obtained **Fair** agreement (20% of total).
 - Finally, three statements (6.6%) obtained **Moderate** agreement.
- There was no agreement for any statement at the **Low** certainty level.

Table 5: Categorization Consistency of Statements (by Statement number) for Survey S1

Agreement Level	High	% of Corpus	Medium High	% of Corpus	Medium Low	% of Corpus	Low	% of Corpus
Almost Perfect [0.81-1.00]	29	2.2%	0	0%	0	0%	0	0%
Substantial [0.61-0.8]	25, 27, 30	6.6%	5	2.2%	0	0%	0	0%
Moderate [0.41-0.6]	4, 28, 42	6.6%	19, 35, 37, 40, 45	11.11%	21, 36, 44	6.6%	0	0%
Fair [0.21-0.4]	3, 15, 22, 38	8.9%	2, 8, 9, 16, 17, 20, 34, 39	17.7%	1, 6, 10, 11, 12, 14, 18, 26, 33	20%	0	0%
Poor [≤ 0.2]	13	2.2%						
Double Classified	41, 43	4.4%	7, 23, 24, 31, 32, 41, 43	15.5%	7, 23, 24, 31, 32,	11.11%		

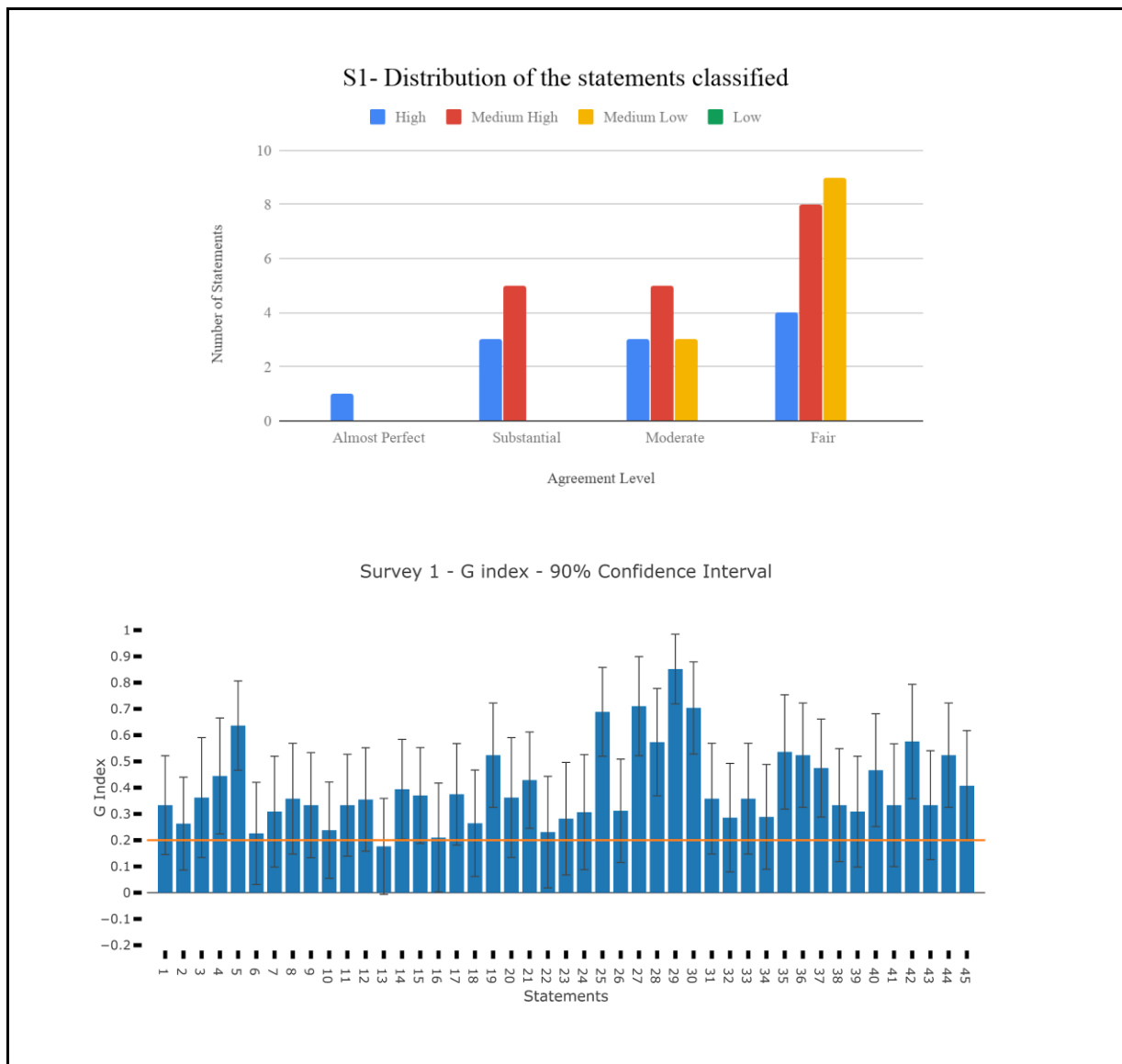


Figure 9: Distribution of the statements classified and G index agreement in S1. The orange line in the lower panel of the figure represents the minimum agreement necessary to consider.

Survey 2: Survey 2 had 246 raters, with 150 completing the Survey (average of 16 responses per participant). 65 partial responses and 31 unanswered. Participants who provided no answers were excluded from the analysis. Again, participation was anonymous, and no demographic information was collected. Disposition of Certainty categories and agreement levels are shown in Table 6 and Figure 10. Seven of the totals of 45 statements (15.5%; items 2, 7, 8, 17, 20, 26, 35) did not reach significant agreement for any Confidence/Certainty level.

- **Relatively High** was selected for 19 of 45 statements (42.2%).
 - three of 19 obtained **Fair** agreement (6.6% of total).
 - **Moderate** agreement was observed for three statements (6.6% of total).
 - Seven statements (16% of total) acquired **Substantial** agreement.
 - Agreement for **Almost Perfect** (13% of total) was achieved by six statements
- The remaining statements (42.2% of total) were selected as **Relatively Low**.
 - 10/19 (22.22 % of total) reached **Fair** agreement.
 - 7/19 statements (15.5 % of the total) obtained **Moderate** agreement.
 - Two statements achieved **Almost Perfect** agreement (2/19, 4.4% of total).

Table 6: Categorization Consistency of Statements (by Statement number) for Survey S2

Agreement Level	Relatively High	% of Corpus	Relatively Low	% of Corpus
Almost Perfect [0.81-1.00]	25, 27, 28, 29, 30, 41	13.32%	36, 44	4.4%
Substantial [0.61-0.8]	3, 15, 22, 38, 40, 42, 43	15.5%	0	0%
Moderate [0.41-0.6]	5, 6, 9,	6.6%	10, 11, 14, 18, 31, 33, 39	15.5%
Fair [0.21-0.4]	4, 37, 45	6.6%	1, 12, 13, 16, 19, 21, 23, 24, 32, 34	22.2%
Poor [≤ 0.2]	2, 7, 8, 17, 20, 26, 35	15.5%		
Double Classified	0	0%		

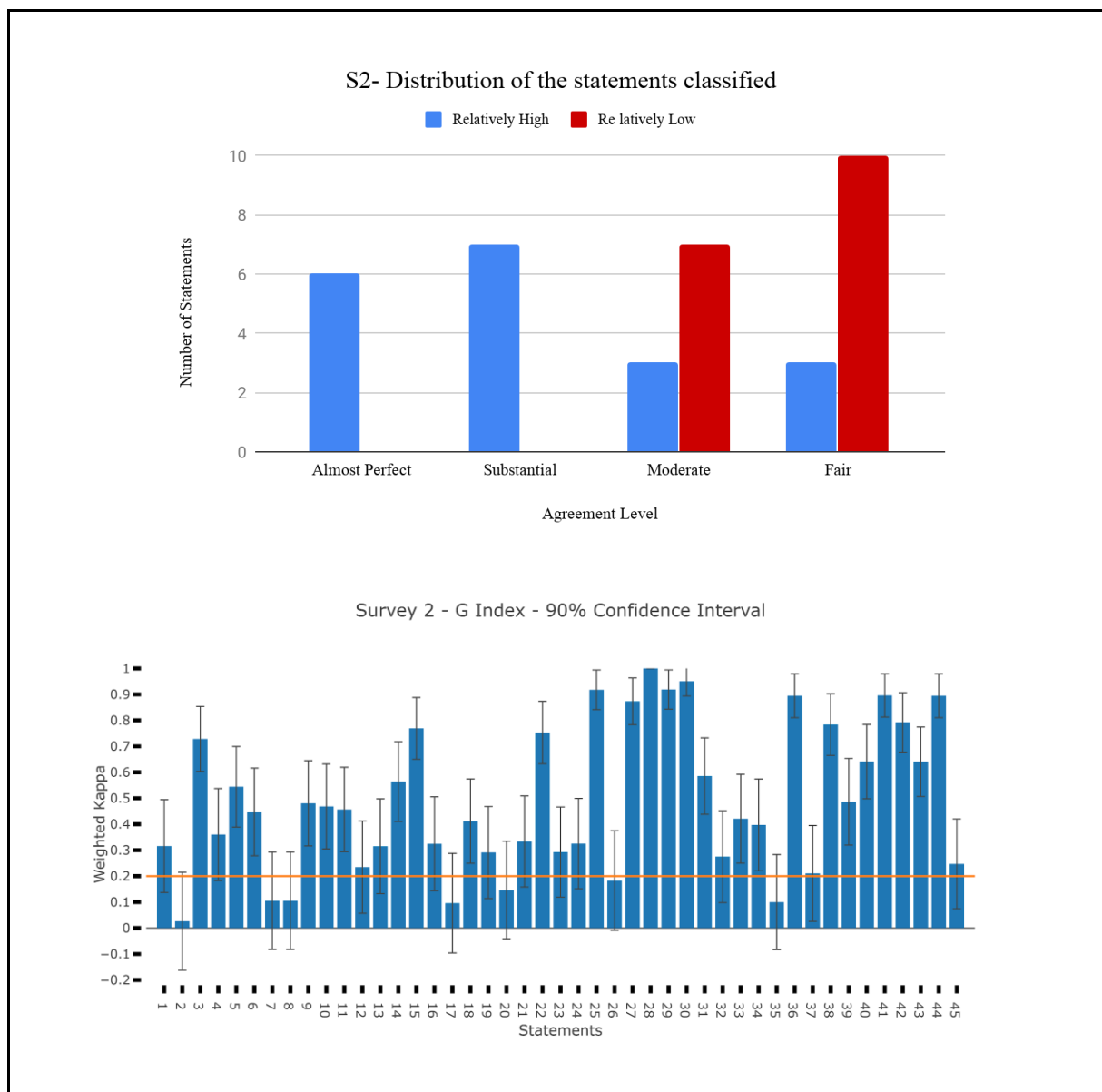


Figure 10: Distribution of the statements classified and G index agreement in S2. The orange line in the lower panel of the figure represents the minimum agreement necessary to consider.

Survey 3: Table 7 and Figure 11 summarize the levels of agreement and certainty classifications observed in Survey 3. 48 of 78 evaluators finished the entirety of Survey 3 (average of 18 responses per participant). 21 submissions had no answer and nine completed S3 partially. Evaluators who provided no responses were excluded from the study. Categories were labelled numerically with the category “1” representing the highest level of certainty, “2” representing some intermediate certainty, and “3” the lowest level of certainty. Minimum agreement ($G = 0.21$) or superior was observed in 41 of 45 statements (91.1%) with no doubly-classified statements, indicating little to no annotator perception of overlap between the presented categories. Four of 45 statements (8.8%) did not obtain agreement for any certainty category.

- **Category 1** was selected for 13 of 45 statements that compose the Survey.
 - **Fair** agreement is represented by one of 13 statements (2.2% of the total).
 - 5/13, (~11% of total), achieved **Moderate** agreement.
 - 5/13 (~11% of total) obtained **Substantial** agreement.
 - Finally, 2/13 (~4% of total) were **Almost Perfect** agreement.
- 24 of 45 (53.3% of total) were selected with **Category 2**.
 - **Fair** agreement was obtained for 14/24 (31.1% of total).
 - **Moderate** agreement for 10/24 (22.2% of total).
- **Category 3** was selected for four out of 45 sentences (8.8% of total).
 - All with **Fair** agreement.

Table 7: Categorization Consistency of Statements (by Statement number) for Survey S3

Agreement Level	Category 1	% of Corpus	Category 2	% of Corpus	Category 3	% of Corpus
Almost Perfect [0.81-1.00]	3, 15, 27	4.4%	0	0%	0	0%
Substantial [0.61-0.8]	28, 29, 38, 42	11.11%	0	0%	0	0%
Moderate [0.41-0.6]	4, 25, 30, 41, 43	11.11%	2, 16, 17, 23, 26, 33, 34, 35, 37, 40	22.22%	0	0%
Fair [0.21-0.4]	22	2.2%	1, 6, 8, 9, 10, 11, 12, 13, 18, 19, 20, 31, 32, 45	31.11%	21, 24, 36, 44	8.8%
Poor [≤ 0.2]	5, 7, 14, 39	8.8%				
Double Classified	0	0%				

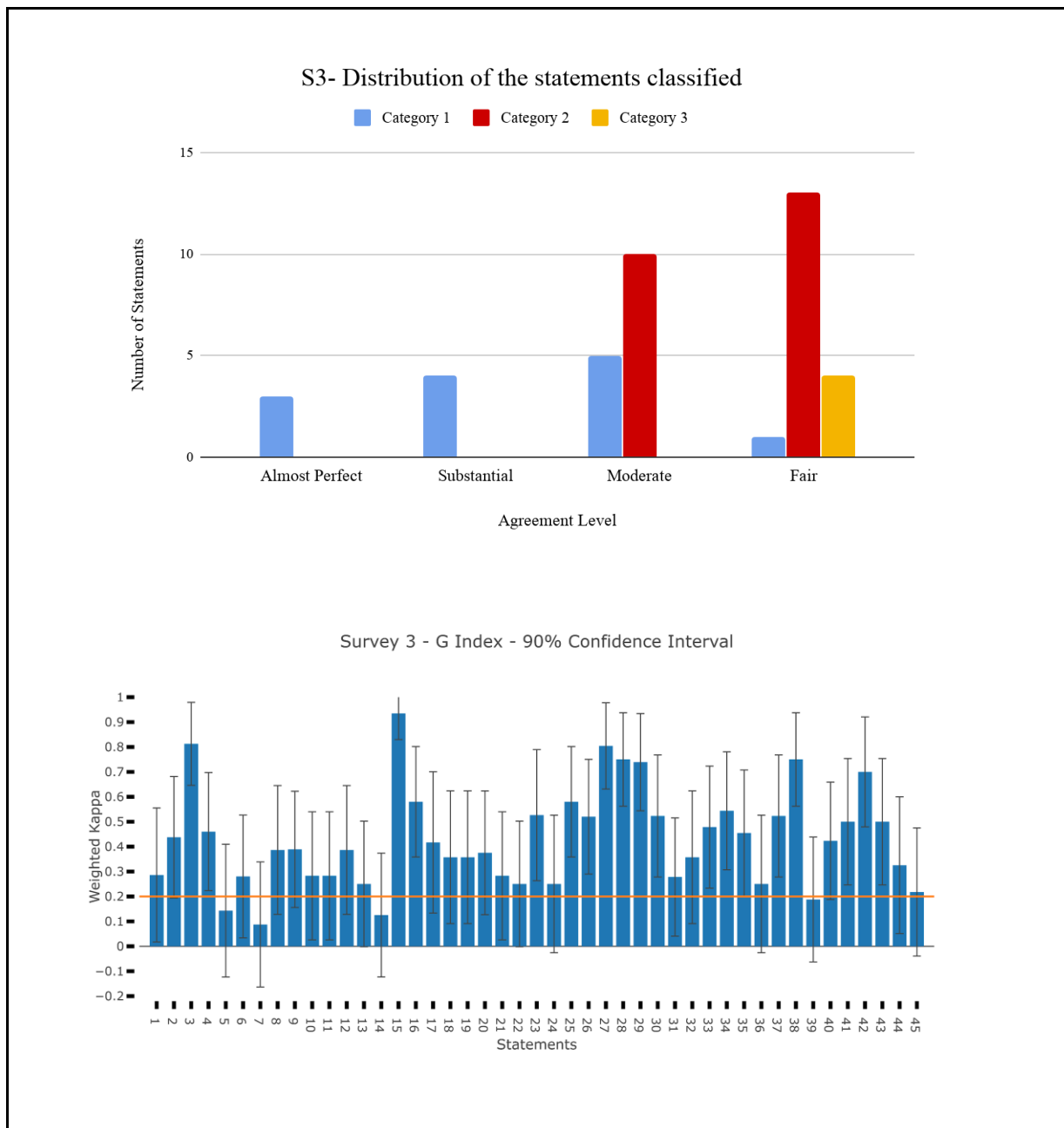


Figure 11: Distribution of the statements classified and G index agreement in S3. The orange line in the lower panel of the figure represents the minimum agreement necessary to consider.

5.2. Clustering of Statements by Survey Results

The Spearman correlation coefficient considers the weight and direction of the relationship between two variables, while Hierarchical clustering analysis (HCA) discovers clusters of similar elements.

Table 8: Interpretation of Spearman correlation [87,101].

Very Low	Low	Moderate	High	Very High
≤ 0.2	≤ 0.5	≤ 0.7	≤ 0.9	> 0.9

As shown in Fig.12, HCA and Spearman correlation rank revealed three primary clusters.

- The first branch of the HCA, left-top side of Fig.12 clustered S3-1, S1-High and S2-Relatively High indicating these categories show significant Spearman correlation ($r = 0.81$, $p\text{-value} = 2.3\text{e-}11$ S1-High/S2-Relatively High; $r = 0.72$, $p\text{-value} = 3.2\text{e-}8$ S1-High/S3-1; $r = 0.79$, $p\text{-value} = 1.3\text{e-}10$ S2-Relatively High/S3-1).
- The Second branch of the hierarchical tree is split again into two main sub-trees, including the center and right sides of the Figure.
 - The central cluster in Fig.12, differentiated by excellent Spearman correlation, contains S1-Medium Low, S2-Relatively Low and S3-3 categories ($r = 0.78$, $p\text{-value} = 2.5\text{e-}10$ S1-Medium Low/S3-3; $r = 0.81$, $p\text{-value} = 1.3\text{e-}11$ S2-Relatively Low/S3-3; $r = 0.83$, $p\text{-value} = 1.5\text{e-}12$ S1-Medium Low/S2-Relatively Low).
 - Finally, the smaller cluster on the right side of Fig.12 comprises S1-Medium High and S3-2 with moderate Spearman correlation ($r = 0.55$, $p\text{-value} = 7.8\text{e-}5$), confirming that a third certainty category has sufficiently strong support.

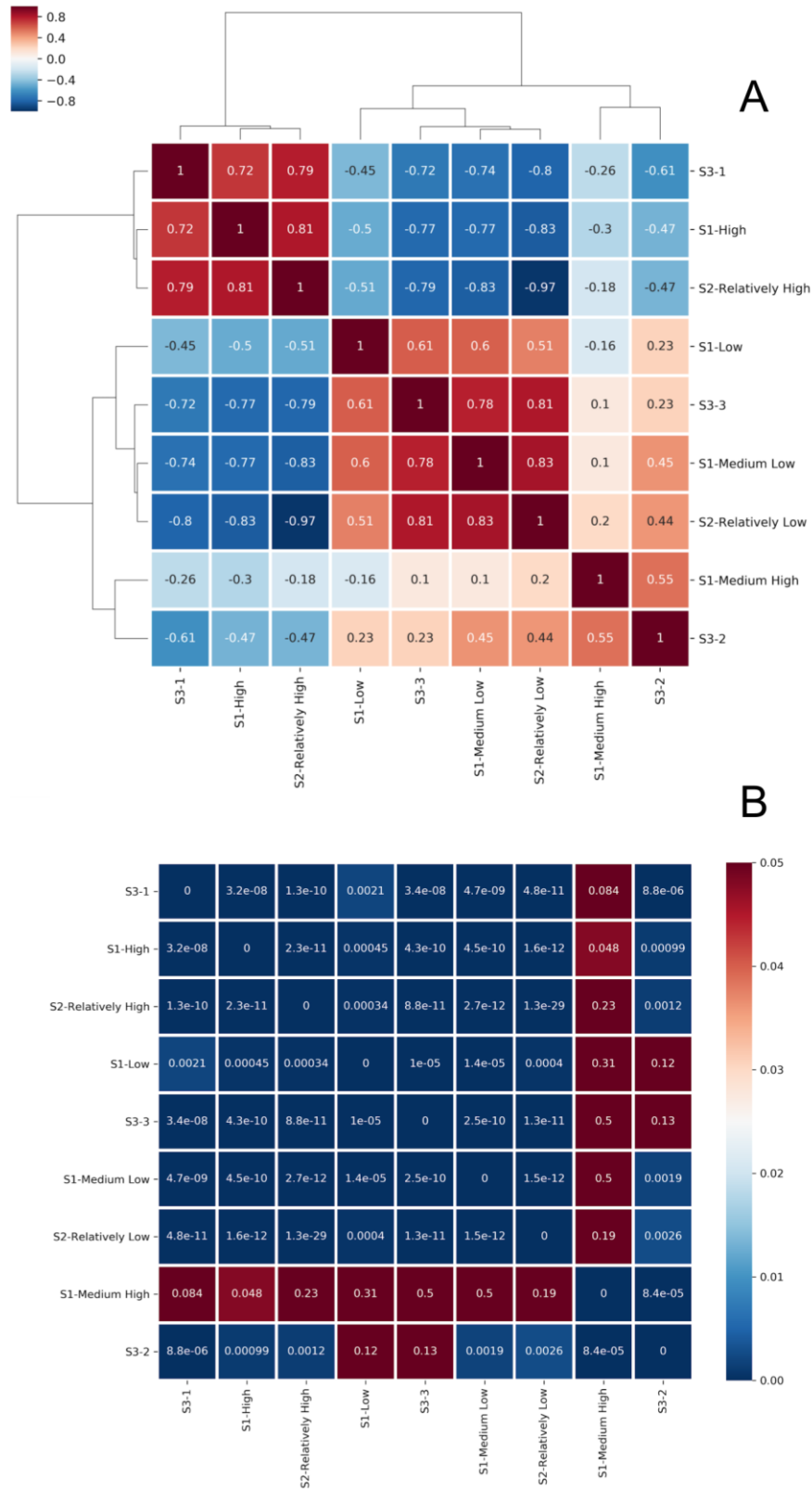


Figure 12: Spearman Rank Correlation, hierarchically-clustered heatmap (A) and their respective *p-values* (B) comparing the statements assigned to the Certainty Categories among all three questionnaires. The clustering tree and heatmap are based on participant's responses from questionnaires S1, S2 and S3.

Supporting previous cluster testing, according to the majority rule, NbClust results

(Fig.13) indicate that:

- 16 indices proposed three as the best number of clusters for the results of S1 (Fig.13.A)
- 11 indices proposed two as the best number of clusters for the results of S2 (Fig.13.B)
- 6 indices proposed three as the best number of clusters for the results of S2 (Fig.13.B)
- 14 indices proposed three as the best number of clusters for the results of S3 (Fig.13.C)

Figure 13 shows the optimal number of clusters for each Survey. Note that, surprisingly, the second-most optimal number of clusters for Survey 2 was three (Fig. 13.B), despite S2 having only two possible responses. This will be examined further in the Discussion section.

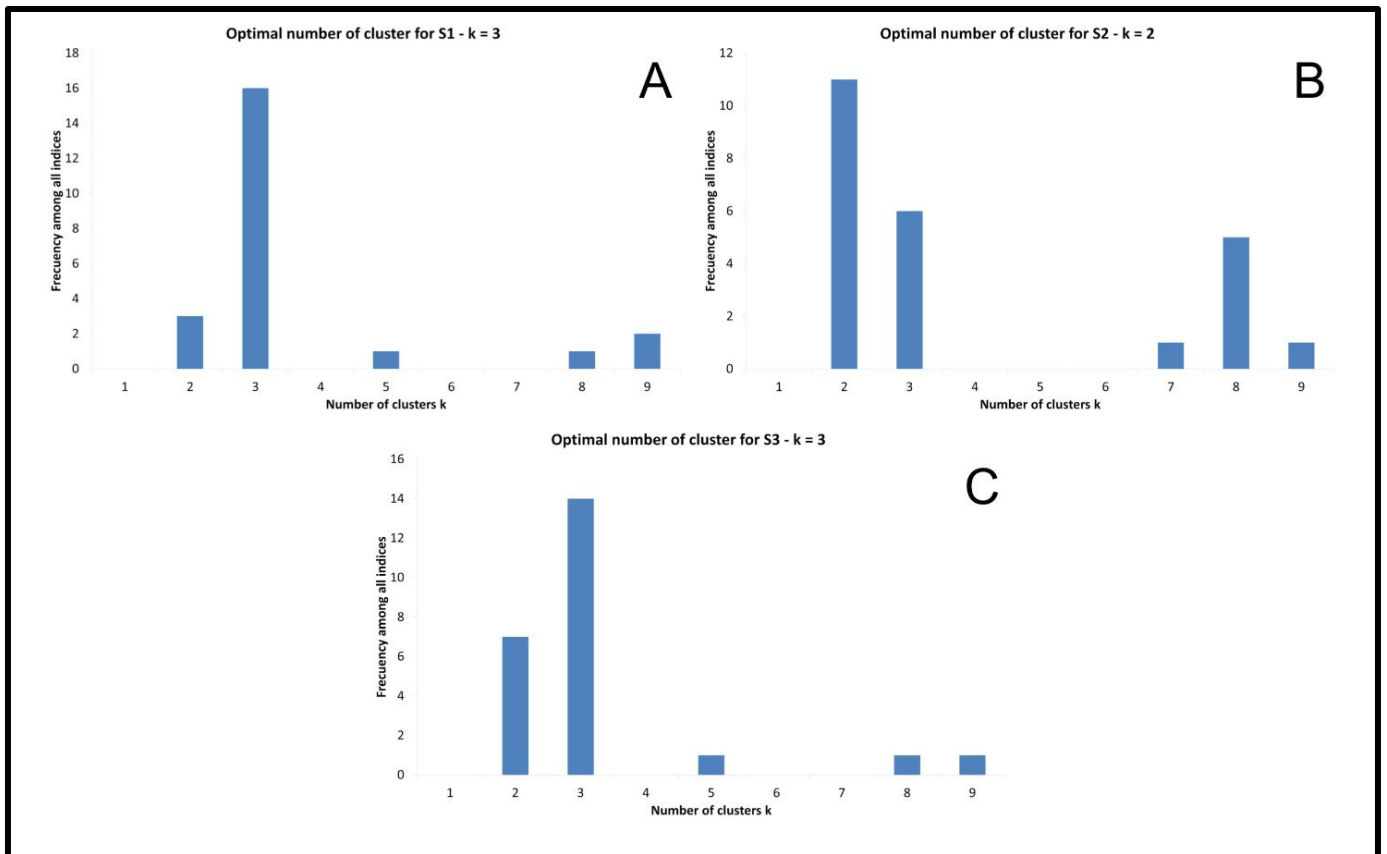


Figure 13: Majority rule output for deciding the optimal number of clusters (k) in the three Surveys. (A) Majority rule indicates three clusters for Survey 1. (B) Majority rule indicates two clusters for Survey 2, though there is notable support for three clusters. (C) Majority rule indicates three clusters for Survey 3, with notable support for two clusters.

Table 9: Jaccard Similarity clusters resulting from K-Means applied to questionnaire results.

Jaccard similarity index on k-means results from scaled questionnaire responses. The score is the result from statements' labels pairwise comparison. A dash indicates that it is not possible to compare due to differing cluster size.

	S1-S2	S1-S3	S2-S3
Cluster 1-1	0,923	0,923	0,786
Cluster 1-2	-	-	-
Cluster 1-3	0	0	-
Cluster 2-1	-	-	-
Cluster 2-2	0,474	0,737	0,833
Cluster 2-3	-	-	-
Cluster 3-1	0	0	-
Cluster 3-2	-	-	-
Cluster 3-3	0,846	0,692	0,684

The Jaccard similarity index showed measures, in terms of statement composition, that exceed 0.68 when comparing the same cluster between surveys S1, S2 and S3, with the exception of cluster 2 between S1 and S2, that shows a Jaccard similarity index of 0.47.

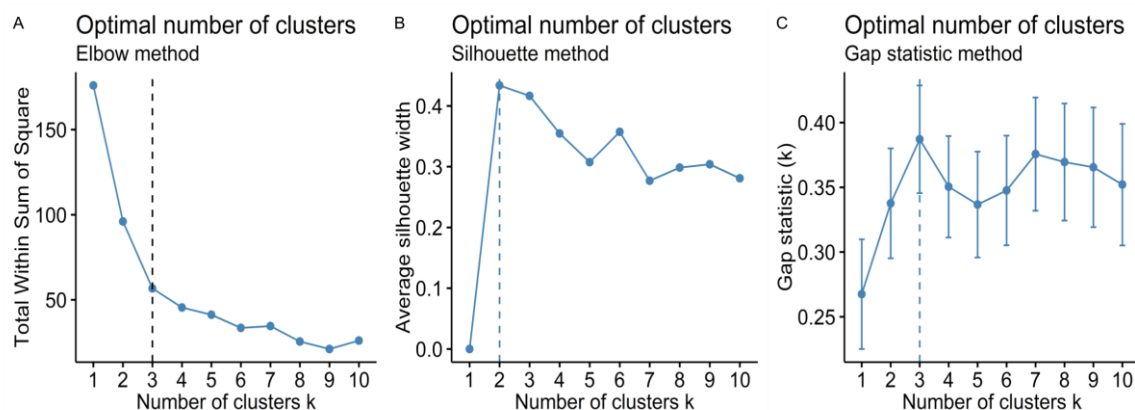


Figure 14: Clustering analysis of k-means results for S1 using Elbow, Silhouette and GAP statistic method. Dotted lines represent the optimal cluster chosen by each method.

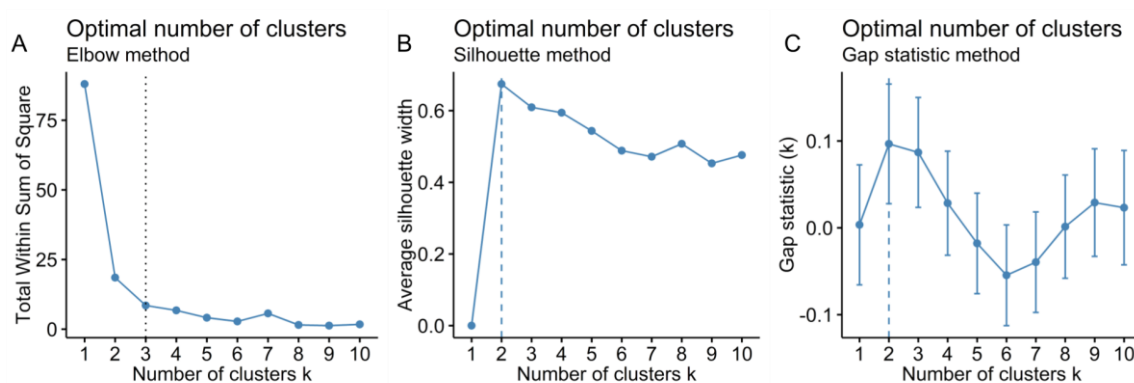


Figure 15: Clustering analysis of k-means results for S2 using Elbow, Silhouette and GAP statistic method. Dotted lines represent the optimal cluster chosen by each method.

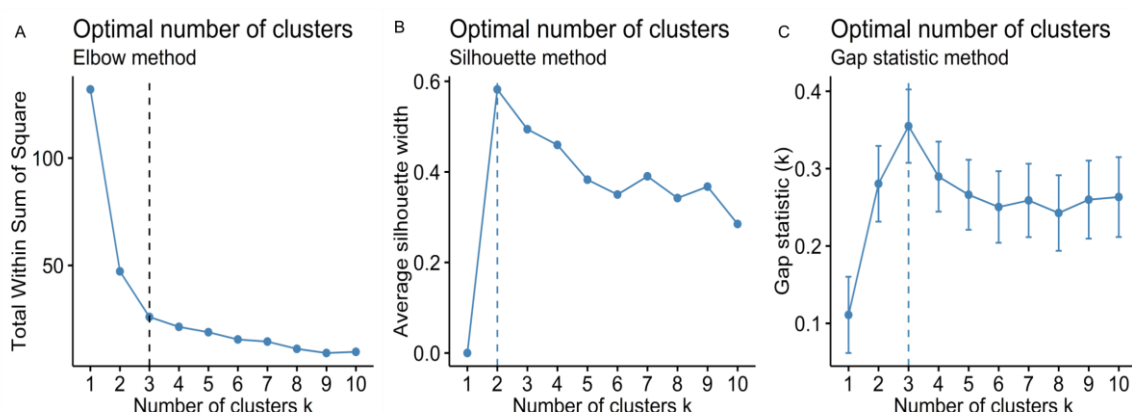


Figure 16: Clustering analysis of k-means results for S3 using Elbow, Silhouette and GAP statistic method. Dotted lines represent the optimal cluster chosen by each method.

More in-depth analysis based on the study of k-means results, shows that the optimal number of clusters is three. Using the Elbow method (identifying the point at which there

is a maximal change in the rate of decrease – the “elbow”), Figures 14, 15 and 16 (Survey 1, 2 and 3 respectively) present the values of the number of optimal clusters for k-means as $k = 3$. In addition, according to the GAP Statistics criteria, the most suitable number of clusters is three (Figure 14 and 16, Survey 1 and 3), since in choosing the optimal K, we by necessity must take the highest value of the GAP rate. Finally, Silhouette criteria is a widely used method for selecting the optimal number of clusters. It is a more conservative method that measures how well each sample fits into the k-means clusters. In the three cases, the Silhouette criterion obtained a value of $k = 2$.

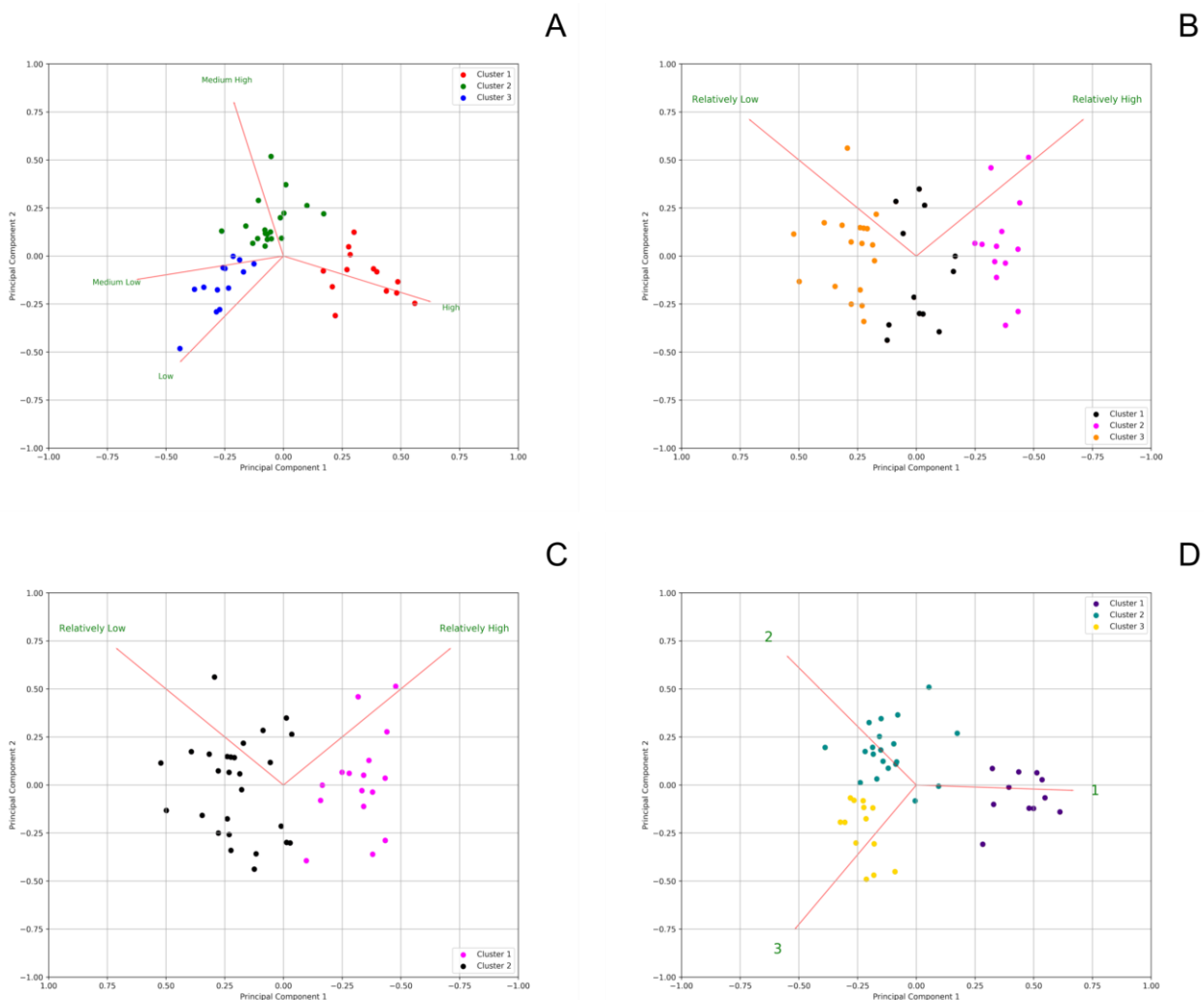


Figure 17. Principal component analysis of responses to Question 1.1 in the three Surveys.

Bi-plot of certainty level distribution over results from k-means clustering (colors) for: Survey 1 (A), Survey 2 with three clusters (B), Survey 2 with two clusters (C) and Survey 3 (D). Each dot represents a statement. Red lines are the eigenvectors for each component

Table 10: Analysis of Principal Components of Question 1.1 for S1 Statement Classifications

Principal Components:	Comp.1	Comp.2	Comp.3	Comp.4
High	0.620	-0.235	0.185	0.725
Medium High	-0.210	0.795	0.473	0.317
Medium Low	-0.618	-0.121	-0.480	0.611
Low	-0.436	-0.546	0.716	0.013
Component variances	2.238	1.275	0.382	0.105
Standard deviation	1.496	1.129	0.618	0.324
Proportion of Variance	0.560	0.319	0.095	0.026
Cumulative Proportion	0.560	0.878	0.974	1.000

Standard Deviation (row 1), Proportion of Variance (row 2) and Cumulative Proportion (row 3) are summarized in Table 10 for S1, for each principal component. Table 10 additionally supplies the information to explain each component and its relative weighting, requisite to understanding all components. The first three components explain 97% of the variance of the data. Fig. 17A shows the graph resulting from a principal component analysis (PCA) of Q1 responses to statements from S1, clustered by K-Means (colored dots). Red lines represent the eigenvectors of each variable (here the certainty categories) for PC1 against PC2. A coefficient close to 1 or -1 indicates that variable strongly influences that component. Thus, the High category has a strong influence on PC1 (0.62), Medium High primarily influences PC2 (0.79), and Medium Low and Low have a notably strong negative relationship with PC1 (-0.62 and -0.44, respectively). Additionally, Low influences PC2 in a negative relationship (-0.54). The same approach was followed for S2 and S3, with the results shown in Fig. 17B, C, and D (tabular data for PCA of S2 and S3 not shown). For Survey S2, we show the K-Means clustering results for both a three-cluster solution (Fig. 17B), and a two-cluster solution (Fig. 17C).

5.3. Comparison Between Surveys

Considering now the second part of the Survey - that is, the statement-ranking question (Question 2) - the objective was to provide a second, distinct way of capturing perceived certainty in order to compare it to the results of Question 1 as a form of internal control, but

also to ensure that comparison between the three Surveys is valid. By presenting statements using a different approach to statement evaluation (ranking) and a different interface (drag-and-drop), that is shared among all three Surveys, Question 2 can corroborate that participants answer consistently; effectively, if respondents are answering consistently, then a statement categorized as high certainty should consistently appear “above” a low certainty statement in a ranking test, and this should be true over all three Surveys. We had concerns about the design of Question 2 (Q2) for Survey 1 given its task of being an internal control. The error lies in our pre-selection of groups of five statements to be presented to the respondent for the ordering task, which would almost certainly introduce bias, and would fail to test all-against-all (v.v. statements) as is necessary for an adequate internal control. As such, while this pre-grouping was necessitated by the limitations of the questionnaire platform used in that Survey, all data from Q2 of S1 will be ignored for the purpose of analysis. Nevertheless, as they remain somewhat informative, the data from S1 Q2 will be included in some of the visualizations and tables shown below.

To determine if there were significant differences in the responses obtained between the different Surveys (and different respondent populations), a Wilcoxon and/or Kruskal-Wallis (K-W) test was performed to compare between two groups, or three groups, respectively.

- H_0 = Population response distribution is the same for S2 and S3. Same population.
- H_1 = Population responses distribution are NOT the same for S2 and S3. Different populations.

Table 11: Analysis of differences between groups using Kruskal-Wallis (> 2 groups) and U-Mann Whitney test (2 groups)

Question 1 of Surveys:	$\chi^2 // W$	Degrees of Freedom	P-value
S1-S2	1189	-	0.1544
S1-S3	1249	-	0.0563
S2 - S3	957.5	-	0.6572
S1 - S2 - S3	3.7343	2	0.1546
Question 2 of Surveys:	$\chi^2 // W$	Degrees of Freedom	P-value
S2 - S3	1004	-	0.9453

The result of the Wilcoxon test for Question 1 between S1 and S2 is $p = 0.154$, for S1 and S3 is $p = 0.056$ and for S2 and S3 is $p = 0.657$. Likewise, the result for Question 2 between S2 and S3 is $p = 0.945$. The result of the Kruskal-Wallis test when we compared Question 1 of the three Surveys is $p = 0.154$. As all p -values are higher than 0.05, there is insufficient evidence to reject the null hypothesis. Therefore, we accept H_0 that all three Surveys are being answered as if from the same population. Specifically, the results of the tests for Question 1 show that the different respondent populations follow the same tendency when classifying each statement, despite being offered a different number of levels of certainty in each Survey. Moreover, when each group responded to an identical question (Question 2 - rank five statements), the difference in the response tendency among all populations becomes smaller (p -value > 0.9), indicating a very close match in response behaviors throughout all three Surveys.

As a *post-hoc* test to corroborate these results, and to avoid any kind of bias from using the result of the *Relative Importance Index* score in the KW and Wilcoxon test, we performed an *Independence test* for Question 2 between Surveys 2 and 3.

Table 12: Independence Test

Question 2 of Surveys:	maxT	Degrees of Freedom	<i>P</i> -value
S2-S3	15.297	-	$< 2.2e-16$

- H_0 or "Null hypothesis" = The responses to the questionnaires are independent.
- H_1 or "Alternative hypothesis" = The answers to the questionnaires are not independent.

The result of the independence test is a p -value much less than 0.05, and therefore the null hypothesis is rejected. That is, the answers from S2 and S3 are not independent, and therefore the respondents are behaving nearly identically. This is consistent with the previous result and confirms that the statistical analysis performed using RII score did not bias the obtained result.

5.4. Comparison Between Questions

Table 13: Analysis of differences between question 1 and question 2 using U-Mann Whitney test (2 groups)

Question 1 vs. Question 2	χ^2	Degrees of Freedom	<i>P-value</i>
S1	1344	1	0.0074
S2	1096.5	1	0.4979
S3	1148	1	0.2741

The null hypothesis in this case is that the median of the populations, when comparing Q1 against Q2, are the same. Since the p-value is > 0.05 for both S2 and S3 we do not reject the null hypothesis and conclude that the average distribution of answers to both questions are the same in S2 and S3 (we also note that the p-value from S1 supports our belief that we introduced significant bias into S1 Q2 by pre-grouping the sets of five statements, and therefore supports our exclusion of this data from the analysis). We can also conclude that the answers to both Question 1 and Question 2 were not selected randomly by the Survey respondents.

Our findings demonstrated that results from S2 and S3 are statistically comparable responses. Following this, we decided to merge S2 and S3 results to simplify analysis. We then applied Relative Importance Index and arranged them from highest to lowest certainty (column S2-S3 of Figure 18). Figure 18 shows the most suitable correlation between statements and certainty category. It revealed that the set of statements that are at the top of column S2-S3, have a high correlation with the categories of "high" certainty (High, Category 1, Relatively High). While those statements located in the lower range of column S2-S3, have high correlation with the categories of "low" certainty

The results show that participants did appear to be acting consistently, since a similar distribution is obtained from the three questionnaires when comparing Q1 (certainty classification) with Q2 (ordering by certainty).

5.5. Evaluating the respondents indication of ‘basis’

Question 1.2 asked the readers to indicate what they felt was the ‘basis’ for each of the statements. For example, was it speculation, or was it supported by direct evidence? Though this, we wished to examine the connection between perception of certainty, and the perceived basis of argumentation the reader felt they were being exposed to.

With respect to Question 1.2, we first need to determine the degree of agreement between annotators. In this case, unlike the certainty portion of the Survey, we set-forth that the categories were to be one of: Direct Evidence, Indirect Evidence/Reasoning, Speculation, Citation, or Unknown.

Survey 1: All statements obtained the minimum agreement ($G = 0.21$), except for statements #35 and #45 (4.4% as Poor). Nine of 45 statements (20%) obtained inter-annotator agreement in Direct Evidence. Seven of 45 (15.5%) were classified as Indirect Evidence/Reasoning; 11 of 45 statements (24.4%) were categorized as Speculation and 10 (22.2%) as Citation. Additionally, six of 45 (13.3%) revealed inter-annotator agreement in two categories simultaneously. Basis distributions are shown in Table 14.

Table 14: Responses to the question of “Basis” in S1

Agreement Level	Direct Evidence	% of Corpus	Indirect Evidence or Reasoning	% of Corpus	Speculation	% of Corpus	Citation	% of Corpus	I don't know
Almost Perfect [0.81-1.00]	27, 28	4.4%	0	0%	44	2.2%	30	2.2%	0
Substantial [0.61-0.8]	25	2.2%	0	0%	14, 24, 36, 39	8.88%	29, 33	4.4%	0
Moderate [0.41-0.6]	3, 15, 37, 38	8.8%	16, 34	4.4%	1, 10, 11, 13, 21	11.11%	23, 26, 40, 43	8.8%	0
Fair [0.21-0.4]	4, 5	4.4%	6, 9, 18, 19, 20	11.11%	12	2.2%	2, 7, 17	6.6%	0
Poor [≤ 0.2]	35, 45	4.4%							
Double Classified	8, 22, 41, 42	8.8%	8, 31, 32, 41	8.8%	31, 32	4.4%	22, 42	4.4%	

Survey 2: Table 15 summarize basis classification observed in S2. Six of the 45 statements achieved inter-annotator agreement in two categories simultaneously (13.3%). Direct Evidence was selected in six of the 45 statements (13.3%). We found agreement for 11 of 45 statements (24.4%) in Indirect Evidence / Reasoning. 10/45 (22.2%) were chosen as Speculation. Citation was selected for 10 of 45 statements (22.2%). The two residual statements (4.4%) did not achieve any level of agreement.

Table 15: Responses to the question of “Basis” in S2

Agreement Level	Direct Evidence	% of Corpus	Indirect Evidence/ Reasoning	% of Corpus	Speculation	% of Corpus	Citation	% of Corpus	I don't know
Almost Perfect [0.81-1.00]	27,28	4.4%	0	0%	44	2.2%	0	0%	0
Substantial [0.61-0.8]	3, 25	4.4%	0	0%	36	2.2%	29, 30, 33	6.6%	0
Moderate [0.41-0.6]	15	2.2%	16,20, 34	6.6%	1, 11, 13, 14, 21, 24, 39	15.55%	22, 26, 40, 42	8.8%	0
Fair [0.21-0.4]	38	2.2%	5, 6, 8, 9, 17, 19, 37, 45	17.77%	35	2.2%	2, 7, 23	6.6%	0
Poor [≤ 0.2]	4, 12	4.4%							
Double Classified	41	2.2%	10, 18, 31, 32, 41, 43	13.33%	10, 18, 31, 32	8.8%	43	2.2%	

Survey 3: Minimum agreement ($G = 0.21$) or superior was observed in 34 of 45 statements (75.5%) with six doubly-classified statements (13.3%), indicating annotator perception of overlap between the submitted categories. Seven of 45 statements (15.5%) achieved agreement for Direct Evidence category. Indirect Evidence / Reasoning was selected for eight of 45 statements (17.7%). Seven of the totals of 45 (15.5%) were chosen using the Speculation Category. Finally, Citation was selected for eleven out of 45 sentences (24.4%). Agreement and basis disposition are shown in Table 16.

Table 16: Responses to the question of “Basis” in S3.

Agreement Level	Direct Evidence	% of Corpus	Indirect Evidence/ Reasoning	% of Corpus	Speculation	% of Corpus	Citation	% of Corpus	I don't know
Almost Perfect [0.81-1.00]	3	2.2%	0	0%	36,44	4.4%	33	2.2%	0
Substantial [0.61-0.8]	15, 25, 27, 28	8.8%	16	2.2%	13, 21, 24	6.6%	7, 26, 29, 30, 40, 42	13.33%	0
Moderate [0.41-0.6]	38	2.2%	8, 18, 19, 20	8.8%	1,11	4.4%	22, 23, 43	6.6%	0
Fair [0.21-0.4]	5	2.2%	32, 34, 45	6.6%	0	0%	2	2.2%	0
Poor [≤ 0.2]	4, 6, 12, 17, 35	11.11%							
Double Classified	37, 41	2.2%	10, 14, 31, 37, 39, 41	13.33%	10, 14, 31, 39	8.8%			

5.6. Basis and Certainty Correlation

We now attempt to identify relationships, using spearman correlation, between the results from the “certainty classification” question, and the “basis of certainty” question. Figure 19, 20 and 21 show, below the diagonal, a scatter plot (lower-left half of the diagram), with “basis” represented in the X axis, and level of certainty on the Y axis. The red line represents a LOESS curve - a method for fitting a smooth curve between the two variables²². The histograms on the diagonal represent the frequency of the variable, with the variable included in that row/column being indicated in the text above that histogram. The X axis is

²² https://www.statsdirect.com/help/nonparametric_methods/loess.htm

divided automatically by ranges and bar length indicates the numbers of statements that fit inside that range. Spearman correlation value is represented above the diagonal (top-right) following the “SPLOM” visualization paradigm²³. To simplify visual inspection, a “good” correlation is indicated by an ascending red line in one of the lower-left scatter plot diagrams, and the Spearman Correlation value of that correlation is indicated in the top-right, with the notable values highlighted. Analysis of certainty and basis correlation of Survey 1, 2 and 3, shown in Figure 19, 20 and 21 respectively, reveals moderate correlation (S1: $r = 0.65$, S2: $r = 0.78$ and S3: $r = 0.65$) between statements that exhibit High certainty (High, Relatively High and Category 1) and having a basis in “Direct Evidence”. We also found moderate and high correlation between the statements that have “low” certainty and the basis of “Speculation” ([Medium Low = 0.77; Low = 0.64] for S1, Relatively Low = 0.87 for S2 and Category 3 = 0.9 in S3). In the four previous correlations between “low” certainty and “Speculation” a clear positive trend can be observed in their respective dot chart. Finally, there is also a positive tendency between Medium High (S1) and Indirect Evidence / Reasoning ($r = 0.55$). Additionally, although it does not reach minimum levels of correlation, there is also a positive correlation between Category 2 (S3) and Indirect Evidence / Reasoning ($r = 0.39$).

23 <https://www.rdocumentation.org/packages/psych/versions/1.8.12/topics/pairs.panels>

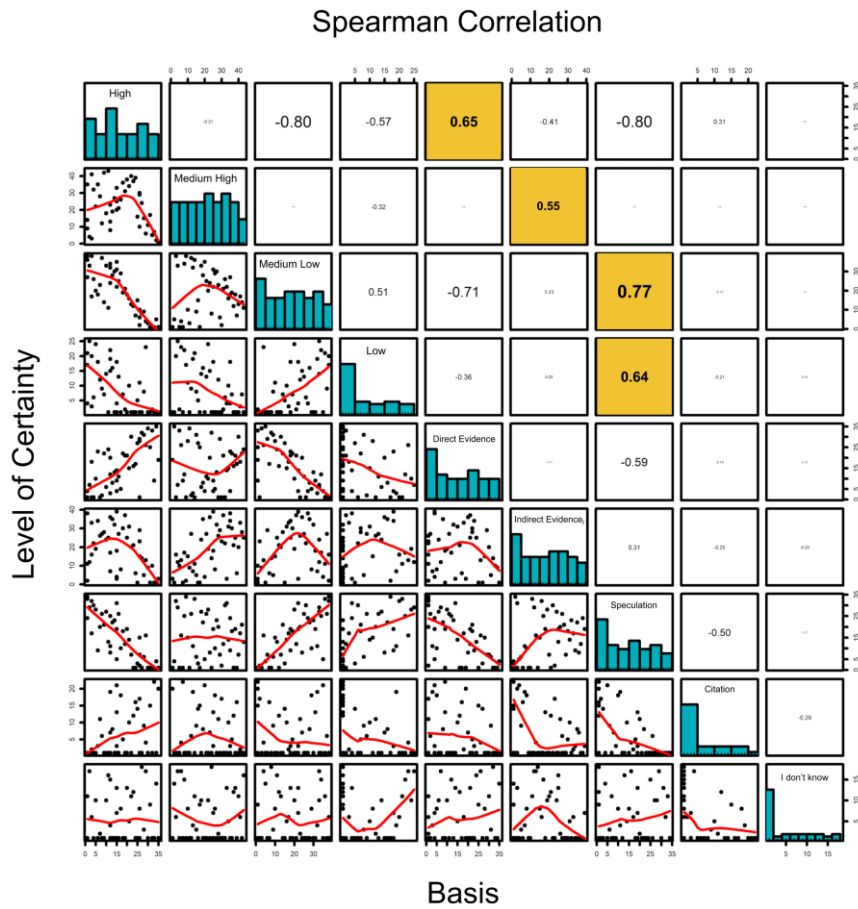


Figure 19: Basis and certainty correlation for S1.

Analysis of certainty and basis correlation for Survey 1 shows (Figure 19) moderate correlation (S1: $r = 0.65$) between High certainty and “Direct Evidence”. We also found a moderate correlation between the statements that have “low” certainty and the basis of “Speculation” (Medium Low = 0.77; Low = 0.64). In the correlations between Medium-Low-Certainty/“Speculation” and Low-Certainty/“Speculation”, a clear positive trend can be observed in their respective scatter plot chart (Figure 19). Finally, there is also a positive tendency between Medium High (S1) and Indirect Evidence/Reasoning ($r = 0.55$).

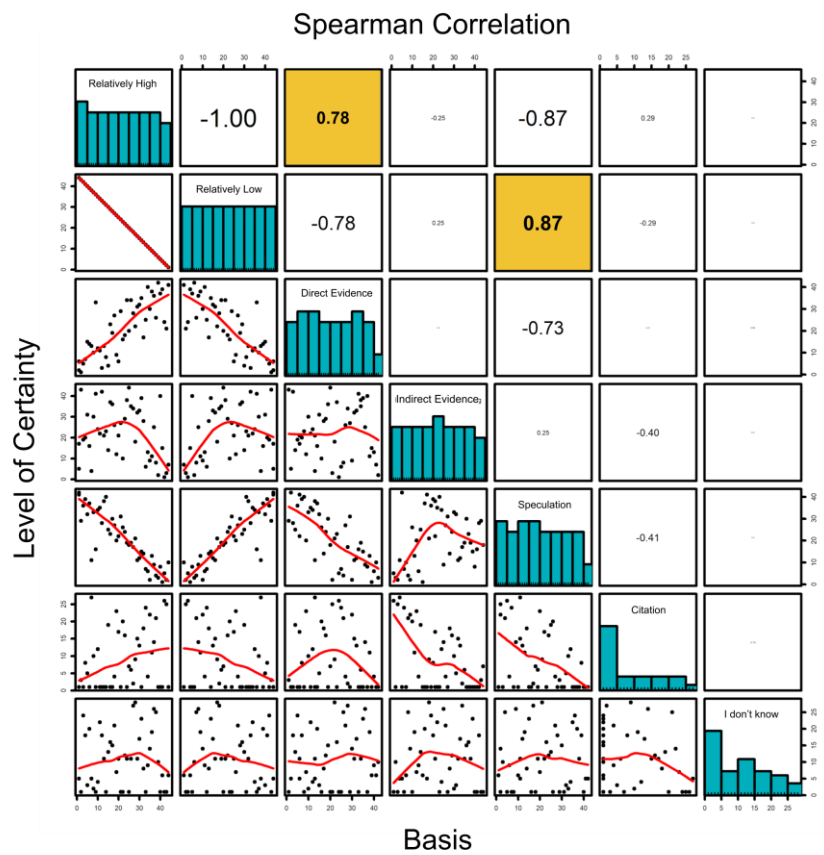


Figure 20: Basis and certainty correlation for S2.

Analysis correlation between basis and certainty of Survey 2, shown in Figure 20, displays moderate correlation (S2: $r = 0.78$) between Relatively High “Direct Evidence”. Moderate correlation is found between Relatively Low and “Speculation” basis ($r = 0.87$).

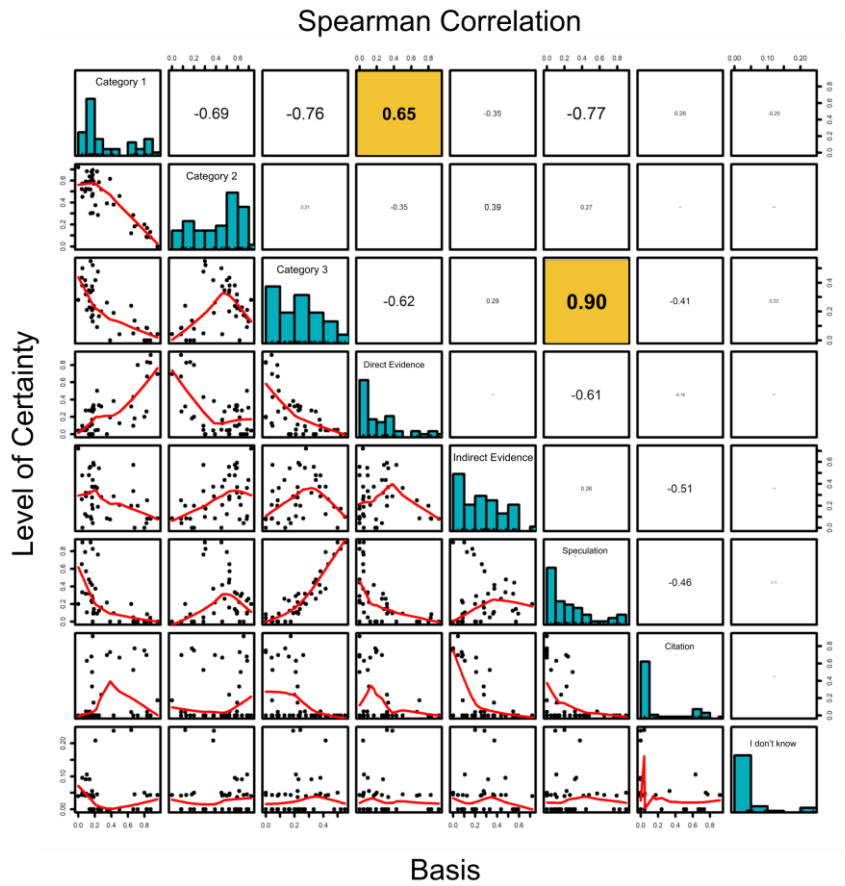


Figure 21: Basis and certainty correlation for S3.

Finally, for S3 shown in Figure 21, the analysis of certainty and basis correlation reveals moderate correlation ($r = 0.65$) between the statements of Category 1 and “Direct Evidence”. Category 3 reveals high correlation, having a basis in “Speculation” with Spearman r of 0.9. Additionally, although it does not reach minimum levels of correlation, there is also a positive correlation between Category 2 (S3) and Indirect Evidence / Reasoning ($r = 0.39$).

5.7. Identifying Reference Spans

Considerable effort – nearly one third of the duration of this thesis - was invested in an attempt to identify reference spans. Unfortunately, the results obtained were never satisfactory. All algorithms used: TF-IDF, Word2Vec, TF-IDF Embedding Vectorizer and Doc2vec, obtained an accuracy lower than 30%. Document preprocessing slightly improved the accuracy but our scores still remained below 30%. These levels of accuracy are similar to those obtained by other authors in previous and ongoing studies [128] [84]. After due consideration, we concluded that we were unlikely to make any significant advances in this domain, and moreover, that other groups were already well-along in their own investigations, and would likely always be ahead of our own efforts. As such, we terminated this aspect of the project.

Given that the main objective of this line of investigation was to automatically generate citation chains which could then be analyzed for “hedging erosion” by our machine learning model, we decided that it would be sufficient for our needs to identify a few such chains through manual investigation of the literature. As such, though the effort was extensive, we will not discuss the results any further here.

5.8. Machine Learning Model

As a first step toward creating a machine-learning model, we generated a manually-classified corpus of about 3000 statements following a methodology described in the Materials and Methods section. We validated that this self-annotated corpus was reflective of the results from the best publicly-classified corpus (Survey 3) in several ways, to ensure that the final machine-learning model would be reflective of those public annotations. In particular, we self-annotated the 45 statements presented in the three Surveys, and executed a variety of statistical comparisons to determine if our annotations were reflective of the general population.

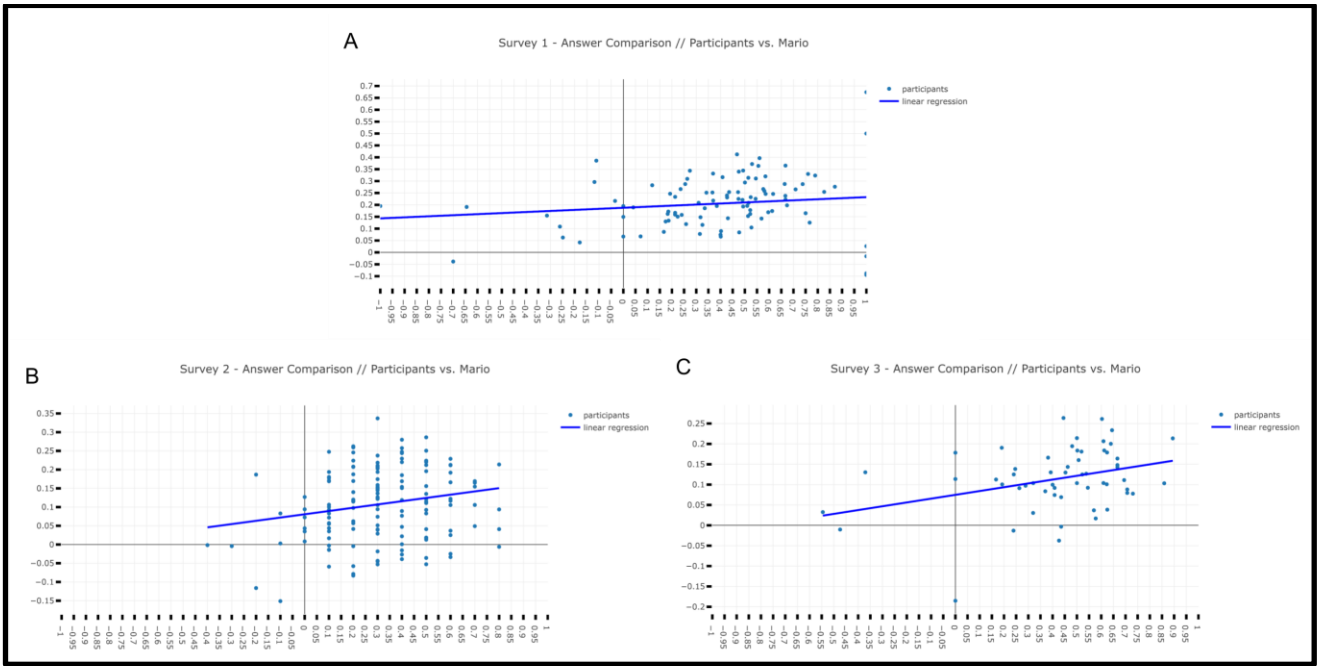


Figure 22: Comparison of the participants' responses against our manual classification for S1 (A), S2 (B) and S3 (C). The X axis shows the similarity measured using cosine similarity. Y axis shows the average of the *Relative Importance Index* score of all statements answered by a participant. Any dot in the top-right quadrant is indicative of commonality between our classification and the public classification. The regression line drawn in blue shows the trend in the level of certainty of the participants' answers increasing as the certainty level of our manual classification increases, indicating that agreement increases for more highly certain statements.

Figure 22A shows that, for S1, 89.1% of the participants classified an average of 70% of their responses in accordance with our annotations using the same categorization system. Figure 22B shows for S2 an average coincidence of 62,6% in the responses with 86,5% of the participants and in S3 an average of 64.4% of coincidence with the 94% of the participants (Figure 22C). In addition, the linear regression of the three Surveys shows that the higher the level of certainty of the statements, the greater the degree of coincidence with our classification, consistent with our suggestion that the high certainty category is much more distinct and uniformly-perceived than the lower categories. Table 17, 18 and 19 show the results of precision, recall, F-score and Overall Accuracy for S1, S2 and S3 comparing our classification performance against the median of the population. The result is an accuracy of 80% for S1, 66.7% for S2 and 85.1% for S3. Our manual classification yielded F-scores of 63.1% for S1, 66.4% for S2 and 74.8% in Survey 3.

Table 17: Performance of our manual classification for S1 vs. the publicly-annotated 45 statements

	Recall	Precision	F-Score	Overall accuracy
High	0,750	0,923	0,827	0,889
Medium	0,563	0,529	0,545	0,667
High	0,750	0,400	0,521	0,756
Medium	0,000	-	-	0,889
Low				
Mean	0,516	0,617	0,631	0,800

Table 18: Performance of our manual classification for S2 vs. the publicly-annotated 45 statements

	Recall	Precision	F-Score	Overall accuracy
Relatively High	0,722	0,565	0,634	0,667
Relatively Low	0,630	0,773	0,694	0,667
Mean	0,676	0,669	0,664	0,667

Table 19: Performance of our manual classification for S3 vs. the publicly-annotated 45 statements

	Precision	Recall	F-Score	Overall accuracy
Category 1	0,857	0,923	0,889	0,933
Category 2	0,692	0,947	0,800	0,800
Category 3	1,000	0,385	0,556	0,822
Average	0,849	0,751	0,748	0,851
Kappa	0,649			

Our answers in the three questionnaires appear to not be significantly biased, since there is a substantial group of participants (> 80%) that correspond to our manual classification. This suggests that the ~3000-statement training set we created to build the machine learning model closely reflects the perceptions of the general population. As such, it is appropriate to use it as input to create a machine learning model intended to reflect the public perception of certainty.

The machine-learning model was generated as described in the Materials and Methods section. This model was validated using a 20-fold CV due to the size of the dataset, with the result indicating $89\% \pm 1,43\%$ accuracy. None of the cross-validation folds shows overfitting by exceeding a difference greater than 0.1 between Loss and Validation Loss. A test of its performance relative to the Survey 3 majority rule classification of the original publicly-annotated 45 statements showed 82.2% accuracy (see Table 21). A further test was done to validate the author-categorized corpus, compared to this same Survey 3 dataset (see Table 19). Majority rule vs. the author's classification gave a kappa value of 0,649 (substantial), while comparison with the model's classification gave a kappa of 0,512 (moderate).

Table 20: 20-fold cross-validation results from the machine learning model.

Folds	Accuracy	Accuracy of the validation	Loss	Loss of the validation
1	0.901	0.900	0.246	0.271
2	0.896	0.892	0.259	0.269
3	0.894	0.875	0.273	0.312
4	0.899	0.873	0.256	0.314
5	0.900	0.881	0.256	0.292
6	0.900	0.900	0.256	0.262
7	0.899	0.903	0.254	0.256
8	0.899	0.890	0.260	0.269
9	0.896	0.892	0.255	0.273
10	0.900	0.894	0.258	0.268
11	0.891	0.900	0.277	0.269
12	0.891	0.898	0.273	0.237
13	0.900	0.915	0.255	0.214
14	0.903	0.904	0.248	0.273
15	0.903	0.863	0.246	0.337
16	0.898	0.888	0.262	0.276
17	0.905	0.885	0.248	0.315
18	0.898	0.879	0.262	0.302
19	0.894	0.888	0.259	0.327
20	0.895	0.854	0.261	0.340

Table 21: Performance of the neural network model on the original publicly-annotated 45 statements

	Precision	Recall	F-Score	Overall accuracy
Category 1	0,786	0,786	0,786	0,867
Category 2	0,778	0,808	0,793	0,756
Category 3	0,250	0,200	0,222	0,844
Average	0,604	0,598	0,600	0,822
Kappa	0,512			

5.9. Capturing Certainty in Formal Logics and Data Structures

To be used in FAIR data publications, it is necessary to formally encode our categorization system into an ontological framework, and publish it using a globally unique and persistent ID. Our assessment of the ORCA ontology of De Waard [129] suggested that it would be reasonable to extend this to include the certainty categories identified in this work as being the most accurate - that is, the three un-labelled categories identified in Survey 3. We did so, by publishing an ontology (“orca-x”) that inherits from ORCA.

Because we are unable to identify appropriate labels for our three categories, we define a set of classes called CategoryA, CategoryB, and CategoryC that inherit from the ORCA “ConfidenceLevel” class, and indicate with a label that these classes share similarities with Doxastic, Dubitative, and Hypothetical knowledge, respectively. We then published this ontology in GitHub, and utilized the W3ID redirection system to provide these new ontological categories with globally unique and persistent URLs, as per the FAIR requirements. An example of the use of these ontology classes can be seen in the example NanoPublication in Figure 23, which is the output from our machine-learning annotation software (described below).

```

@prefix this: <http://linkeddata.systems/nanopubs_mario/CertID_1> .
@prefix sub: <http://linkeddata.systems/nanopubs_mario/CertID_1#> .
@prefix void: <http://rdfs.org/ns/void#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix dcelem: <http://purl.org/dc/elements/1.1/> .
@prefix np: <http://www.nanopub.org/nschema#> .
@prefix pav: <http://purl.org/pav/2.3/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix schema: <https://schema.org/> .
@prefix certainty: <http://w3id.org/orca-x#> .
@prefix thispub: <https://dx.doi.org/10.1083/jcb.200404108> .

sub:Head {
  this: np:hasAssertion sub:assertion ;
  np:hasProvenance sub:provenance ;
  np:hasPublicationInfo sub:pubinfo ;
  a np:Nanopublication .
}
sub:assertion {
  certainty: asserts-1 rdf:singletonPropertyOf certainty:asserts .
  thispub: certainty:asserts-1 'These results demonstrate that the Shank3 protein forms a specific complex with
    the Ret9 isoform through a novel Ret9 PDZ-binding motif.' .
  thispub: certainty:hasConfidenceLevel certainty:CategoryA .
}
sub:provenance {
  sub:assertion dcterms:author "Certainty Classifier" ;
  dcterms:title "Automated Certainty Classification of Statement from https://dx.doi.org/10.1083/jcb.200404108" ;
  dcterms:license <https://creativecommons.org/publicdomain/zero/1.0/> ;
  schema:identifier this: ;
  dcat:distribution sub:_1 .

  sub:_1 dcelem:format "application/pdf" ;
  a void:Dataset , dcat:Distribution ;
  dcat:downloadURL <https://dx.doi.org/10.1083/jcb.200404108> .
}
sub:pubinfo {
  this: dcterms:created '2019-03-15'^^xsd:date ;
  foaf:primaryTopic sub:assertion ;
  dcterms:rights <https://creativecommons.org/publicdomain/zero/1.0> ;
  dcterms:rightsHolder <https://orcid.org/0000-0002-9416-6743> ;
  pav:authoredBy "Mario Prieto" , <https://orcid.org/0000-0002-9416-6743> ;
  pav:versionNumber "1" ;
  prov:wasGeneratedBy "Mario Prieto's Certainty Classifier" .
}

```

Figure 23: An exemplar prototype NanoPublication including certainty annotations. The Figure shows how certainty classifications could be used as additional, and important metadata when added to text-mining pipelines.

In this NanoPublication (representing statement #29 in this study) the concept being asserted (the creation of a Shank3/Ret9 complex) is captured using ontologically-based concepts in the “assertion” block of the NanoPublication. We wish to attach the certainty annotation to the assertion, and this is done using `singletonProperties`[130], a design choice that allows us to efficiently and in a self-contained way, annotate individual triples when Named Graphs are already being utilized for another purpose (as in the case of NanoPubs, where the Assertion triples are already in a Named Graph as part of the structural requirements of a NanoPublication). In this case, the assertion triples are annotated to belong to `CategoryA` from the `orca-x` ontology (red text). Such annotations could be used, in particular by machines, to filter a large number of assertions (for example, in a database query) based on their degree of certainty. The “provenance” block then carries a variety of information about the original text – that is, who originally generated the publication that is being annotated by our software. The final block, `PubInfo`, contains authorship, license, and citation information for the NanoPublication itself, expressing the terms of usage of this metadata, and who to cite. In this case, the `PubInfo` block contains the citation information for our Machine-learning/NanoPub generating software, as the source of this NanoPublication. This entire structure can be interpreted by automated agents, and fully complies with the FAIR Data Principles, as demonstrated below.

We have created a repository of nanopublications representing the ~3000 statements used in this thesis (<https://github.com/Guindillator/thesis/tree/master/Nanopublications>).

In parallel with the creation of these NanoPublications, a software library was created that was capable of automatically interacting with the recently published FAIR Evaluator framework [131] (with whom we collaborated during this thesis). The FAIR Evaluator consumes a metadata record, and executes up to 22 distinct tests on the content of that record, determining which aspects of the FAIR Principles can be detected by a machine within that record. Using our Evaluator interaction library, an objective FAIRness evaluation was executed on these NanoPublications. The result shows that they achieve a score of between 19 or 20 “passes” out of 22 tests (the result is stochastic – 19 or 20 – based on which triples from the nanopublication are selected by the Evaluator to be tested, which occurs randomly). Of the ~400 FAIR Evaluations executed so far by the public – spanning several dozen high-profile data repositories including Zenodo, Figshare, Dryad, and major biomedical databases such as EBI - the NanoPublications generated by our system have the highest “FAIRness” score obtained to-date. As such, we believe that this facet of the thesis has also been achieved at an objectively quantifiable high level of quality relative to other international efforts.

5.10. Applying Automated Certainty Annotations

A straightforward python package called *biocertainty* has been developed. It incorporates the machine learning model on certainty described earlier (see Results section 4.8). The function *biocertainty.Certainty()* provides the level of confidence assigned to one statement. It has two additional functions that transform the statement into a formal publication in the form of either a Nanopublication (described in detail above) or a Micropublication.

Biocertainty has been added to the repository of software for the Python programming language, Python Package Index (PyPI). It is currently available as open source software.

Figure 24 shows the result of using the package to categorize statements in a series of citation chains. Early statements (those lower in the Figure panels) were found to have a low level of certainty; however, along successive steps in the citation chain, the certainty levels increase. Figure 24, discussed earlier, is the output from the *biocertainty.Nanopublication()* function, assigning a level of certainty to the input and creating a machine-readable nanopublication containing that novel metadata.

A

<i>"We have previously demonstrated that accumulation of AβPP epitopes precedes other abnormalities in IBM muscle fibers" [37]</i>
<i>"βAPP accumulation is considered to play a major role in the pathogenesis of IBM and AD and is thought to precede other changes in both diseases"[132]</i>
<i>"Those muscle fibers, widely prevalent in our one case of hereditary IBM, may represent early changes of IBM and therefore be analogous to the finding in AD brains where PAP accumulations in the "diffuse" Congo-red-negative plaques seem to represent early changes" [133]</i>

B

<i>"We have previously demonstrated that accumulation of AβPP epitopes precedes other abnormalities in IBM muscle fibers" [37]</i>
<i>"Increased βAPP-mRNA and increased accumulation of βAPP epitopes appear to precede other abnormalities in IBM muscle fiber" [134]</i>
<i>"One possibility is that one protein is accumulated first, due to excessive synthesis, e.g., excessive transcription of mRNA in the IBMs is known for beta APP"[135]</i>

C

<i>Recently it was reported that s-IBM vacuolated muscle fibers, and those in some other vacuolar myopathies, contain a marker of autophagosomes, but only in s-IBM is it colocalized with AβPP[18]. [136]</i>
<i>Overexpression of amyloid precursor protein (APP) and subsequent accumulation of cleaved fragments including β-amyloid in vacuolated muscle fibers is considered a central mechanism in the pathogenesis of s-IBM.[2] [137]</i>
<i>it is now established that Aβ/AβPP is also abnormally accumulated in muscle fibers of s-IBM patients, where they are considered to play an important pathogenetic role[4,5,6,7] [138]</i>
<i>A possibility that excessive accumulation of AβPP/Aβ induces inflammation has been proposed by us and by others.[1-3,7,10] [139]</i>
<i>Deposition of the Aβ fragment of the amyloid precursor protein is a feature of affected muscle in IBM (see below) and it has been shown that muscle cells can secrete Aβ. [10] Interaction of Aβ with muscle cells in turn can stimulate IL-6 production by these cells [19]... [40]</i>
<i>However, in some abnormal muscle fibers in IBM, the accumulation of βAPP appears to extend outside the muscle fiber boundary. This may have been attributable to a fragility of the fiber's surface membrane, which could have been transiently broken.[25] [140]</i>

Figure 24: Automated classification of scholarly assertions related to the accumulation of beta-APP protein in muscle fibers. Statements are color-coded as green (Category A - highest certainty), orange (Category B - medium certainty) and red (Category C - lowest certainty). (A and B) Two citation chains showing that the degree of certainty expressed in the most recent statement is higher than that in the cited text. (C) A selection of statements identified by Greenberg, 2009, as being potentially indicative of 'citation distortion'. In that panel, there is a general trend to higher certainty over time, with the exception of an early high-certainty statement by Mastaglia in 2003 (second row from the bottom).

6. DISCUSSION

6.1. Evidence to support three levels of certainty in scholarly statements

In S1, we began with a four-category classification system, since this is the highest number presumed in earlier studies (Zerva et al. used a five point numerical scale, but we do not believe they were proposing this as a categorization system). In the absence of any agreed-upon set of labels between these prior studies, and to enable untrained annotators to categorize scholarly statements, we labelled these categories Low, Medium Low, Medium High and High. This Survey revealed a statistically significant categorization agreement for 37 of the 45 statements (82.2% of total), with seven statements being doubly-classified and one statement showing poor inter-annotator agreement, for a total of eight ‘ambiguous’ classifications. The G index (Holley & Guilford, 1964) with only four categories is small, and the statistical probability of chance-agreement in the case of ambiguity is therefore high, which may account for the high proportion of doubly-classified statements. Interestingly, the category Low was almost never selected by the readers. We will discuss that observation in isolation later in this discussion; nevertheless, for the remainder of this discussion we will assume that this category does not exist in our corpus of statements, and will justify this in these later, more detailed arguments.

With respect to the categories themselves, the category of High had robust support using the G index statistic, indicating that it represents a valid category of certainty based on agreement between the annotators on the use of that labelled category. Support for the other two, medium-level, categories was less robust. This could be interpreted in two ways - one possibility is that these two categories are not distinct from one another, and that readers are selecting one or the other “arbitrarily” with statistical significance, because there were only two choices. This would suggest that there are only two certainty categories used in scholarly writing. The other option is that the labels assigned to these two non-high categories do not accurately reflect the perception of the reader, and thus that the categorizations themselves are flawed, leading to annotator confusion.

In Survey 2, with only two categories (Relatively High and Relatively Low), statistical support for these two categories was evident, but deeper examination of the results suggests that these categories may still not accurately reflect the reader’s perception. For example, seven of the 45 statements (15.5%) showed no inter-annotator agreement. Of the remainder, Table 6 and Figure 10 show a clear pattern of association between the strength of certainty perceived by the reader, and the degree to which the readers agreed with one another. Effectively, there was greater agreement on the categorization of high-certainty statements, than low-certainty statements. This mirrors the observations from Survey 1, where the category High generated the highest levels of agreement among annotators. Since

this binary categorization system lacks an intermediate category, the G index in this Survey is 0.5, meaning that agreement by chance is high. It appears that statements that would have been categorized into one of the middle classes from Survey 1 became distributed between the two Survey 2 categories, rather than being categorized uniformly into the lower category. This would indicate that the two-category explanation for Survey 1 is not well-supported, and possibly, that the labelling of the categories themselves in both Survey 1 and Survey 2 confounds the analysis and does not reflect the perception of the reader. In other words, the category High/Relatively High seems to match a perception that exists in the minds of the readers, but the categories labelled Medium High (S1), Medium Low (S1) and Relatively Low (S2) might not correspond to the perception of the readers as they interpret and attempt to classify lower certainty statements, which is why they are less consistent in the selection of these categories.

To reveal patterns that may clarify what defines these lower categories, we utilized a variety of clustering approaches (Figures 12 and 13). That there are three, rather than two, categories is supported by the hierarchical clustering of all three Surveys, shown in Fig.12 (see clusters along the top edge) which reveals three primary clusters in the data, where high is strongly differentiated from non-high categories. The output from NbClust's "majority rule" approach to selecting the optimal number of clusters was executed on individual Surveys. The results for S1 and S2 are shown in Fig. 13A and Fig. 13B. The majority rule indicates that there were three discernable clusters in S1. Survey 2 was assessed by the 30 NbClust indices [102] (Fig.13B). Surprisingly, we found that, while 11 indices recommended only two clusters, which was expected from a survey with only two options, six indices suggested that there were three clusters. Since in this analysis a cluster represents a pattern of "categorization-behavior" among all evaluators, we take these results as further indication that there are three discernable annotator responses when faced with a certainty categorization task.

To further explore the meaning of these clusters, we executed a feature reduction analysis using Principal Components. The PCA of Survey 1 revealed three primary components accounting for ~97% of the variability. The main component, accounting for more than half (~56%) of the variation, is characterized by a strong positive influence from the category labelled High, and a strong negative influence from the categories labelled Medium Low and Low. This lends support to our earlier interpretation that there is little ambiguity among annotators about what statements are classified as highly certain, and moreover, when faced with a high-certainty statement annotators will almost never select one of the low categories. The second and third components (accounting for ~32% and ~10%

respectively) are more difficult to interpret. Component 2 is characterized by a strong positive influence from the category Medium High, and a strong negative influence from the category Low; Component 3's "signature" is distinguished through a positive influence from the category Low, though as stated earlier, this category was rarely selected, and showed no significant agreement among annotators, making this difficult to interpret. The lack of clarity regarding the interpretation of these second and third components may reflect ambiguity arising from the labelling of the non-high certainty categories in the questionnaire; effectively, the words used for the labels may be confusing the readers, and/or not aligning with their impressions of the statements.

In an attempt to gain additional evidence for a three-category classification system, we undertook a third Survey (S3) in which the reader was offered three categories, ordered from higher to lower, but with numerical labels (1, 2, or 3). The rationale for this was twofold. First, we could not think of three suitable labels that would not inherently bias the results (for example, 'high', 'medium', and 'low' would not be suitable because we have already determined that the category 'low' is almost never selected). In addition, we wished to know if category labels were a potential source of bias, and therefore more semantically neutral labels might lead to a stronger correspondence between the annotators. Indeed, Survey 3 generated the most consistent agreement of the three questionnaires, where only four of the 45 statements did not meet the cutoff level for annotator agreement, and none were doubly-classified. It is not possible to disambiguate if this enhanced agreement is due to the annotators being presented with a "correct" number of categories, or if it supports the suggestion that the presentation of meaningful (but non-representative) category labels caused annotators to behave inconsistently in S1 and S2, or perhaps a combination of both. As with S1, NbClust's "majority rule" proposes three clusters for S3 (Fig.13C).

In Fig.12 we present the correlation matrix to show how the categories relate to one another between the three Surveys, using a Spearman Correlation. High (S1) is clearly correlated with Relatively High (S2) and Category 1 (S3). Medium Low (S1), Relatively Low (S2) and Category 3 (S3), are also highly correlated. Low (S1) only has moderate correlation with Relatively Low (S2) and Category 3 (S3). The intermediate values Medium High (S1) and Category 2 (S3), are found on the negative side of Principal Component 1 (Fig.17A & Fig.17D), which supports the interpretation that a High certainty category is strongly supported, and strongly distinct from other categories. The non-high categories appear as distinct blocks within the correlation matrix, but with more ambiguity or inconsistency, though the Jaccard similarity index was sufficient to support the existence of two distinct

lower-certainty categories. Additionally, the clusters identified by the Spearman analysis (three clusters) are supported by the results of the HCA analysis (three branches).

One general source of inconsistency we noted in the data could be described as a “tendency towards the middle”. When a category is removed, statements from that category tend to distribute to adjacent categories. We presume this reflects some form of “central tendency bias”, a behavioral phenomenon earmarked as a preference for selecting a middle option.[141], [142], [143]. Nevertheless, this did not appear to be sufficiently strong in this investigation to mask the detection of distinct clusters of categorization behavior.

In summary, these results suggest that there are three categories of certainty in the minds of the readers of scholarly assertions. One category is clearly distinguished as representing high-certainty statements. The other two categories, representing non-high certainty statements, are distinct from one another in the minds of the annotators, however, seem to not be reflected well by the labels “moderately/relatively + high/low”. Nevertheless, they do appear to represent a higher-to-lower spectrum, since the replacement of textual labels with a numerical range resulted in stronger annotator agreement about these two lower categories.

6.2. The absence of a Low certainty category

Several studies that preceded this one [21], [29], [36] suggested four categories of certainty, with one of those being a category that would represent the lowest certainty. In this study, we identify only three. The category that seems to be absent from our data is this lowest category - generally described as “no knowledge” in these three precedent studies. We examined our corpus and, given the grammatical cues suggested by De Waard (De Waard & Pander Maat, 2012) we identified two statements in our corpus that, by those metrics, should have scored in the Low category. Those are Statement 3, *“However, this was not sufficient for full blown transformation of primary human cells, which also required the collaborative inhibition of pRb, together with the expression of hTERT, RASV12.”*, and Statement 4, *“Hence, the extent to which miRNAs were capable of specifically regulating metastasis has remained unresolved.”*

Looking at the results in Table 5, 6 and 7, these two statements were annotated with considerable agreement as *high-certainty* statements - the opposite of what would have been predicted. One explanation for this is that the statements are making a negative claim, with high certainty, and thus are being categorized as high-certainty assertions by our

annotators. If that is the case, then the category of “no knowledge” may not be a category that lies anywhere on the spectrum of certainty, and may reflect a distinct feature of scholarly communication discourse, or (more likely) a combination of the meta-knowledge facets of “certainty” and “polarity” (in essence, “certainly not”)

6.3. Comparison between Questions and Surveys

The results of the certainty classification question, when comparing the Wilcoxon and Kruskal-Wallis test for S1-S2, S1-S3, S2-S3 and S1-S2-S3, show that despite their having been carried out in different countries over several years, there is no significant difference between them for the 45 statements (Table 11. All p-values > 0.05). Therefore, we might conclude that the results of the three Surveys are, effectively, extracted from the same population, regardless of the native language or the country where the questionnaires were conducted. Certainty appears to be perceived consistently and with the same number of discernable categories, everywhere (at least, in the biomedical community).

The results obtained when comparing Question 2 between S2 and S3, which both use the same methodology (5 possible positions/ranking for each statement) and the Independence test that was carried out by comparing the raw results of Question 2 between S2 and S3 also, further confirms that they are extracted from the same population. Moreover, the Independence test reassures us that applying the score obtained from *Relative Importance Index* to each statement does not distort the sample. Therefore, when we compare Question 1 with Question 2 for the three questionnaires we found that except for S1 (due to bias error) S2 and S3 show a p-value higher than 0.05, and we can conclude that the answers obtained in the questionnaires for both question 1 and question 2 have not been answered randomly.

This assures us that the overall analysis conducted in this thesis – where we examine three distinct surveys from three linguistically and geographically separate communities – is a valid analysis. All three surveys can legitimately be compared to one another, and the results treated as a uniform corpus.

6.4. Basis and Certainty Correlation

The results obtained from Question 1.2 regarding the bases of the statements (Table 14, 15 and 16) show that there was poor agreement in ~6,6% of the statements and ~13,3% had been doubly classified. This low agreement suggests one of two things – that our defined classification system for “basis” was incorrect, or that there is genuinely disagreement between annotators regarding the basis of certain scholarly statements. Speculatively, this might indicate a difference in the level of trust assigned to published scholarly statements, where “trusting” individuals assign a more concrete basis to an ambiguously-sourced statement than “non-trusting” individuals. If this were true, it is possible that inclusion of demographic information (e.g. the seniority of the researcher) might reveal the source of these differences in perception.

In Figure 19, 20 and 21 (S1, S2 and S3, respectively) we provide a correlation matrix to show how Basis relates to certainty categories in the three surveys using Spearman correlation. Direct Evidence clearly shows a relationship with “high” certainty categories (S1-High, S2-Relatively High and S3-Category 1). Speculation show also very important correlation with “Low” certainty categories (S1-Medium Low, S1-Low, S2-Relatively Low and S3-Category 3). Indirect Evidence/Reasoning has moderate correlation with Medium High (S1) and Category 2 although the result was not significant enough to consider valid.

6.5. Machine learning

When we compare our manual classification with the participants' responses, Figure 22A, B and C reveal that the largest group of participants (89, 1% for S1, 86, 5% for S2 and 86.5%) are in the upper right area of the graph. This means that most raters have at least 50% of their responses in-common with our manual annotations. They also responded to medium statements with medium or high certainty (average score for the statements answered >0, Y axis). The regression line suggests that the higher the level of certainty, the greater the similarity between the participants' responses and ours, demonstrating decreasing ambiguity/disagreement with increasing certainty.

Other than the results of S2, when comparing our manual classification with the median of participants for each statement for S1 and S3, we obtain an Overall Accuracy score greater than 80%. This supports the proposition that our manual classification resembles the population trend, and is therefore reasonable to use as the input to a machine-learning model.

The result of the Cross-Validation (Table 20) shows that there is no overfitting in any of the folds and that our learning never stops gain information. Thus, it confirms the performance of the model, and its validity when predicting new data.

6.6. Application of this categorization system

As indicated in the Introduction, a primary motivation for this study is its application to the automated capture of metadata related to the certainty being expressed, particularly in text-mined scholarly assertions, or to identify or monitor ‘hedging erosion’ within identified citation chains. To demonstrate how the outcomes of this study can be applied, we have used the data described here to generate, by machine-learning, an automated certainty classifier capable of assigning new scholarly statements into one of the three certainty categories [127]. Two exemplar outputs from this classification system are shown in Figs. 23 and 24. Figure 24 shows three sets of statements, color-coded by the category of certainty detected by our classifier - green (Category A, associated with High certainty), orange (Category B, non-high/moderate), and red, (Category C non-high/low). Two citation chains relate to the accumulation of beta-APP in muscle fibers of Alzheimer’s Disease patients (Fig.24A and 24B), while Fig.24C shows a longer citation chain identified by Greenberg as being problematic with respect to ‘citation-distortion’[19]. The panels reveal that the degree of certainty can change through citation, becoming higher (Fig.24A and 24B). Fig.24C reveals a similar trend toward increasing certainty, with the exception of one author who used a clearly high-certainty assertion four years before others in the community expressed the same idea with certainty.

6.7. Tools for researchers, authors, reviewers, and data miners

As discussed in the introduction, researchers may lack the knowledge required to assess the legitimacy of claims that are not directly in their domain, or may be unaware of the history of a claim if they have not followed a citation chain to its roots. Similarly, when acting as peer reviewers, there is little tooling to assist them in evaluating the validity of assertions in the submitted manuscript or funding proposal. In parallel with research into automated identification of reference-spans [10], the availability of a certainty classifier would make it possible to automate the creation of annotated citation chains such as shown in Fig.24. Reviewers could then use these to determine if a claim was being made with unusually high

(or low) certainty - like the Magstalia statement from 2003, shown in Fig.24C - and thus enhance the confidence of their reviews.

Though considerable effort was spent in this thesis attempting to improve the state-of-the-art in the identification of reference spans, we concluded that we were not achieving any more success than our peer laboratories, and as a result, we focused on questions for which we had a unique opportunity to make a novel contribution. However, our contribution, while already significant, will become more significant as these peer researchers become increasingly able to automate the process of citation-chain detection.

Tools derived from our work could become an important part of the scholarly planning process. During the preparation of a paper or proposal, researchers could be made aware of dubious assertions, and avoid relying on these as the basis for their hypothesis. In the context of automated data mining, assuming that incremental steps towards certainty should be associated with the existence of supporting data, the automated detection of “certainty inflection points” could be used by data mining algorithms to identify the specific dataset containing data supporting (or refuting) a given claim. Together with the use of certainty classification in the context of text-mining discussed above, the use of such a classification system may become an important part of the scholarly publishing lifecycle.

Based on this, a set of Python libraries were generated, and released in the Open Source, that facilitate the application of our machine learning classifier to new statements. Moreover, these libraries also enable the output from this machine learning model to be published in one of two FAIR, machine-accessible formats – NanoPublications, and Micropublications. From this library, we have generated >3000 annotated scholarly statements, also published in the Open Source, that may be used by other researchers for, e.g., linguistic evaluations of various certainty classifications.

Additional tools derived, in whole or in part, from this thesis work were used to validate our machine-readable outputs with respect to their “FAIRness”. These include our participation in the development of an objective evaluation system for FAIR – called the Evaluator – which executes a series of publicly-generated tests for certain “behaviors” within a (meta)data record that are compliant with the FAIR Principles. In addition to our core efforts on the Evaluator, we also generated Python libraries that facilitate automated, high-throughput interaction with the Evaluator. Since it is not scholarly to simply declare that a research output is “FAIR” without evidence, these libraries were used to test the quality of the NanoPublications generated by our machine-learning outputs. The results of this quality-assessment demonstrate that, to date, the NanoPublications generated in this thesis are the

“most FAIR” digital objects ever evaluated, since the emergence of automated FAIR evaluations.

Finally, we should consider what has been learned from these studies at a much higher level, as it has implications for the future of scholarly publishing. In particular, it is noteworthy that these studies imply that narrative scholarly communication is both wasteful and unnecessarily opaque. Ideas are encoded into a myriad of hedging structures that are difficult to interpret by the reader, difficult to interpret by a machine, and yet appear to encode only one of three possible ideas (at least, with respect to certainty, which is the primary purpose of hedging language). They are effectively “grammatical sugar” – of little benefit to healthy scholarly communication, while at the same time making scholarly output seemingly more pleasant to consume. There is clearly a need, and a move toward, *ab initio* publication for mechanized consumption – that is, a move away from narrative publications to machine-first publications. It is reassuring, therefore, that there are so few categories of certainty that need to be distinguished/represented in these machine-readable resources, which will simplify the creation of, and fidelity of, these novel scholarly publication frameworks.

6.8. Future investigations to elucidate perceptions of certainty

A variety of future studies could provide additional insight into how researchers communicate and perceive certainty. First, it is noteworthy that all of these analyses were undertaken without any attempt to “cleanse” the data. That is, while overall there was no indication of annotators behaving inconsistently, we did not test if *individual* annotators were behaving inconsistently. This was because, at the outset, we felt that we could not definitively determine what “inconsistent” meant (beyond providing the same answer to every question, which we of course filtered-out as being invalid responses). As a result, the survey data was treated as raw data throughout this investigation. We have no doubt that, after determining what certainty categories exist, we could return to the raw data to identify annotators who consistently respond outside of the consensus. Removal of these oddly-behaving annotators would almost certainly improve the overall statistics of the study, and would likely also improve the statistics of the machine learning algorithm’s accuracy.

The results presented here seem to suggest that words like “medium” and “low” do not align well with the perception held by researchers as they read statements that fall into non-high certainty categories. Future studies could extract additional information in the

questionnaire, such as questions related to the basis upon which an assertion was made (e.g. speculation, direct or indirect observation, etc.), as it may be that the distinction between the two lower certainty categories is being made based on other kinds of implicit information, rather than being specifically “medium” or “low” expressions of certainty. Using the data, we have made openly available, professional Linguists could undertake a deeper analysis of the grammatical structure of the members of our three categories, and would likely be able to create even more accurate automated classification tools (and we would encourage a data cleansing step, noted above, before this analysis is done). Finally, it would also be interesting to capture demographic information, to determine if perception of certainty changes as a researcher becomes more experienced, if it differs between different linguistic groups, or if it is associated with other demographic variables.

7. CONCLUSIONS

- *Certainty is not a continuum, but rather, can be thought of as a set of distinguishable categories that are uniformly perceived by readers of scholarly literature in the bio-sciences.*
- *There are three categories of certainty, though these can be expressed in a myriad of different grammatical structures through the use of hedging.*
- *Artificial intelligence models can detect these certainty levels, and thus can assign certainty annotations to novel scholarly statements.*
- *Application of certainty annotation to citation chains reveal “inflexion points”, where the certainty level changes.*
- *It is possible to formalize the capture and exchange of these certainty assignments through semantic technologies such as Resource Description Framework and Web Ontology Language, and formal knowledge-publication frameworks such as Nanopublications.*
- *Such NanoPublications represent among the most FAIR-compliant digital objects created to-date.*

8. REFERENCES

- [1] Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016;533:452–4.
- [2] Ioannidis JPA, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, et al. Repeatability of published microarray gene expression analyses. *Nat Genet* 2009;41:149–55.
- [3] De Waard A, Maat HP. Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features. *Proceedings of the Workshop on Detecting* 2012.
- [4] Sethunya R J, Hlomani H. Natural Language Processing: A Review. *International Journal of Research in Engineering and Applied Sciences* n.d.
- [5] Smaoui C. Linguistic Features to Compile a Successful Scientific Discourse: Have Tunisian Novice Researchers Ever Seen Such Features during their Educational Career. *IJLL* 2015;3:8.
- [6] Lippincott T, Séaghdha DÓ, Korhonen A. Exploring subdomain variation in biomedical language. *BMC Bioinformatics* 2011;12:212.
- [7] Fanelli D. Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data. *PLoS One* 2010;5:e10271.
- [8] Sarewitz D. The pressure to publish pushes down quality. *Nature* 2016;533:147–147.
- [9] Blanco P, Fidalgo E, Alegre E, Al-Nabki MW. Detecting textual information in images from onion domains using text spotting. *Actas de Las XXXIX Jornadas de Automática, Badajoz, 5-7 de Septiembre de 2018* 2018.
- [10] Saggion H, Ronzano F, Others. Trainable citation-enhanced summarization of scientific articles. *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, 2016, p. 175–86.
- [11] Prieto M, Deus H, De Waard A, Schultes E, García-Jiménez B, Wilkinson MD. Data-driven classification of the certainty of scholarly assertions. *PeerJ Preprints*; 2019. doi:10.7287/peerj.preprints.27829v1.
- [12] Salager-Meyer F. Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes* 1994;13:149–70.
- [13] Hyland K. Writing Without Conviction? Hedging in Science Research Articles. *Appl Linguist* 1996;17:433–54.
- [14] Lakoff G. Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. In: Hockney D, Harper W, Freed B, editors. *Contemporary Research in Philosophical Logic and Linguistic Semantics: Proceedings of a Conference Held at the University of Western Ontario, London, Canada, Dordrecht: Springer Netherlands*; 1975, p. 221–71.
- [15] Skelton J. Comments in Academic Articles. 1988.
- [16] Myers G. "In this paper we report ...": Speech acts and scientific facts. *Journal of Pragmatics* 1992;17:295–313. doi:10.1016/0378-2166(92)90013-2.
- [17] Hashemi MR, Shirzadi D. The use of hedging in discussion sections of applied linguistics research articles with varied research methods. *Journal of Teaching Language Skills* 2016;35:31–56.
- [18] Masnick AM, Zimmerman C. Evaluating Scientific Research in the Context of Prior Belief: Hindsight Bias or Confirmation Bias? *Journal of Psychology of Science and Technology* 2009;2:29–36. doi:10.1891/1939-7054.2.1.29.
- [19] Greenberg SA. How citation distortions create unfounded authority: analysis of a citation network. *BMJ* 2009;339:b2680.
- [20] Agami R, Bernards R. Distinct initiation and maintenance mechanisms cooperate to induce G1 cell cycle arrest in response to DNA damage. *Cell* 2000;102:55–66.
- [21] Wilbur WJ, Rzhetsky A, Shatkay H. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics* 2006;7:356.
- [22] Lorés R. On RA abstracts: from rhetorical structure to thematic organisation. *English for Specific Purposes* 2004;23:280–302. doi:10.1016/j.esp.2003.06.001.
- [23] Rubinstein A, Harner H, Krawczyk E, Simonson D, Katz G, Portner P. Toward fine-grained annotation of modality in text. *Proceedings of the IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, 2013, p. 38–46.
- [24] Light M, Qiu XY, Srinivasan P. The language of bioscience: Facts, speculations, and statements in between. *HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases*, 2004.
- [25] Malhotra A, Younesi E, Gurulingappa H, Hofmann-Apitius M. "HypothesisFinder": a strategy for the detection of speculative statements in scientific text. *PLoS Comput Biol* 2013;9:e1003117.
- [26] Zerva C, Batista-Navarro R, Day P, Ananiadou S. Using uncertainty to link and rank evidence from biomedical literature for model curation. *Bioinformatics* 2017;33:3784–92.
- [27] Voorhoeve PM, le Sage C, Schrier M, Gillis AJM, Stoop H, Nagel R, et al. A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. *Cell* 2006;124:1169–81.
- [28] Zimmermann N, Colyer JL, Koch LE, Rothenberg ME. Analysis of the CCR3 promoter reveals a regulatory region in exon 1 that binds GATA-1. *BMC Immunol* 2005;6:7.
- [29] Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for

- clinical radiology. *J Am Med Inform Assoc* 1994;1:161–74.
- [30] Vincze V, Szarvas G, Farkas R, Móra G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics* 2008;9 Suppl 11:S9.
 - [31] Thompson P, Nawaz R, McNaught J, Ananiadou S. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics* 2011;12:393.
 - [32] Coates J. EPISTEMIC MODALITY AND SPOKEN DISCOURSE. *Transactions of the Philological Society* 1987;85:110–31. doi:10.1111/j.1467-968x.1987.tb00714.x.
 - [33] Ciglič BP. Glagolski kazalniki epistemične modalnosti in njihova vloga pri razvoju sporazumevalnih jezikovnih zmožnosti. *Linguistica* 2014;54:381–95.
 - [34] Holmes J. Expressing Doubt and Certainty in English. *RELJ* 1982;13:9–28.
 - [35] Thompson P, Venturi G, McNaught J, Montemagni S, Ananiadou S. Categorising modality in biomedical texts. *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, 2008, p. 27–34.
 - [36] De Waard A, Schneider J. Formalising uncertainty: An ontology of reasoning, certainty and attribution (ORCA). *Proceedings of the Joint Workshop on Semantic Technologies Applied to Biomedical Informatics and Individualized Medicine (SATBI+ SWIM 2012)*, Boston MA, USA: 2012, p. 8–15.
 - [37] Askanas V, Engel WK, Alvarez RB, McFerrin J, Broccolini A. Novel Immunolocalization of α -Synuclein in Human Muscle of Inclusion-Body Myositis, Regenerating and Necrotic Muscle Fibers, and at Neuromuscular Junctions. *J Neuropathol Exp Neurol* 2000;59:592–8.
 - [38] Paciello O, Wójcik S, Engel WK, McFerrin J, Askanas V. Parkin and its association with alpha-synuclein and AbetaPP in inclusion-body myositis and AbetaPP-overexpressing cultured human muscle fibers. *Acta Myol* 2006;25:13–22.
 - [39] Voorhoeve PM, le Sage C, Schrier M, Gillis AJM, Stoop H, Nagel R, et al. A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. *Adv Exp Med Biol* 2007;604:17–46.
 - [40] Mastaglia FL, Garlepp MJ, Phillips BA, Zilko PJ. Inflammatory myopathies: clinical, diagnostic and therapeutic aspects. *Muscle Nerve* 2003;27:407–25.
 - [41] GENIA Event Extraction (GENIA) - BioNLP Shared Task n.d. <http://2011.bionlp-st.org/home/genia-event-extraction-genia> (accessed May 13, 2019).
 - [42] Coates J. The expression of root and epistemic possibility in English. *Modality in Grammar and Discourse* 1995;55:66.
 - [43] Weiss V, Medina-Rivera A, Huerta AM, Santos-Zavaleta A, Salgado H, Morett E, et al. Evidence classification of high-throughput protocols and confidence integration in RegulonDB. *Database* 2013;2013:bas059.
 - [44] Teufel S, Carletta J, Moens M. An Annotation Scheme for Discourse-level Argumentation in Research Articles. *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics; 1999, p. 110–7.
 - [45] Palmer FR. *Mood and Modality*. Cambridge University Press; 2001.
 - [46] Kuhn TS. *The Structure of Scientific Revolutions: 50th Anniversary Edition*. University of Chicago Press; 2012.
 - [47] De Waard A, Breure L, Kircz JG, Van Oostendorp H. Modeling rhetoric in scientific publications. *International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006*, 2006.
 - [48] Fuller S. Book Review : *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*, by Charles Bazerman. Madison: University of Wisconsin Press, 1988. *Sci Technol Human Values* 1991;16:122–5.
 - [49] Latour B, Woolgar S. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press; 2013.
 - [50] Kloosterman WP, Plasterk RHA. The diverse functions of microRNAs in animal development and disease. *Dev Cell* 2006;11:441–50.
 - [51] Yabuta N, Okada N, Ito A, Hosomi T, Nishihara S, Sasayama Y, et al. Lats2 is an essential mitotic regulator required for the coordination of cell division. *J Biol Chem* 2007;282:19259–71.
 - [52] Okada N, Yabuta N, Suzuki H, Aylon Y, Oren M, Nojima H. A novel Chk1/2-Lats2-14-3-3 signaling pathway regulates P-body formation in response to UV damage. *J Cell Sci* 2011;124:57–67.
 - [53] de Waard A. These Results Suggest That...', Knowledge Attribution in Scientific Discourse. Slideshare n.d. <https://es.slideshare.net/anitawaard/these-results-suggest-that-knowledge-attribution-in-scientific-discourse/2>.
 - [54] Nigel Gilbert G. Referencing as Persuasion. *Soc Stud Sci* 1977;7:113–22.
 - [55] Kovanis M, Porcher R, Ravaud P, Trinquart L. The Global Burden of Journal Peer Review in the Biomedical Literature: Strong Imbalance in the Collective Enterprise. *PLoS One* 2016;11:e0166387.
 - [56] Brun RE, Senso JA. Artículo Minería textual. *El Profesional de La Información* 2004;13:11.
 - [57] Cohen AM, Hersh WR. A survey of current work in biomedical text mining. *Brief Bioinform* 2005;6:57–71.
 - [58] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 2016;3:160018.
 - [59] Groth P, Gibson A, Velterop J. The anatomy of a nanopublication. *Inf Serv Use* 2010;30:51–6.
 - [60] Clark T, Ciccarese PN, Goble CA. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. *J Biomed Semantics* 2014;5:28.
 - [61] About | nanopub.org n.d. http://nanopub.org/wordpress/?page_id=65 (accessed July 1, 2019).

- [62] G7 Science and Technology Ministers. Tsukuba Communiqué G7 Science and Technology Ministers' Meeting. 2016.
- [63] Big Data to Knowledge | NIH Common Fund n.d. <https://datascience.nih.gov/commons> (accessed June 25, 2019).
- [64] Directorate-General For Research. H2020 Programme Guidelines on FAIR Data Management in Horizon 2020. EUROPEAN COMMISSION ; 2016.
- [65] G20 Nation Leaders. G20 Leaders' Communique Hangzhou Summit. 6 September, 2016.
- [66] SEC.gov | The Role of Machine Readability in an AI World 2018. <https://www.sec.gov/news/speech/speech-bauguess-050318> (accessed July 1, 2019).
- [67] Mjolsness E, DeCoste D. Machine learning for science: state of the art and future prospects. *Science* 2001;293:2051–5.
- [68] Veronika V. Uncertainty Detection in Natural Language Texts. phd. szte, 2015.
- [69] Tema 3: Análisis de Componentes Principales n.d. <http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema3am.pdf> (accessed July 3, 2019).
- [70] Lindsay S. A tutorial on Principal Components Analysis 02, 2002.
- [71] `scipy.cluster.hierarchy.linkage` — SciPy v1.3.0 Reference Guide n.d. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html> (accessed July 3, 2019).
- [72] Pitarque A, Ruiz JC, Roy JF. Las redes neuronales como herramientas estadísticas no paramétricas de clasificación. *Psicothema* 2000;12:459–63.
- [73] Wikipedia contributors. Deep learning. Wikipedia, The Free Encyclopedia 2019. https://en.wikipedia.org/w/index.php?title=Deep_learning&oldid=904635431 (accessed July 5, 2019).
- [74] Deng L, Yu D. Deep Learning: Methods and Applications. *Found Signal Process Commun Netw* 2014;7:197–387.
- [75] Sak H, Senior A, Beaufays F. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. *arXiv [csNE]* 2014.
- [76] Necula S-C. Testing the Quality of a Semantic Web Database. *ResearchGate*, 2012.
- [77] Sándor Á, de Waard A. Identifying claimed knowledge updates in biomedical research articles. *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, Association for Computational Linguistics*; 2012, p. 10–7.
- [78] Teufel S, Moens M. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics* 2002;28:409–45. doi:10.1162/089120102762671936.
- [79] Cohan A, Soldaini L, Goharian N. Matching Citation Text and Cited Spans in Biomedical Literature: a Search-Oriented Approach. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2015*. doi:10.3115/v1/n15-1110.
- [80] Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* 2019;571:95.
- [81] Qi T, Wu Y, Zeng J, Zhang F, Xue A, Jiang L, et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat Commun* 2018;9:2282.
- [82] Burnier M, Phan O, Wang Q. High salt intake: a cause of blood pressure-independent left ventricular hypertrophy? *Nephrol Dial Transplant* 2007;22:2426–9.
- [83] Cooke DW, Bennett BL, Farnum EH, Hults WL, Sickafus KE, Smith JF, et al. SiOx luminescence from light-emitting porous silicon: Support for the quantum confinement/luminescence center model. *Appl Phys Lett* 1996;68:1663–5.
- [84] Min-Yen KAN. The Computational Linguistics Scientific Summarization Shared Task (CL-SciSumm 2018) n.d. <http://wing.comp.nus.edu.sg/~cl-scisumm2018/> (accessed May 27, 2019).
- [85] Holley JW, Guilford JP. A Note on the G Index of Agreement. *Educ Psychol Meas* 1964;24:749–53.
- [86] Jolliffe I. Principal Component Analysis. *International Encyclopedia of Statistical Science*, 2011, p. 1094–6.
- [87] Dunham MH. *Data Mining: Introductory And Advanced Topics*. Pearson Education India; 2006.
- [88] Prieto M. Certainty Corpus March, 5, 2019. https://github.com/Guindillator/Certainty/blob/master/Corpus/Complete_statements.txt (accessed May 17, 2019).
- [89] Aziz N, Zain Z, Mafuzi RMZR, Mustapa AM, Najib NHM, Lah NFN. Relative importance index (RII) in ranking of procrastination factors among university students. vol. 1761, Author(s); 2016, p. 020022.
- [90] Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;70:213–20.
- [91] Xu S, Lorber MF. Interrater agreement statistics with skewed data: Evaluation of alternatives to Cohen's kappa. *J Consult Clin Psychol* 2014;82:1219–27.
- [92] Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med* 2005;37:360–3.
- [93] Landis JR, Richard Landis J, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 1977;33:159.
- [94] Deery C, Wagner ML, Longbottom C, Simon R, Nugent ZJ. The prevalence of dental erosion in a United States and a United Kingdom sample of adolescents. *Pediatr Dent* 2000;22:505–10.

- [95] Lix LM, Yogendran MS, Shaw SY, Burchill C, Metge C, Bond R. Population-based data sources for chronic disease surveillance. *Chronic Dis Can* 2008;29:31–8.
- [96] Narayanan A, Greco M, Powell H, Bealing T. Measuring the quality of hospital doctors through colleague and patient feedback. *Journal of Management & Marketing in Healthcare* 2011;4:180–95.
- [97] Campbell J, Narayanan A, Burford B, Greco M. Validation of a multi-source feedback tool for use in general practice. *Educ Prim Care* 2010;21:165–79.
- [98] Sauvageot N, Guillaume M, Albert A, Others. Validation of the food frequency questionnaire used to assess the association between dietary habits and cardiovascular risk factors in the NESCAV study. *J Food Sci* 2013;3.
- [99] Narayanan A, Greco M, Reeves P, Matthews A, Bergin J. Community pharmacy performance evaluation: Reliability and validity of the Pharmacy Patient Questionnaire. *International Journal of Healthcare Management* 2014;7:103–19.
- [100] Gauthier TD. Detecting Trends Using Spearman's Rank Correlation Coefficient. *Environ Forensics* 2001;2:359–62.
- [101] Raithel J. *Quantitative Forschung: Ein Praxiskurs*. Springer-Verlag; 2008.
- [102] Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: AnRPackage for Determining the Relevant Number of Clusters in a Data Set. *J Stat Softw* 2014;61. doi:10.18637/jss.v061.i06.
- [103] Chouikhi H, Charrad M, Ghazzali N. A comparison study of clustering validity indices. 2015 Global Summit on Computer Information Technology (GSCIT), 2015, p. 1–4.
- [104] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011;12:2825–30.
- [105] Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics* 2003;19:1699–706.
- [106] Chen T, Guestrin C. XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016. doi:10.1145/2939672.2939785.
- [107] Pavlov YL. *Random Forests*. VSP; 2000.
- [108] Werbos P. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. 1974.
- [109] Brownlee J. *A gentle introduction to XGBoost for applied machine learning* 2016.
- [110] Brownlee J. How to Tune the Number and Size of Decision Trees with XGBoost in Python. *Machine Learning Mastery* 2016. <https://machinelearningmastery.com/tune-number-size-decision-trees-xgboost-python/> (accessed July 22, 2019).
- [111] basucode n.d. <https://www.kaggle.com/pranav93/basucode> (accessed July 22, 2019).
- [112] Koehrsen W. Random Forest in Python - Towards Data Science. Medium 2017. <https://towardsdatascience.com/random-forest-in-python-24d0893d51c0> (accessed July 23, 2019).
- [113] Wang Y, Huang M, Zhao L, Others. Attention-based LSTM for aspect-level sentiment classification. *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, p. 606–15.
- [114] Baziotis C, Pelekis N, Doukeridis C. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 2017, p. 747–54.
- [115] Ma Y, Peng H, Cambria E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [116] Hersh W, Voorhees E. TREC genomics special issue overview. *Inf Retr Boston* 2008;12:1–15.
- [117] Cohan A, Goharian N. Scientific document summarization via citation contextualization and scientific discourse. *International Journal on Digital Libraries* 2018;19:287–303.
- [118] Chiu B, Crichton G, Korhonen A, Pyysalo S. How to Train good Word Embeddings for Biomedical NLP. *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, 2016. doi:10.18653/v1/w16-2922.
- [119] Crestan E, Pantel P. Web-scale Knowledge Extraction from Semi-structured Tables. *Proceedings of the 19th International Conference on World Wide Web*, New York, NY, USA: ACM; 2010, p. 1081–2.
- [120] Snow R, O'Connor B, Jurafsky D, Ng AY. Cheap and Fast---but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics; 2008, p. 254–63.
- [121] Lewis RJ. An introduction to classification and regression tree (CART) analysis. *Annual meeting of the society for academic emergency medicine in San Francisco, California*, vol. 14, 2000.
- [122] Chollet F, Others. Keras. Keras 2015. <https://keras.io> (accessed May 25, 2019).
- [123] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *CoRR* 2016;abs/1603.04467.
- [124] Tong S, Koller D. Support Vector Machine Active Learning with Applications to Text Classification. *J Mach Learn Res* 2001;2:45–66.
- [125] Ghoneim S. Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on? Medium 2019. <https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124> (accessed July 22, 2019).
- [126] Garg R, Oh E, Naidech A, Kording K, Prabhakaran S. Automating Ischemic Stroke Subtype Classification Using Machine Learning and Natural Language Processing. *J Stroke Cerebrovasc Dis* 2019.

- doi:10.1016/j.jstrokecerebrovasdis.2019.02.004.
- [127] Prieto. Guindillator/Certainty. GitHub Guindillator/Certainty 2019. <https://github.com/Guindillator/Certainty> (accessed May 29, 2019).
 - [128] Hounbo KH. Investigating Citation Linkage Between Research Articles. Western University, 2017.
 - [129] De Waard A, Schneider J. Formalising uncertainty: An ontology of reasoning, certainty and attribution (ORCA). On Semantic Technologies Applied to 2012.
 - [130] Nguyen V, Bodenreider O, Sheth A. Don't Like RDF Reification? Making Statements about Statements Using Singleton Property. Proc Int World Wide Web Conf 2014;2014:759–70.
 - [131] FAIRtools. PyPI n.d. <https://pypi.org/project/FAIRtools/> (accessed August 6, 2019).
 - [132] Askanas V, McFerrin J, Baqué S, Alvarez RB, Sarkozi E, Engel WK. Transfer of beta-amyloid precursor protein gene using adenovirus vector causes mitochondrial abnormalities in cultured normal human muscle. Proc Natl Acad Sci U S A 1996;93:1314–9.
 - [133] Askanas V, Engel WK, Alvarez RB. Light and electron microscopic localization of beta-amyloid protein in muscle biopsies of patients with inclusion-body myositis. Am J Pathol 1992;141:31–6.
 - [134] Askanas V, McFerrin J, Alvarez RB, Baqué S, Engel WK. β APP gene transfer into cultured human muscle induces inclusion-body myositis aspects. Neuroreport 1997;8:2155.
 - [135] Askanas V, Engel WK. New advances in the understanding of sporadic inclusion-body myositis and hereditary inclusion-body myopathies. Curr Opin Rheumatol 1995;7:486–96.
 - [136] Askanas V, Engel WK. Inclusion-body myositis, a multifactorial muscle disease associated with aging: current concepts of pathogenesis. Curr Opin Rheumatol 2007;19:550–9.
 - [137] Lünemann JD, Schmidt J, Schmid D, Barthel K, Wrede A, Dalakas MC, et al. Beta-amyloid is a substrate of autophagy in sporadic inclusion body myositis. Ann Neurol 2007;61:476–83.
 - [138] Askanas V, Engel WK. Inclusion-body myositis: a myodegenerative conformational disorder associated with Abeta, protein misfolding, and proteasome inhibition. Neurology 2006;66:S39–48.
 - [139] Askanas V, Engel WK. Proposed pathogenetic cascade of inclusion-body myositis: importance of amyloid-beta, misfolded proteins, predisposing genes, and aging. Curr Opin Rheumatol 2003;15:737–44.
 - [140] Baron P, Galimberti D, Meda L, Scarpini E, Conti G, Cogiamanian F, et al. Production of IL-6 by human myoblasts stimulated with Abeta: relevance in the pathogenesis of IBM. Neurology 2001;57:1561–5.
 - [141] Hollingworth HL. The Central Tendency of Judgment. The Journal of Philosophy, Psychology and Scientific Methods 1910;7:461–9.
 - [142] Huttenlocher J, Hedges LV, Vevea JL. Why do categories affect stimulus judgment? J Exp Psychol Gen 2000;129:220–41.
 - [143] Duffy S, Huttenlocher J, Hedges LV, Crawford LE. Category effects on stimulus estimation: shifting and skewed frequency distributions. Psychon Bull Rev 2010;17:224–30.

9. SUPPLEMENTAL INFORMATION

Horn’s Parallel Analysis

The optimal number of principal components was selected using Horn’s parallel analysis to the certainty categories of the 3 questionnaires. Analysis was carried out using *paran* function of the R package *paran*. Details of the function are specified in a Jupyter Notebooks on Github. Fig. 25, 26 and 27 show the result of the Horn’s parallel analysis.

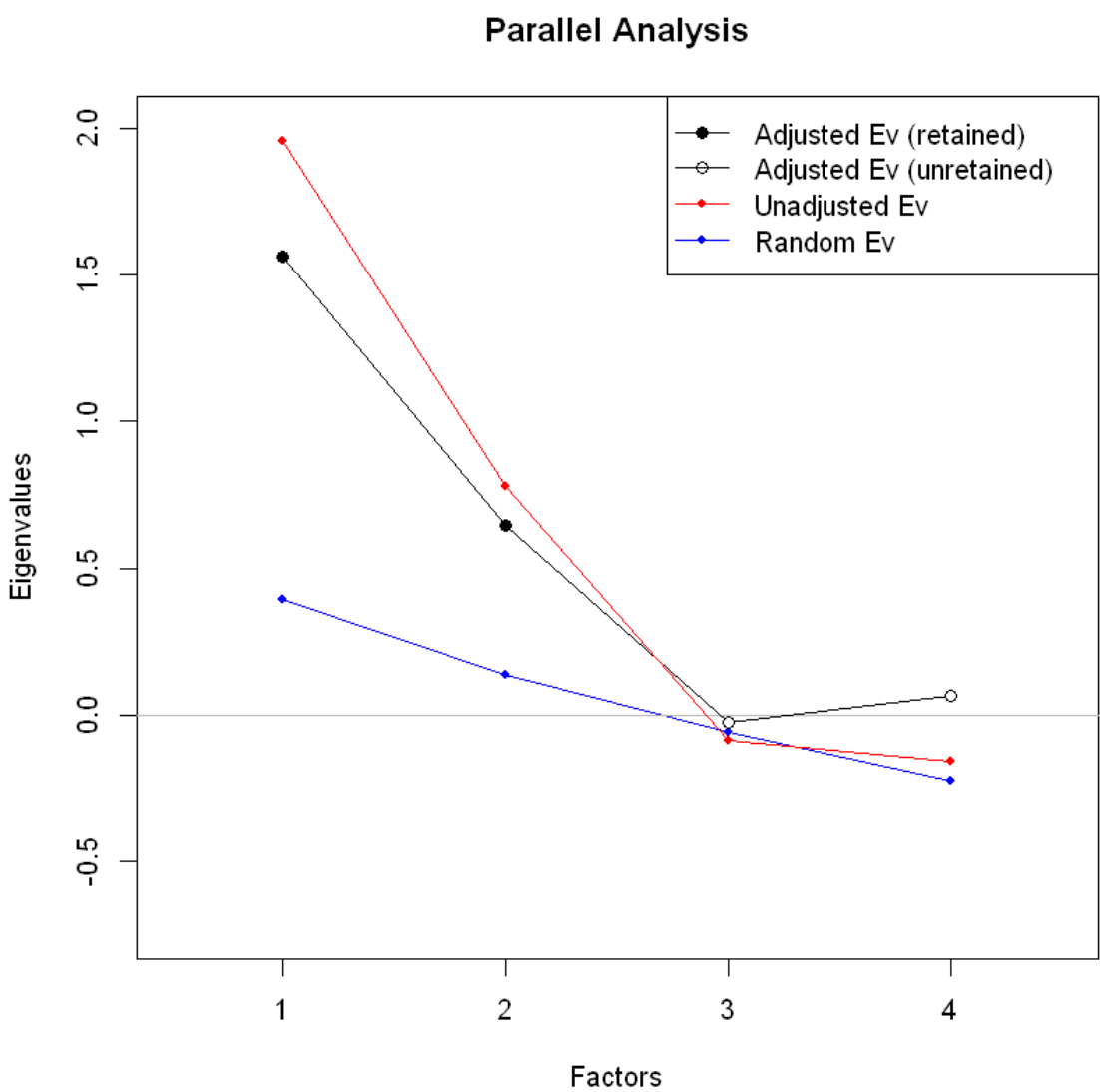


Figure 25. Supplemental Information. Horn’s parallel analysis result for S1

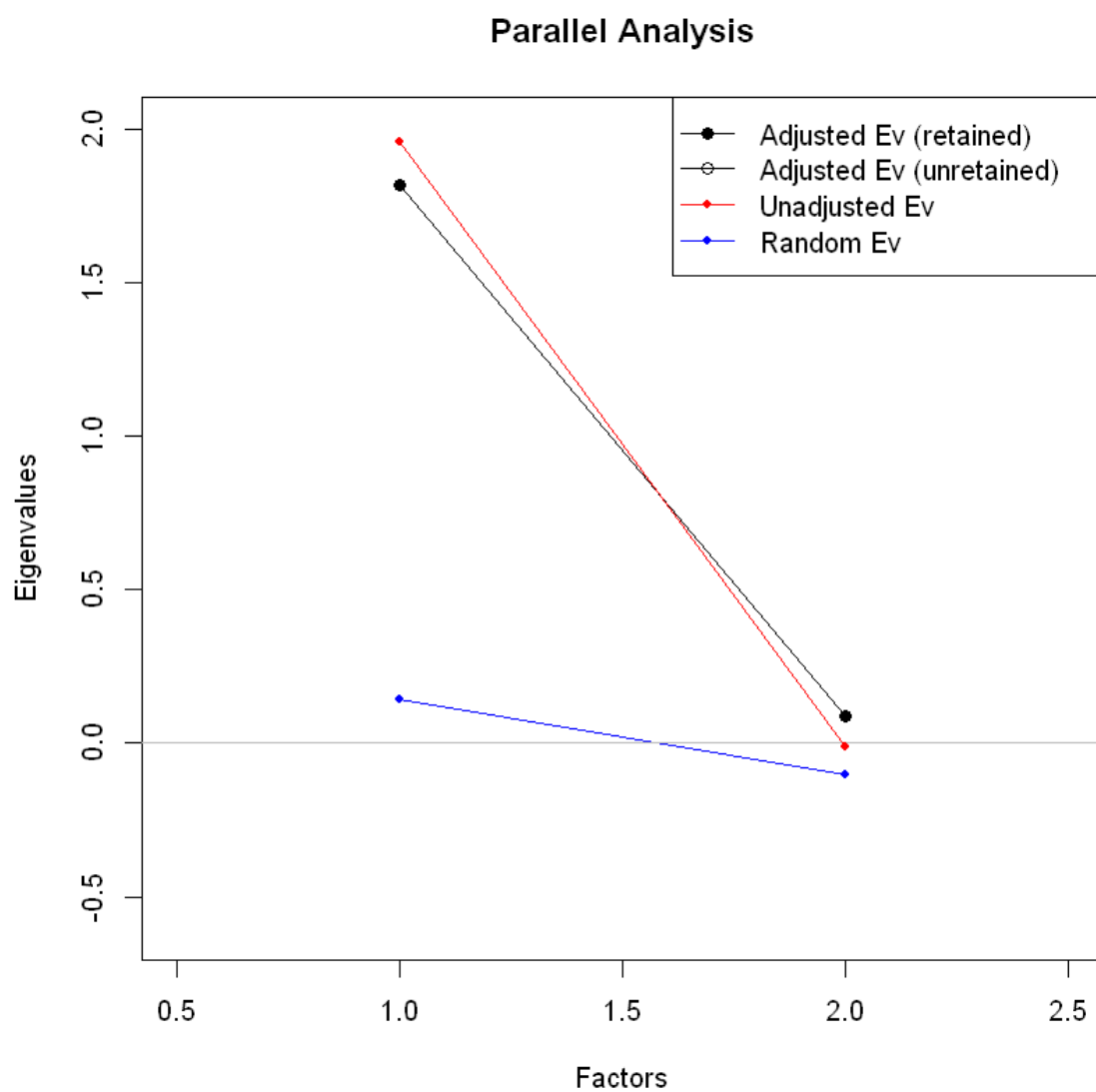


Figure 26. Supplemental Information. Horn's parallel analysis result for S2

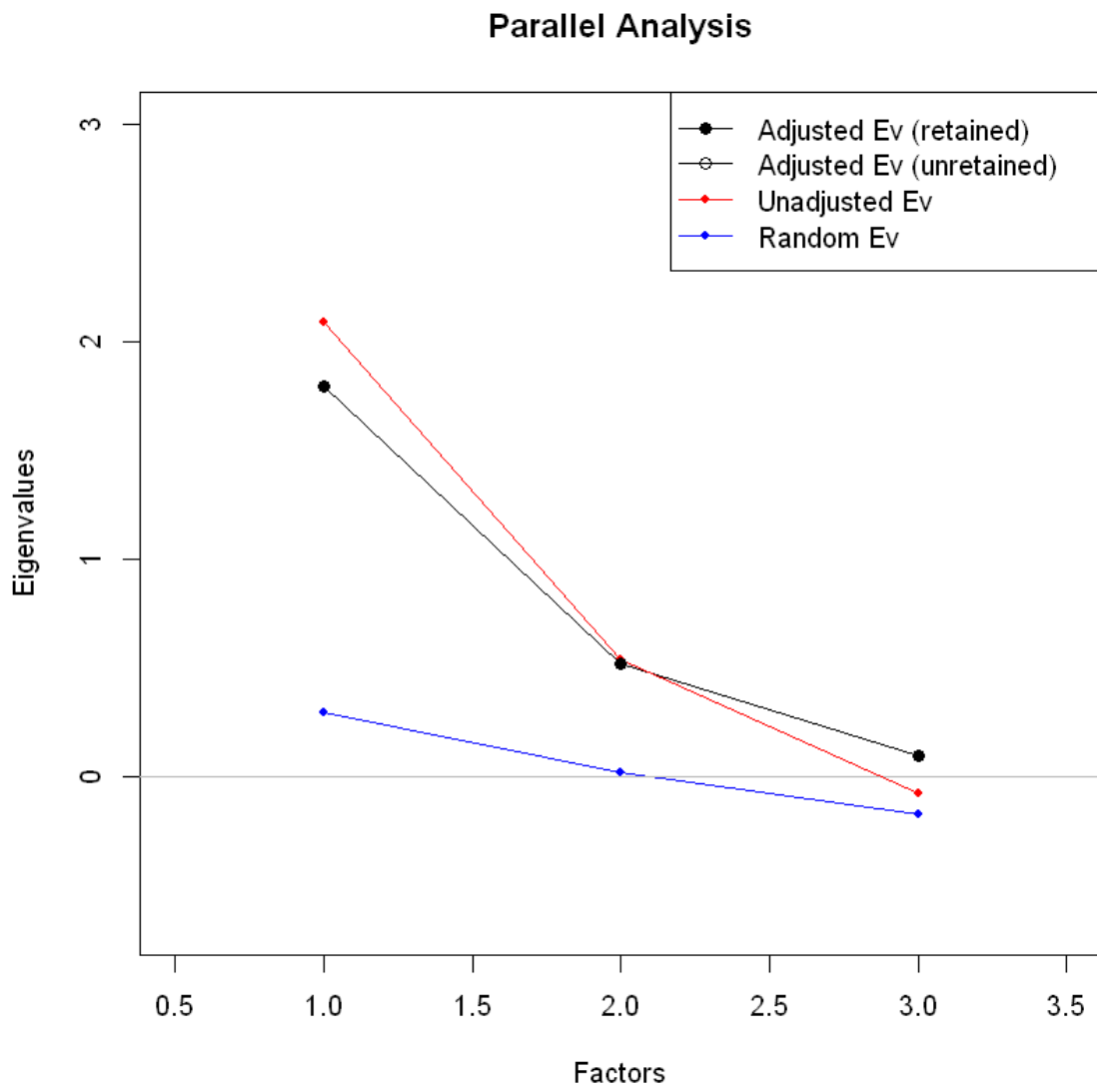


Figure 27. Supplemental Information. Horn's parallel analysis result for S3

9.1. List of Tables

Table 1: Comparison of corpora and approaches used in prior investigations into scholarly certainty.

Table 2: Passive voice versus active voice examples in experimental sciences.

Table 3: Hypothesis and their corresponding certainty research areas.

Table 4: Final machine learning model parameters. *Loss and Optimizer are required parameters to compile the model.

Table 5: Categorization Consistency of Statements (by Statement number) for Survey S1

Table 6: Categorization Consistency of Statements (by Statement number) for Survey S2

Table 7: Categorization Consistency of Statements (by Statement number) for Survey S3

Table 8: Interpretation of Spearman correlation [87,101].

Table 9: Jaccard Similarity clusters resulting from K-Means applied to questionnaire results.

Table 10: Analysis of Principal Components of question 1 for S1 Statement Classifications

Table 11: Analysis of differences between groups using Kruskal-Wallis (> 2 groups) and U-Mann Whitney test (2 groups).

Table 12: Analysis of differences between groups using Independence Test.

Table 13: Analysis of differences between question 1 and question 2 using U-Mann Whitney test (2 groups).

Table 14: Responses to the question about “basis” in S1.

Table 15: Responses to the question about “basis” in S2.

Table 16: Responses to the question about “basis” in S3.

Table 17: Performance of our manual classification for S1 on the original publicly-annotated 45 statements

Table 18: Performance of our manual classification for S2 on the original publicly-annotated 45 statements

Table 19: Performance of our manual classification for S3 on the original publicly-annotated 45 statements

Table 20: Cross-validation results of machine learning model.

Table 21: Performance of the neural network model (right) and our manual classification for S3 (left) on the original publicly-annotated 45 statements

9.2. List of Figures

Figure 1: The Flow of Scholarly Discourse.

Figure 2: How a claim becomes a fact.

Figure 3: Neural Network structure.

Figure 4: Example of the Survey 1 questionnaire introduction.

Figure 5: Example of the Survey 1 questionnaire interface. Question 1.1.

Figure 6: Example of the Survey 1 questionnaire interface. Question 1.2.

Figure 7: Example of the Survey 1 questionnaire interface's final question.

Figure 8: Cosine Similarity explanation.

Figure 9: Distribution of the statements classified and G index agreement in S1.

Figure 10: Distribution of the statements classified and G index agreement in S2.

Figure 11: Distribution of the statements classified and G index agreement in S3.

Figure 12: Spearman Rank Correlation, hierarchically-clustered heatmap (A) and their respective *p-values* (B) comparing the statements assigned to the Certainty Categories among all three questionnaires.

Figure 13: Majority rule output for deciding the optimal number of clusters (k) in the three Surveys.

Figure 14: Clustering analysis of k-means results for S1 using *Elbow*, *Silhouette* and *GAP statistic* method. Dot-line represent the optimal cluster chosen by each method.

Figure 15: Clustering analysis of k-means results for S2 using *Elbow*, *Silhouette* and *GAP statistic* method. Dot-line represent the optimal cluster chosen by each method.

Figure 16: Clustering analysis of k-means results for S3 using *Elbow*, *Silhouette* and *GAP statistic* method. Dot-line represent the optimal cluster chosen by each method.

Figure 17: Principal component analysis of questionnaire responses of question 1 in the three Surveys.

Figure 18: Principal component analysis of Question 2 responses from the three Surveys.

Figure 19: Basis and certainty correlation for S1.

Figure 20: Basis and certainty correlation for S2.

Figure 21: Basis and certainty correlation for S3.

Figure 22: Comparison of the participants' responses against our manual classification for S1 (A), S2 (B) and S3 (C).

Figure 23: An exemplar prototype NanoPublication including certainty annotations.

Figure 24: Automated classification of scholarly assertions related to the accumulation of beta-APP protein in muscle fibres, color coded as green (Category A - highest certainty), orange (Category B - medium certainty) and red (Category C - lowest certainty).

Figure 25: Supplemental Information. Horn's parallel analysis result for S1.

Figure 26. Supplemental Information. Horn's parallel analysis result for S2.

Figure 27. Supplemental Information. Horn's parallel analysis result for S3.

