9

Reading Certainty across Sources

Benjamin Miller

Abstract. Witness testimony provides the first draft of history and requires a kind of reading connecting descriptions of events from many perspectives and sources. This chapter examines one critical step in that connective process, namely, how to assess a speaker's certainty about the events they describe. By surveying a group of approximately 300 readers and their approximately 28,000 decisions about speaker certainty, this chapter explores how readers may think about factual and counterfactual statements and how they interpret the certainty with which a witness makes their statements. Ultimately, this chapter argues that readers of collections of witness testimony were more likely to agree about event descriptions when those providing the description were certain and that readers' abilities to accept gradations of certainty were better when a witness described factual, rather than counterfactual or negated events. These findings lead to a suggestion for how researchers in natural language processing could better model the question of speaker certainty, at least when dealing with the kind of narrative nonfiction one finds in witness testimony.

9.1 Introduction

Understanding and researching the impact of human rights violations, environmental disasters, and other types of collective trauma rely on reading large collections of witness statements and connecting the stories therein. These stories and the events they describe are evocative because of both the individual events they relate and their potential connections as they enrich, substantiate, or contradict the stories of others. Reading across witness statements and other sources is an essential task, one that requires making many small judgments about features like the relevance and reliability of sources (Martin, 2017).

Though that kind of connective reading could be supported by computational approaches, these types of sources, namely, witness statements, present

https://doi.org/10.1017/9781108854221.012 Published online by Cambridge University Press

a few challenges. First, they often indicate space, time, and entities indirectly more so than absolutely. Additionally, they do so with a fragmented syntactic structure and highly referential semantics. These features make them resistant to techniques reliant on named entity or temporal recognition. For example, often, such as in testimonies provided by first responders to the World Trade Center attacks of September 11, 2001, a witness either does not know where or when something specifically happened or they do not have the language with which to talk about it. In "World Trade Center Task Force Interview No. 9110335," an EMT says, "That's when we noticed a whole bunch of police cars responding somewhere" ("The September 11 Records", 2005). "We immediately jumped back into the vehicle," the same EMT says, "back into my car, and we get to the station"; at no point in that interview does the EMT indicate the specific name of the station she means. Although she relates the event with *certainty*, a critical term for this study, the geography of the event is poorly specified.

However difficult these stories are to read, these same witness testimonies from traumatic events such as environmental disasters, industrial accidents, social injustice, and attacks of terror provide the material with which the first draft of history is written. In the aftermath of these events, statements from witnesses often provide the public with their first glimpse into what happened. At later stages of individual and collective processing of traumatic events, long after the events they describe are concluded, these statements can serve to anchor collective understanding. By anchoring collective understanding, they work to limit people's ability to dissimulate and misuse events for political or personal gain, while providing emblematic voices that help communities process and survive these assaults.

Speaker certainty is one critical predicate for understanding the events that comprise these stories. Certainty, also known as veridicality, is the extent to which the speaker is certain about the statement they put forward. Many questions about the computational reading of event language is conditioned on whether or not it might be something an algorithm can be trained to identify, in addition to whether that perspective is meaningful relative to understanding the witness, their statement, and the event to which they offer testimony. A second critical predicate in relation to this material's role as an anchor of collective memory is the speaker's statement's facticity, or whether an event is being described or negated. Combined, these two measures provide a first step in ascertaining whether the description of an event in one witness testimony can be legitimately connected to an event description from another testimony. It can also provide a sense for how witnesses narrate and think about the events they observed or survived. Scholars of trauma and witnessing such as

Caruth (2016) and Herman (2015) describe how the linguistic features of a witness's testimony, such as the degree of specificity of their references and their degree of certainty, indicates aspects of their cognitive and mental state. In combination with observations of how readers interpret these stories of events, this sense can help us understand how readers approach this kind of difficult historical material.

To explore this idea, a large study was conducted that posed two related questions. First, how are certainty and uncertainty indicated in the language of witness statements? Second, how do readers interpret those statements? This study builds on the work in the area of using natural language processing (NLP) to quantify a speaker's certainty about their statements (Hyland, 2005; Saurí and Pustejovsky, 2009; De Marneffe et al., 2012; Wan and Zhang, 2014; Lee et al., 2015; Stanovsky et al., 2017). Prior work specifically on witness testimony is limited, with examples focusing on emotions (Truong et al., 2014), information extraction (Divitaa et al., 2018), collocations (Nugumanova and Bessmertny, 2013), cross-document coreference (Miller et al., 2013), narrative segmentation (Miller et al., 2015), domain-specific problems (Gibbons, 2014), social media (Soni et al., 2014), and news (Wan and Zhang, 2014), but, to my knowledge, none on the messier domain of witness statements and veridicality. These studies undertook a variety of approaches.

Some, like Soni et al. (2014) and Hahn and Engelmann (2014), pursued rule-based approaches that focused on quantifying usage of hedges, modals, and modifiers. What is valuable about these two studies is their focus on linguistic markers of certainty and uncertainty and their shared argument that interpretive discrepancies may emerge from "the interpretative hardness of the linguistic items in question" (Hahn and Engelmann, 2014). Others, like Wan and Zhang (2014), began with the annotated corpus from Saurí and Pustejovsky (2009) but then proceeded to build a classifier relying on their own enriched annotated data developed with a simplified five-point schema ranging from *Very Certain* to *Very Uncertain*. Though their distribution of scores is interesting, with only 1.3% of their 1,000 examples falling into the *Very Uncertain* category, they only used two raters per item, and their schema may oversimplify a critical aspect of veridicality. Namely, it ignores whether the event being described is being posited, and so took place, or negated, and so did not take place. As I will argue below, that difference seems to have an impact on the degree of certainty a speaker communicates. Along with the first finding of this chapter, a proposal for a new schema for the annotation of veridicality, that recognition of a kind of interpretive difference when considering posited versus negated events forms the second key finding of this chapter.

An additional challenge to those looking to adopt a computational approach to reading across witness statements is that they are often in nonstandard English or a mixture of languages. This polyphony makes them resistant to syntactical and semantic approaches, such as one reliant on quantifying uses of hedges, such as *maybe*, and modals, such as *could be*. In one testimony to the South African Truth and Reconciliation Commission, the witness says about her unlawful detainment, "When I was about to – to give a response to one of the questions, the other one said I am *spoggerig* – they kept on interrogating me for hours on end" (SABC, 1996). And, because they are often stories of traumatic, or at least challenging, events, their grammatical and narrative composition can reflect the psychologically difficult moments of their creation.

Scale presents an additional challenge. Collections such as the one resulting from the South African Truth and Reconciliation Commission contain thousands of testimonies, the above-referenced World Trade Center Task Force interviews are comprised of approximately 17,000 question and answer pairs, and government repositories like the Guatemalan National Police Historical Archive contain many millions of witness statements. Combined, these various challenges make the understanding of event language in real-world documents a meaningful but difficult research domain in computational linguistics. Nevertheless, the scale of these collections and the very close reading required to read across documents necessitate a computational approach. The effort put forward by various truth and reconciliation processes indicates that these materials serve vitally important functions for people and communities.

To assess the viability of current approaches to veridicality and facticity in the domain of witness testimony, I undertook a multistep process of (1) corpus building that drew on many different collections of real witness statements, (2) event detection and annotation, (3) cloud labor-based veridicality annotation, (4) preliminary categorization and interpretation by simple interrater reliability, (5) a more robust categorization by k-means clustering of mean ratings per item, and, finally, (6) a comparative interpretation of the results.

As a result of this research on 27,800 ratings of 2,490 witness descriptions of events, I propose two findings: first, a revision of an existing classification schema for veridicality and, second, a conceptual finding about how readers process descriptions of events differently when the events are about facts or negations of facts.[1]

[1] Project data are available at https://github.com/bjmiller16/witness-veridicality

https://doi.org/10.1017/9781108854221.012 Published online by Cambridge University Press

Table 9.1. *Veridicality categories*

Positive or nonnegated		
1	Certain+	According to the speaker, it is certainly the case that
2	Probable+	According to the speaker, it is possibly the case that
3	Possible+	According to the speaker, it is possibly the case that
Negative or negated		
4	Certain–	According to the speaker, it is certainly not the case that
5	Probable–	According to the speaker, it is probably not the case that
6	Possible–	According to the speaker, it is possibly not the case that
Underspecified		
7	Certain but underspecified	The speaker knows but does not fully communicate whether or not it is the case that
8	Uncertain	The speaker does not know or does not commit to whether or not it is the case that
Error		
9	Error	There is something wrong with the sentence

The first finding is that the nine-element annotation schema put forward by Saurí and Pustejovsky (2009, 2012) does not describe this evidence. That schema, based on the work of Horn (1989), offers two types of polarity: facts and counterfacts. Each polarity has three degrees of certainty. Additionally, there are two more categories for partially or fully underspecified statements. The schema is further described in Table 9.1. Though theoretically sound, it presents a perspective on speaker certainty implying that statements, witnesses, and readers possess equal and opposite gradations for factual and counterfactual statements. Based on the evidence of the annotated witness statements, I argue that witnesses, and their annotators, have a greater sensitivity to gradations of certainty of facts, or positive evidence, than they do to certainty of counterfacts, or negative evidence. The second finding is that annotators, and the statements they read, are more sure of their ratings of facts than they are of uncertain statements or of counterfacts.

These observations suggest that an unbalanced categorization of veridicality and broader categories for more uncertain and counterfactual events would be more reflective of how events are presented by witnesses to traumatic events.

https://doi.org/10.1017/9781108854221.012 Published online by Cambridge University Press

These findings suggest that NLP and computational social sciences should process assessments of veridicality with more allowances for counterfactual or uncertain statements and more gradations of certainty and positive evidence. To that end, I propose a modification to Saurí and Pustejovsky's schema. This new nine-element classification schema does away with one gradiation of certainty for counterfacts, adds one additional gradation of certainty for facts, and revises the underspecified categories as certain unknowns and uncertain unknowns. This new schema better reflects the evidence from this study and potentially better describes how readers process information about events in witness testimony.

9.2 Background

The problem addressed here involves the linguistic concept of event certainty or veridicality. This concept arises from the observation that besides communicating propositional information, such as who did what to whom, language users routinely communicate other information about their propositions such as their attitudes toward the propositional information (Hyland, 2005). In particular, though every description of an event contains some combination of propositional information, specifying actors, acts, time, and location, language users make use of various linguistic mechanisms that allow them to commit more or less strongly to the information they present in their utterances.

For example, Hyland (2005) distinguishes between hedges and boosters. Speakers and writers use hedges to communicate that they are not fully certain about a proposition. Hedges are commonly communicated through features like modal verbs (e.g., *might* and *may*) or adverbs (e.g., *perhaps* and *possibly*). In contrast, boosters allow the speaker to more fully commit to being certain about a proposition. Commonly, boosters are adverbs such as "definitely." These examples, however, are far from exhaustive. Prior research (Saurí and Pustejovsky, 2009; De Marneffe et al., 2012) suggests that a large inventory of linguistic features is necessary to even begin approaching a comprehensive description of how language users communicate veridicality. Stanovsky et al. (2017) go further to suggest that this type of dictionary approach is of limited generalizability and a phraseological approach would be of more value.

More specifically for the purpose of making quantitative generalizations about the contents of corpora, veridicality presents an issue for projects attempting to use NLP, a common research tool in the computational social

sciences. For example, our own research attempts to derive summative analyses of human rights violation events by automatically processing the testimony of many eyewitnesses stored in digitized corpora. One potential threat to the validity of these analyses is that tools for named event recognition identify potentially relevant elements in discourse without regard for whether the speaker is positing a fact or counterfact about an event. Thus, such tools may not make any distinction between the events represented by the verb "kill" in Examples 9.1 and 9.2, even though for work on witness statements the two need to be distinguished.

He definitely did kill that person. (9.1)

He definitely did **not** *kill that person.* (9.2)

Annotation of the corpora would need to indicate that 9.1 is a report of a human rights violation event, whereas 9.2 is not. Automatically distinguishing between these two requires an NLP tool that is able to make judgments about facticity. Past research has already been undertaken on the general form of this problem by Saurí and Pustejovsky (2009), De Marneffe et al. (2012), and Stanovsky et al. (2017), but the text types used in the first two were constrained to newspaper articles, a well-structured, standardized genre. The third looked to develop a more generalizable method for assessing factuality by moving away from a balanced eight-category approach for assessing factuality to a single numeric value. Unfortunately, that singular value, though easier to calculate, conflates the multidimensionality presented by the problem of veridicality; namely, that it is a measure of both certainty and facticity. These data, transcribed oral interviews, differ greatly in composition from newspaper articles and serve to highlight a potential shortcoming in the initial nine-category schemas used by these projects, namely, that schema are symmetrical, offering the same kinds of choices for readers about statements describing events and statements describing the negation of an event. Though elegant, it is possible that a reader does not have the same ability to discriminate about concrete statements as they do about more ambiguous ones or about statements of facts versus statements about counterfacts. Therefore, to help better understand how people think about speaker certainty in writing about events, this work extends and refines prior research by testing already established computational approaches in a new context in ways that highlight the shortcomings of prior schemas.

https://doi.org/10.1017/9781108854221.012 Published online by Cambridge University Press

9.3 Methods

Following the example of Saurí and Pustejovsky (2009) and De Marneffe et al. (2012), a questionnaire was created to gather judgments of veridicality from Amazon Mechanical Turk users. mTurk is a cloud-labor platform that connects workers with information processing tasks. For this task, first, sentences were extracted from different corpora of interviews from different contexts: the South African Truth and Reconciliation Commission, the Cambodian Khmer Rouge Tribunal, interviews with survivors of the Holocaust, statements from survivors of the Rwandan genocide, and interviews with survivors of ethnic cleansing in the former Yugoslavia. One goal of this study was to use real-world data, rather than simulated data. Though it can be argued that simulated data would allow for a stronger statement to be made about the quantitative findings and sources of variation, it would not reflect how witnesses use language or how readers grapple with the complex problem of understanding witness testimony.

Readers' attention was focused on the events described by each statement. Using EVITA (Saurí et al., 2005), events in each sentence were tagged. The author chose to use EVITA, as opposed to a more contemporary event tagger, because the tool was sufficient for generating candidate event sentences. Only sentences containing *OCCURRENCE* or *STATE* event tags were retained as candidates for the questionnaire. An *OCCURRENCE* tag implied that the event was a specific action that took place over a defined period of time. A *STATE* event describes an ongoing condition, rather than an action. A random sample of sentences containing about 800 events was taken from each of the five corpora, leading to a total of approximately 4,000 events.

Second, sentences were preprocessed so that each event in a multi-event sentence was identified and a different iteration of the sentence was created to highlight (through bolding) the individual event. In this manner, a sentence that contained two events like Example 9.3 would be rendered as two items in our questionnaire: Examples 9.4 and 9.5.

He ran up the street and jumped on the train. (9.3)

He **ran** *up the street and jumped on the train.* (9.4)

He ran up the street and **jumped** *on the train.* (9.5)

As previously stated, a total of 2,490 items similar to Examples 9.4 and 9.5 were contained in our overall bank of items for our questionnaires.

Third, a series of different versions of the questionnaire were created that first trained mTurk users in how to make judgments about certainty in the

https://doi.org/10.1017/9781108854221.012 Published online by Cambridge University Press

sentences and then asked them to place items like Examples 9.4 and 9.5 above in one of nine categories presented in Table 9.1. A total of 227 unique raters participated in the study, providing a total of 27,800 individual event ratings. On average, each rater accounted for 0.44% of the total by providing 122.5 ratings with a standard deviation of 193.3. Each rater was provided with a slate of 55 survey items to categorize, of which 5 were training and norming items, 49 were unknown items to rate, and 1 was a known answer question instructing raters to select a particular option. Informal estimates suggested that a rater fluent in English could comfortably complete 5 items per minute. In all, 556 surveys were completed, of which 56 were rejected for failing to correctly answer the known question item. For each accepted set of ratings, raters were paid six dollars (in 2014). Each survey was limited to 10 accepted assignments, and raters could not work on the same survey twice. For each survey, raters were provided with a description of the task, a consent form, brief descriptions of each of the categories identical to those provided in 9.1, and instructions to rate how certain each speaker is about the key event bolded in each sentence item.

Once the rating results were returned, cleaned, and concatenated with their sentences, the categories were replaced with numeric values. Drawing on the work of Stanovsky et al. (2017) and Saurí and Pustejovsky (2012), values divisible by three were selected for the x-dimension of fact-to-counterfact, or facticity, and the y-dimension of certainty. A *certain+* rating received a value of (30; 30), reflecting maximal facticity and maximal certainty. A *certain–* rating received a value of (−30; 30). Examples of *uncertain* received a rating of (0; −30). Item means were calculated based on these values and the resulting data clustered using these two means. The number of clusters was increased until the graph of clustered to unclustered items revealed an "elbow" shape (Kodinariya and Makwana, 2013). One of the oldest methods for determining the correct number of clusters, k, the elbow method relates the potential number of clusters to sum of squares distances from the data points to the cluster centers. As one increases the number of possible clusters, the sum of squares decreases toward zero. Often there is a value of k at which the value of the sum of squares slows its rate of decrease. That change shows up as an "elbow" in the graph. This visual cue indicates that increasing the number of clusters beyond a certain number barely improves the clusters' description of the data.

9.4 Results

Table 9.2 shows the results of using a simple standard of majority agreement for interrater reliability where 6 out of 10 raters selected the same rating

https://doi.org/10.1017/9781108854221.012 Published online by Cambridge University Press

Table 9.2. *Simple majority interrater reliability agreement results*

Certain+	1083
Probable+	370
Possible+	21
Certain−	18
Probable−	5
Possible−	0
Certain but underspecified	0
Uncertain	0
Error	0

Table 9.3. *K-means clustering results*

Cluster	Within SS	Variance	SD	Nomenclature	Certainty	Counter/ Fact	No. of Items
6	4,596.2	71.8	8.5	Certain Fact	27.6	26.6	483
3	7,625.9	119.2	10.9	Probable Fact	22.9	20.2	581
5	5,062.1	79.1	8.9	Likely Fact	16.6	15.6	451
7	4,798.2	75.0	8.7	Possible Fact	10.9	10.0	336
8	3,843.4	60.1	7.7	Certain Unknown	21.9	5.1	117
2	4,106.8	64.2	8.0	NA/Other	2.6	4.6	173
1	3,846.9	60.1	7.8	Uncertain Unknown	−9.7	−0.9	64
9	5,068.0	79.2	8.9	Possible Counterfact	9.5	−7.9	113
4	8,555.1	133.7	11.6	Certain Counterfact	23.4	−16.5	172

category from the eight-bin classification schema. In this schema, "+" denotes a positing of an event as fact and "−" denotes a negation of an event as a counterfact. Using this majority threshold, only 60.1% of results, or 1,497 out of 2,490 items, were categorized. Of those, 72.3% were in the same category, *Certain*+, and 98.5% were in the three categories of fact, *Certain*, *Probable*, or *Possible*. Additionally, of the counterfact or uncertain categories, only *Certain*− and *Probable*− contained any items at all.

Instead of discarding items that prompted disagreement among raters, I propose incorporating those results in the model using the central tendencies of each rated item. Figure 9.1 presents a visualization of the results from Table 9.3 that were based on taking the average and standard deviation of the valid ratings for each item. It yields a different, more complete picture than Table 9.2. Of principal interest is that the items are more evenly distributed

https://doi.org/10.1017/9781108854221.012 Published online by Cambridge University Press

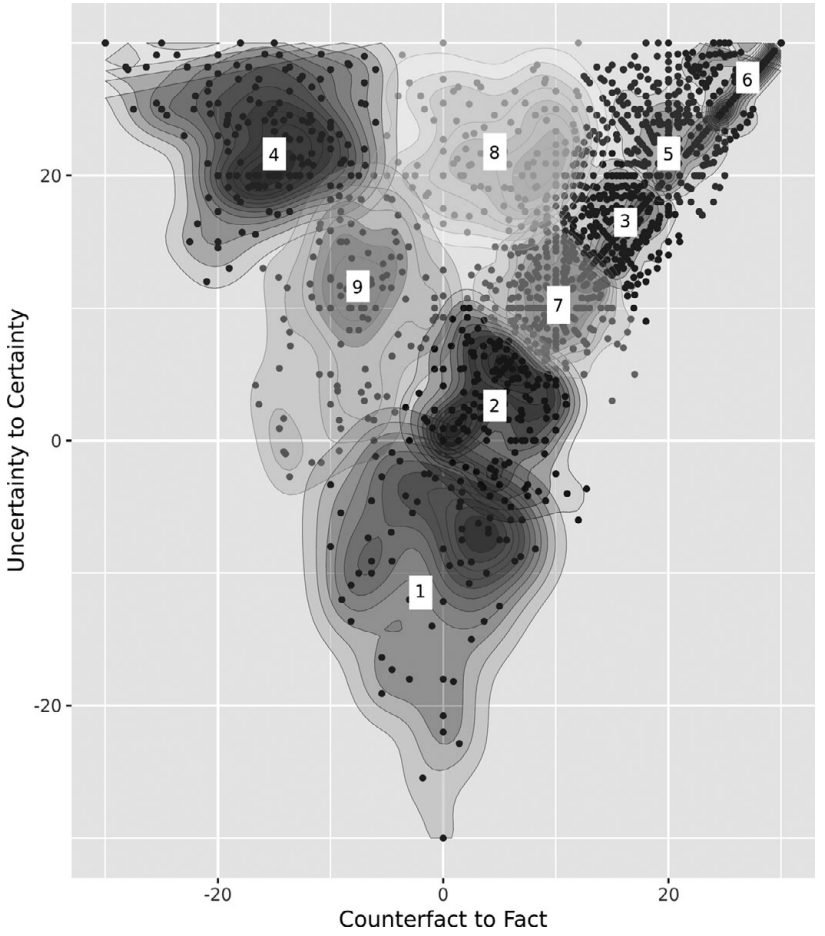

Figure 9.1 Clusters of veridicality.

across the categories and that items that prompted disagreement among raters can still be categorized. Most critical, these aggregate ratings show that in order to describe a reader's perception of veridicality, an additional category of certainty is required for facts and one fewer for counterfacts; that both counterfactual and uncertain categories are well populated with items; and that cluster variance is not directly associated to the number of items per cluster. The fitness of the clustering model as a ratio of the Between sum of squares (SS) /Within SS values is 91.9%.

https://doi.org/10.1017/9781108854221.012 Published online by Cambridge University Press

9.5 Discussion

The simple explanation that the relatively lower number of observations of uncertain or counterfactual statements resulted in a less refined categorization schema is arguable, except for four points. First, the variance within each of the proposed categories does not correlate with the number of observations per category. Rather, the highest variation occurs within the second most certain category, *Probable Fact* (SD = 10.9), and the lowest variation occurs within the *Certain* and *Uncertain Unknown* categories (SD = 7.7, SD = 7.8), each of which possess among the smallest number of observations (n = 64 and n = 117). Second, it ignores the higher cognitive demands negations place on readers and evaluators (Marsh, 1986), demands that increase the likelihood of generating results with higher dispersion. Third, the language of the witnesses and the literature on modality such as Hacquard (2011) indicate that there is a broader range of articulations for uncertainty than there is for certainty. And fourth, there is no apparent reason why speakers would have a balanced observational scale of either fact versus counterfact or of certainty versus uncertainty.

Based on these assertions, I find it more likely that the observation of asymmetry in the coding of witnesses' veridicality is due to features of the language of testimony. It is these features that lead to the categorizations featured in Table 9.4. Consider the first example in the table, a quote from the testimony of Alexander Ehrmann, a survivor of the Holocaust. "We didn't know, of course, at that point what it meant, we were hoping that he is being sent to uh, maybe a camp for elderly peo … for older people and uh, he's going to be treated according to his age" (Ehrmann and Bolkosky, 1983). Readers' evaluations of the the certainty with which the witness spoke are arrayed to the right of the sentences, indicating the number of ratings per category. This statement, in terms of its syntactic and semantic style, is typical of witness testimony. It indicates the complex stances adopted by witnesses and demonstrates some of the challenges this material presents to readers and computational approaches for the assessment of veridicality. Ehrmann speaks both of himself and of a collective "we." He hedges, self-corrects, and speaks of beliefs while communicating the underlying conflict of an evolving, difficult, contemporary understanding and a collective past hope. Readers responded to this complexity – this challenge – by evaluating the sentence as belonging to any of six different veridicality categories. The approach taken by this study suggests that the collective understanding of this sentence best describes it as communicating an *Uncertain Unknown.*

Table 9.4. *Recategorized "lost" sentences*

Sentence ID	Sentence	Cluster No.	Cluster Name	Certain–	Certain+	Possible–	Possible+	Certain but underspecified	Probable–	Probable+	Uncertain	NA/Error
704	We didn't know, of course, at that point what it meant, we were hoping that he is being sent to uh, maybe a camp for elderly peo. . . for older people and uh, he's going to be treated according to his age.	1	Uncertain Unknown	1	1	1	4			2	1	
2306	The medics who drew the blood, there were the two, as far as I know clearly.	5	Likely Fact		5		2			3		
809	But you know that uh, I'm sure you know that it has come into somebody's mind, you know, we have Neo-Nazi organizations.	7	Possible Fact		1	1	1	1	1	5	2	

The approach taken by this study provides for categorization of 2,317 items out of 2,490 (93%), versus a simple majority interrater reliability (IRR) approach that only categorized 1,497 out of 2,490 (60.1%). Additionally, this study's approach recognizes that the variation in rater assessments is actually indicative of their responses to an item's own uncertainty and the psychological difficulty posed by these event sentences, rather than a failure to accurately classify those items.

https://doi.org/10.1017/9781108854221.012 Published online by Cambridge University Press

9.6 Conclusion

The ultimate goal of this work is to better understand a fundamental aspect of narration – a speaker's belief in their own statements. That belief, for a reader, is most obviously encoded in the use of hedges, modals, and attitude markers. Though one goal for work like this would be a tool that can assign a classification with regards to interpreted speaker veridicality of events in the domain of witness testimony, an equally significant finding has to do with the distribution of results. Noting that the preponderance of results tended toward descriptions of events, rather than of nonevents, and that annotators showed far greater agreement in examples seen to be more certain suggests something about how people perceive testimonial writing; namely, that we are possibly more attuned to reading about events, rather than about negations of events, and that readers may be more precise in their agreement about more certain events and nonevents than they are about more uncertain events and nonevents. In developing training data to support that goal, it was discovered that simple IRR did not adequately describe how raters were evaluating contextual examples.

To better capture the relative certainty and facticity implied by those ratings, an alternative to simple majority agreement was implemented. Instead, the means of all valid ratings were calculated and then plotted where the x-axis denoted a continuum from counterfact, or event-negation, to fact, or event, and the y-axis denoted certainty. Those plotted results were clustered using a number of clusters determined by the elbow method, wherein the number of clusters is increased until an elbow appears in the graph. Based on these clusters, a new nine-element schema for the evaluation of veridicality that better reflects how readers interpreted real-world testiomny was proposed. That schema draws upon the study's evidence and suggests that (1) readers and witnesses have a greater sensitivity to gradations of fact than they do to gradations of counterfacts and (2) readers and witnesses exhibit more variation and less surety in their evaluation of uncertain statements. What these two aspects of reader cognition mean for the writing of history has yet to be explored.

Acknowledgments This research was supported in part by the Minerva Research Initiative, U.S. Department of Defense. The author thanks the reviewers for their careful, detailed, knowledgeable, and extremely helpful feedback and Kristopher Kyle, Shakthidhar Gopavaram, and Jennifer Olive, the graduate research assistants who helped collect the initial survey data. Lastly, I want to thank my wife, Elora; she is my most supportive and critical reviewer.

References

Caruth, Cathy. 2016. *Unclaimed Experience: Trauma, Narrative, and History*. Johns Hopkins University Press.

De Marneffe, Marie-Catherine, Manning, Christopher D., and Potts, Christopher. 2012. Did It Happen? The Pragmatic Complexity of Veridicality Assessment. *Computational Linguistics*, **38**(2), 301–333.

Divitaa, Guy, Brignonea, Emily, Carter, Marjorie E., et al. 2018. Extracting Sexual Trauma Mentions from Electronic Medical Notes Using Natural Language Processing. Pages 351–355 of: Gundlapalli, A. V., Jaulent, M.-C., and Zhao, D. (eds.), *MEDINFO 2017: Precision Healthcare through Informatics: Proceedings of the 16th World Congress on Medical and Health Informatics*, Vol. 245. Amsterdam: IOS Press.

Ehrmann, Alexander, and Bolkosky, Sidney M. 1983. *Alexander Ehrmann, First Impressions of Auschwitz*. University of Michigan–Dearborn [Television Services].

Gibbons, John Peter. 2014. *Language and the Law*. Routledge.

Hacquard, Valentine. 2011. *Modality. Semantics: An International Handbook of Natural Language Meaning*, Maienborn, Claudia, von Heusinger, Klaus, and Portner, Paul (eds.).

Hahn, Udo, and Engelmann, Christine. 2014. Grounding Epistemic Modality in Speakers' Judgments. Pages 654–667 of: Pham, Duc-Nghia, and Park, Seong-Bae (eds.), *Pacific Rim International Conference on Artificial Intelligence*. Springer.

Herman, Judith L. 2015. *Trauma and Recovery: The Aftermath of Violence From Domestic Abuse to Political Terror*. London: Hachette.

Horn, Laurence R. 1989. *A Natural History of Negation*. University of Chicago Press.

Hyland, Ken. 2005. Stance and Engagement: A Model of Interaction in Academic Discourse. *Discourse Studies*, **7**(2), 173–192.

Kodinariya, Trupti M., and Makwana, Prashant R. 2013. Review on Determining Number of Cluster in K-Means Clustering. *International Journal*, **1**(6), 90–95.

Lee, Kenton, Artzi, Yoav, Choi, Yejin, and Zettlemoyer, Luke. 2015. Event Detection and Factuality Assessment with Non-expert Supervision. Pages 1643–1648 of: Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Marsh, Herbert W. 1986. Negative Item Bias in Ratings Scales for Preadolescent Children: A Cognitive-Developmental Phenomenon. *Developmental Psychology*, **22**(1), 37–49.

Martin, Nora. 2017. Journalism, the Pressures of Verification and Notions of Post-truth in Civil Society. *Cosmopolitan Civil Societies: An Interdisciplinary Journal*, **9**(2), 41–56.

Miller, Ben, Shrestha, Ayush, Derby, Jason, et al. 2013. Digging into Human Rights Violations: Data Modelling and Collective Memory. Pages 37–45 of: Lin, Tsau Young, Raghavan, Vijay, and Wah, Benjamin (eds.), *2013 IEEE International Conference on Big Data*. IEEE.

Miller, Ben, Olive, Jennifer, Gopavaram, Shakthidhar, and Shrestha, Ayush. 2015. Cross-document Non-fiction Narrative Alignment. Pages 56–61 of: Caselli,

Tommaso, van Erp, Marieke, Minard, Anne-Lyse, et al. (eds.), *Proceedings of the First Workshop on Computing News Storylines*.

Nugumanova, Aliya, and Bessmertny, Igor. 2013. Applying the Latent Semantic Analysis to the Issue of Automatic Extraction of Collocations from the Domain Texts. Pages 92–101 of: Klinov, Pavel, and Mouromtsev, Dmitry (eds.), *International Conference on Knowledge Engineering and the Semantic Web*. Springer.

Saurí, Roser, and Pustejovsky, James. 2009. FactBank: A Corpus Annotated with Event Factuality. *Language Resources and Evaluation*, **43**(3), 227–268.

Saurí, Roser, and Pustejovsky, James. 2012. Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text. *Computational Linguistics*, **38**(2), 261–299.

Saurí, Roser, Knippen, Robert, Verhagen, Marc, and Pustejovsky, James. 2005. Evita: A Robust Event Recognizer for QA Systems. Pages 700–707 of: Mooney, Raymond J. (ed.), *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

The Sept 11 Records. 2005. *The New York Times*, November 30. https://archive.nytimes.com/www.nytimes.com/indexes/2005/11/30/nyregion/nyregionspecial3/index.html

Soni, Sandeep, Mitra, Tanushree, Gilbert, Eric, and Eisenstein, Jacob. 2014. Modeling Factuality Judgments in Social Media Text. Pages 415–420 of: Toutanova, Kristina, and Wu, Hua (eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Volume 2. Short Papers*.

The South African Broadcasting Corporation. 1996 (Jun). Human Rights Violation Hearings, Case Number CT/00530.

Stanovsky, Gabriel, Eckle-Kohler, Judith, Puzikov, Yevgeniy, Dagan, Ido, and Gurevych, Iryna. 2017. Integrating Deep Linguistic Features in Factuality Prediction over Unified Datasets. Pages 352–357 of: Barzilay, Regina, and Kan, Min-Yen (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Volume 2. Short Papers*.

Truong, Khiet P., Westerhof, Gerben J., Lamers, Sanne M.A., and de Jong, Franciska. 2014. Towards Modeling Expressed Emotions in Oral History Interviews: Using Verbal and Nonverbal Signals to Track Personal Narratives. *Literary and Linguistic Computing*, **29**(4), 621–636.

Wan, Xiaojun, and Zhang, Jianmin. 2014. CTSUM: Extracting More Certain Summaries for News Articles. Pages 787–796 of: Geva, Shlomo, Trotman, Andrew, Bruza, Peter, Clarke, Charles L. A., and Järvelin, Kal (eds.), *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*.