“Reading Certainty: Evidence from a Large Study on NLP and Witness Testimony”

Ben Miller
b.j.miller@emory.edu

Witness testimony provides the first draft of history, and requires a kind of reading that connects descriptions of events from many perspectives and sources. "Reading Certainty" examines one critical step in that process, namely how a group of approximately 230 readers decided whether a statement about an event is credible and factual. That examination supports an exploration of how readers of primary evidence think about factual and counterfactual statements, and how they interpret the certainty with which a witness makes their statements. This presentation argues that readers of collections of witness testimony were more likely to agree about event descriptions when those providing the description are certain, and that the ability of readers to accept gradations of certainty were better when a witness described factual, rather than counter-factual events. These findings lead to a suggestion for how researchers in linguistics and the humanities could better model the question of speaker certainty, at least when dealing with the kind of narrative non-fiction one finds in witness testimony.

As represented by the effort that has gone in to various truth and reconciliation processes, such as occurred in Canada from 2008-2015 about treatment of First Nations in the residential schools systems, or in South Africa that from 1996 to 1998 convened hearings on abuses perpetrated during the Apartheid period, these materials serve vitally important functions for people and communities. Thus, despite their challenges, they need to be included in the types of material made more readable by computational methods. Absent any quantitative approaches, these many tens of thousands of witness statements remain mostly unread, perceived only in the aggregate form of commissions’ reports.

Various challenges make the understanding of event language in real-world documents a meaningful, but difficult research problem. Often, witnesses indicate space, time, and entities referentially more so than absolutely, so their stories are resistant to techniques like named entity or temporal recognition. Additionally, these types of stories are frequently feature fragmented syntactic structure and highly referential semantics. Also, it is common that a witness either does not know where or when something specifically happened, they elide that detail, or they do not have the language with which to talk about it. In testimony provided by a first responder to the World Trade Center attacks of September 11, 2001 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" (Times, 2004). The resistance engendered by ambiguity hinders the critical work of associative reading, where descriptions of events by one witness are correlated to descriptions of events by additional witnesses.

While the kind of associative reading historians, linguists, and those studying narrative non-fiction could be supported by computational approaches, a critical framework to facilitate interpretation of event-related language is essential. One critical predicate for understanding the events that comprise these stories is the extent to which the speaker is certain about the statement they put forward. Any question 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. This measure of speaker certainty is also known as veridicality. A second critical predicate in relation to this material's role as an anchor of collective memory is the speaker's statement's facticity. Combined, these 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.

To explore this idea, a large study was conducted that asked; how are certainty and uncertainty indicated in the language of witness statements, and how do readers interpret those statements. This research builds on the work of (Hyland, 2005); (De Marneffe et al., 2012); (Saurí and Pustejovsky, 2009, 2012); (Lee et al., 2015); and (Stanovsky et al., 2017) in the area of using NLP to quantify a speaker's certainty about their statements. Following the above examples, 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. While 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 wouldn't reflect how witnesses use language, or how readers grapple with the complex problem of understanding witness testimony. A random sample of sentences containing about 800 events was taken from each of the seven corpora leading to a total of approximately 4,000 events. A total of 227 unique raters participated in the study providing a total of 27,800 individual event ratings. Those ratings were categorical, based upon the existing best model for assessing certainty according to the literature referenced above. However, a majority of raters were only likely to agree on the category when the event was considered to be Certain Fact, or Probable Fact. Table 1 below indicates the initial categories and when at least half of the raters of an item agreed.

Table 1: Simple Majority Interrater Agreement

Certain +	1083
Probable +	370
Possible +	21
Certain -	18
Probable -	5
Possible -	0
Certain by under-specified	0
Uncertain	0
Error	0

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 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, 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, shown in Figure 1, 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, as shown in Table 2, I propose a new nine-element schema for the evaluation of veridicality. This new schema better reflects that better reflects how readers interpreted real-world testimony and the evidence from this study, and potentially better describes how readers process information about certainty and events in witness testimony.

Table 2: K-means Clustering of Survey Results

Cluster	Within SS	Variance	SD	Nomenclature	Certainty	Counter/Fact	No. of Items
6	4596.2	71.8	8.5	Certain Fact	27.6	26.6	483
3	7625.9	119.2	10.9	Probable Fact	22.9	20.2	581
5	5062.1	79.1	8.9	Likely Fact	16.6	15.6	451
7	4798.2	75.0	8.7	Possible Fact	10.9	10.0	336
8	3843.4	60.1	7.7	Certain Unknown	21.9	5.1	117
2	4106.8	64.2	8.0	NA / Other	2.6	4.6	173
1	3846.9	60.1	7.8	Uncertain Unknown	-9.7	-0.9	64
9	5068.0	79.2	8.9	Possible Counterfact	9.5	-7.9	113
4	8555.1	133.7	11.6	Certain Counterfact	23.4	-16.5	172

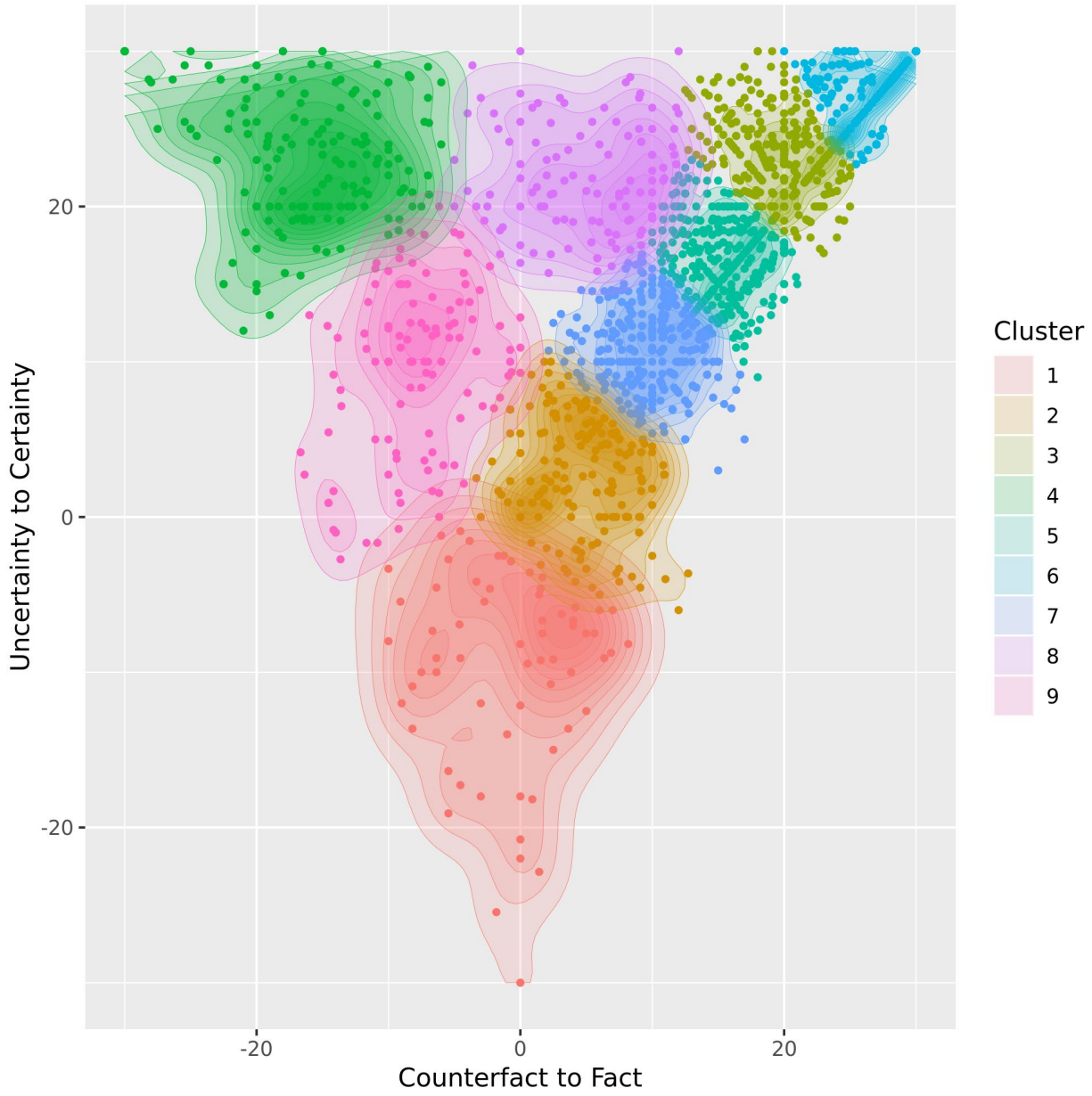

Figure 1 Clusters of Certainty to Uncertainty and Factual to Counterfactual Statements

With this theoretically and empirically developed instrument, linguists, historians, and readers of narrative non-fiction such as witness statements, may be better able to move forward with methods for computational approaches to associative reading, and better able to reengage with the primary evidence of our societies truth and reconciliation processes.

Works Cited

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.

Hyland, Ken. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse studies*, 7(2), 173-192.

Lee, Kenton, Artzi, Yoav, Choi, Yejin, and Zettlemoyer, Luke. 2015. Event detection and factuality assessment with non-expert supervision. Pages 1643-1648 of: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Saurí, Roser, and Pustejovsky, James. 2009. FactBank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3), 227.

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.

Stanovsky, Gabriel, Eckle-Kohler, Judith, Puzikov, Yevgeniy, Dagan, Ido, and Gurevych, Iryna. 2017. Integrating deep linguistic features in factuality prediction over uni_ed datasets. Pages 352-357 of: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers).

Times, The New York. 2004. The Sept 11 Records. *The New York Times*, Aug.