Pace Law Review

Volume 32
Issue 2 *Spring 2012*

Article 1

April 2012

Inconsistent Responsiveness Determination in Document Review: Difference of Opinion or Human Error?

Maura R. Grossman

Gordon V. Cormack
David R. Cheriton School of Computer Science, University of Waterloo

Inconsistent Responsiveness Determination in Document Review: Difference of Opinion or Human Error?

Maura R. Grossman* and Gordon V. Cormack**

Abstract

This Article analyzes the inconsistency between different document review efforts on the same document collection to determine whether that inconsistency is due primarily to ambiguity in applying the definition of responsiveness to particular documents, or due primarily to human error. By examining documents from the TREC 2009 Legal Track, the Authors show that inconsistent assessments regarding the same documents are due in large part to human error. Therefore, the quality of a review effort is not simply a matter of opinion; it is possible to show objectively that some reviews, and some review methods, are better than others.

I. Introduction

In responding to a request for production in civil litigation, the goal is typically to produce, as nearly as practicable, *all* and *only* the non-privileged documents that are *responsive* to the request.[1]

It has been observed that independent reviewers, when asked to identify all and only the responsive documents in a large collection, will not identify precisely the same set of documents.[2] It has been suggested

---

* Maura R. Grossman is counsel at Wachtell, Lipton, Rosen & Katz. She is co-coordinator of the Legal Track of the National Institute of Standards and Technology's Text Retrieval Conference ("TREC"), an adjunct faculty member at Columbia Law School, and a member of the Steering Committee of The Sedona Conference® Working Group on Electronic Document Retention and Production. The views expressed herein are solely those of the Author and should not be attributed to her firm or its clients.

** Gordon V. Cormack is a professor in the David R. Cheriton School of Computer Science at the University of Waterloo. He is a member of the Program Committee for TREC, co-coordinator of the TREC Legal Track, and past coordinator of the TREC Spam Track.

1. *See* FED. R. CIV. P. 26(b), (g); FED. R. CIV. P. 34(a); FED. R. CIV. P. 37(a)(4).
2. Peter Bailey et al., *Relevance Assessment: Are Judges Exchangeable and Does it*

that the observed inconsistency between reviewers demonstrates that responsiveness is a matter of subjective opinion rather than fact, and therefore, there can be no *gold standard* against which the effectiveness of search and review efforts may be measured.[3] This Article presents an alternate hypothesis: that inconsistency among reviewers is equally well explained by human error and does not preclude the existence of a gold standard of responsiveness against which review efforts may be evaluated. The alternative hypothesis is supported by two experiments:

> 1. *The Tall T's Game*, a simple, well-defined task for which human results exhibit the same type of inconsistency as for document review; and
>
> 2. *Re-examination of the TREC 2009 adjudication results*, a post-hoc, qualitative analysis of a random sample of cases of disagreement identified during the process of constructing the gold standard for the TREC 2009 Legal Track Interactive Task ("TREC 2009").[4]

The Tall T's Game, while obviously not a document review task, illustrates that human judgments may show substantial inconsistency, even when there is an objectively verifiable correct answer. In other

---

Matter?, 31 PROC. ANN. INT'L ACM SIGIR CONF. ON RES. & DEV. INFO. RETRIEVAL 667 (2008); THOMAS I. BARNETT & SVETLANA GODJEVAC, FASTER, BETTER, CHEAPER LEGAL DOCUMENT REVIEW, PIPE DREAM OR REALITY? (2011), *available at* http://www.umiacs.umd.edu/~oard/desi4/papers/barnett3.pdf; Heting Chu, *Factors Affecting Relevance Judgment: A Report from TREC Legal Track*, 67 J. DOCUMENTATION 264 (2011); Efthimis N. Efthimiadis & Mary A. Hotchkiss, *Legal Discovery: Does Domain Expertise Matter?*, 45 PROC. AM. SOC'Y FOR INFO. SCI. & TECH. 1 (2008); Herbert L. Roitblat et al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y FOR INFO. SCI. & TECH. 70 (2010); Ellen M. Voorhees, *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, 36 INFO. PROCESSING & MGMT. 697 (2000); Jianqiang Wang & Dagobert Soergel, *A User Study of Relevance Judgments for E-Discovery*, 47 PROC. AM. SOC'Y FOR INFO. SCI. & TECH. 1 (2010).

3. BARNETT & GODJEVAC, *supra* note 2, at 2, 12; ELI NELSON, THE FALSE DICHOTOMY OF RELEVANCE: THE DIFFICULTY OF EVALUATING THE ACCURACY OF DISCOVERY REVIEW METHODS USING BINARY NOTIONS OF RELEVANCE (2011), http://www.umiacs.umd.edu/~oard/desi4/papers/nelson.pdf; Ralph C. Losey, *Secrets of Search—Part One*, E-DISCOVERY TEAM (Dec. 11, 2011, 9:23 PM), http://e-discoveryteam.com/2011/12/11/secrets-of-search-part-one.

4. Bruce Hedin et al., Overview of the TREC 2009 Legal Track, *in* THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE PROCEEDINGS (E. M. Voorhees & Lori P. Buckland eds. 2010), *available at* http://trec.nist.gov/pubs/trec18/papers/LEGAL09.OVERVIEW.pdf.

words, the observed inconsistency does not necessarily indicate that the correct answer is a matter of subjective opinion, or that there can be no absolute standard against which to measure human prowess at identifying taller T's.

In re-examining the TREC 2009 adjudication results, the Authors examined a random sample of documents for which the first-pass reviewer's responsiveness determination was reversed by the TREC "Topic Authority"—a senior lawyer familiar with the subject matter—and made their own determination as to whether the document was "clearly responsive," "clearly non-responsive," or "arguable," meaning that it could reasonably be construed as either responsive or not, given the production request and the applicable coding guidelines. More than 90 percent of the time, the Authors' determination was that the document was "clearly responsive" or "clearly non-responsive," meaning that one of the two reviewers was right and the other was wrong. Less than 10 percent of the time was the Authors' determination of the document "arguable," meaning that the disagreement could be due to a reasonable difference of opinion as to responsiveness. Overall, the results suggest that inconsistent assessments of responsiveness may be largely attributed to human error, and that it is reasonable to derive a gold standard for responsiveness.

II. The Tall T's Game

Figure 1 depicts a simple game that illustrates the issue of reviewer inconsistency. The object of the game is to identify the T's that are taller than they are wide. Eleven volunteers—well-known lawyers or judges in the e-discovery realm, as well as a professor published in the area of e-discovery—were asked to identify the taller T's without using a ruler or any other measuring instrument. As shown in Figure 2, the eleven participants identified nine entirely different combinations of the twenty-five T's. The only two pairs of players to agree on results were A and J, who both identified only the one T at position E3 to be taller than it was wide, and C and F, who identified none of the T's as taller than they were wide.

Figure 3 indicates the pairwise *agreement* among the eleven participants. The agreement between two players is defined as the fraction of all examples (T's, in this instance) as to which they agree. For example, D agreed with I that ten particular T's were taller, and that nine particular T's were not taller. That is, they agreed on a total of nineteen of the twenty-five T's. Their agreement is therefore nineteen out of

twenty-five, or 76 percent. It is difficult to glean from Figures 2 and 3, alone, who is right and who is wrong, and the reader may therefore be tempted to conclude that the answer is a matter of opinion, or "too close to call," for many T's.

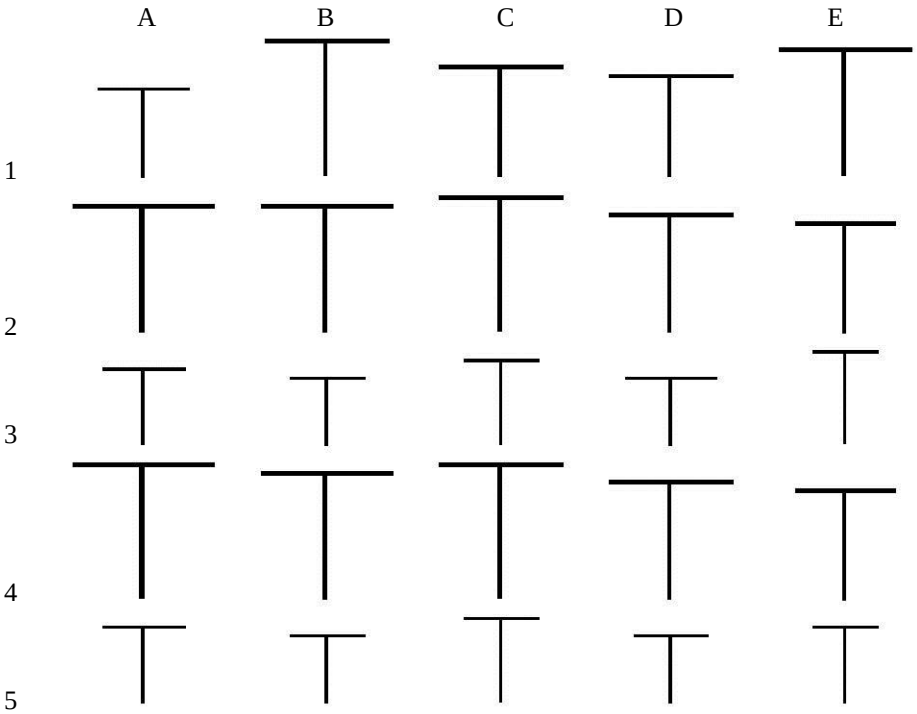


Figure 1: Instructions for The Tall T's Game: Among the twenty-five "T" figures shown above, identify all those that are taller than they are wide. Do not use a ruler or any other measuring device for this purpose.

If the height and width of some of the T's were equal, or so nearly equal that it was impossible to measure the difference using a ruler, the reader might correctly declare some of them to be "too close to call." However, that is not the case here; in this game, the height of each T differs from its width by more than 5 percent, and determining whether or not each T is taller than it is wide is well within human capability—if not by eye alone—most certainly using a ruler.[5] Accordingly, the

5. The answer key for The Tall T's Game is provided in Figure 13 *infra*.

inconsistency among the various players does not suggest that the T's were "too close to call," as indeed they are not.

	A	B	C	D	E
1		BDEGHIK	EI		BDEI
2	E	EI	BDEIK	EI	DEI
3	EI	E	BDEGHIK		ABDEGHIJK
4	BDEGI	EI	BDEGIK	E	BEHIK
5	E	E	BDEGHIK	E	BDEGHIK

Figure 2: T's identified as taller by eleven e-discovery luminaries known to the Authors, labeled arbitrarily as A through K. Note that C and F, who identified none of the T's as taller than they were wide, do not appear in the table.

B	64%									
C	96%	60%								
D	64%	92%	60%							
E	16%	52%	12%	52%						
F	96%	60%	100%	60%	12%					
G	76%	88%	72%	88%	40%	72%				
H	80%	84%	76%	76%	36%	76%	88%			
I	40%	76%	36%	76%	76%	36%	64%	60%		
J	100%	64%	96%	64%	16%	96%	76%	80%	40%	
K	72%	92%	68%	84%	44%	68%	88%	92%	68%	72%
	A	B	C	D	E	F	G	H	I	J

Figure 3: Pairwise agreement (expressed as a percent) among the eleven participants in The Tall T's Game.

Once the tallness of each T is determined, one may score each of the participants by various effectiveness measures. Figure 4 shows *accuracy*, the fraction of all T's that are classified correctly, regardless of whether they are taller or not. The players' scores differ substantially from each other, and those differences are real, not a matter of subjective opinion. Player K—in this particular game—has twice the accuracy of player E. This fact is not apparent from the pairwise agreement scores shown in Figure 3.

The Authors do not mean to suggest that a human's ability to recognize taller T's without a ruler is representative of their ability to

recognize responsive documents. This experiment merely illustrates that human judgment can yield remarkably inconsistent results, even when the correct answer is well defined. Furthermore, when the results are compared to the answer key, it becomes apparent that some human results are considerably better than others.

Player	K	H	D	B	G	I	J	A	F	C	E
Accuracy	96%	88%	88%	88%	84%	72%	68%	68%	64%	64%	48%

Figure 4: Individual accuracy scores (expressed as a percent) for each player in The Tall T's Game.

III. The Document Review Game

The structure of The Document Review Game closely parallels that of The Tall T's Game. Instead of determining which Ts are taller, and which are not, the player (a document reviewer) must determine which documents are responsive to a request for production, and which are not. Previous studies have claimed that there can be no gold standard based on agreement rates between independent document reviewers that are remarkably similar to those we report here for the Tall T's Game.[6]

Figures 5 and 6 show the pairwise agreement of reviewers in these two studies of document review efforts, which may be compared to those of the Tall T's players in Figure 3. Just as we observed with the results of The Tall T's Game, one simply cannot infer from the agreement rates whether responsiveness is a matter of subjective opinion or, as with the tallness of T's, a matter of fact.

To determine the tallness of T's as a matter of fact, one can measure the height and width of the T's with a ruler. To determine the responsiveness of documents, the TREC 2009 Legal Track used a "Topic Authority"—a senior lawyer familiar with the subject matter of the request for production—to prepare formal coding guidelines specifying how the responsiveness of a document was to be assessed, and also to render the final responsiveness determination in cases of disagreement.[7] After each participating TREC 2009 team completed The Document Review Game, submitting to TREC its list of all documents deemed to be responsive to a particular request for production (a "topic" in TREC parlance), TREC used a team of human reviewers to code a sample of the

6. BARNETT & GODJEVAC, *supra* note 2, at 8; Roitblat et al., *supra* note 2, at 74.
7. Hedin et al., *supra* note 4, at 2, 3.

documents as responsive or not, according to the coding guidelines.[8] Participating TREC teams were given the results of this first-pass review and invited to appeal any coding decision with which they disagreed.[9] The Topic Authority adjudicated all documents whose first-pass coding decision was appealed, and issued a final authoritative determination as to responsiveness.[10] These final determinations, along with any first-pass codes that were not appealed, were used as the gold standard (i.e., the answer key) against which the participating teams' submissions were then evaluated.[11]

B	75.06%					
C	83.05%	75.01%				
D	74.51%	65.53%	72.20%			
E	79.91%	71.95%	76.69%	80.32%		
F	76.94%	84.90%	75.21%	68.17%	74.26%	
G	76.94%	75.23%	74.11%	67.39%	73.08%	77.20%
	A	B	C	D	E	F

Figure 5: Pairwise agreement (expressed as a percent) among seven separate reviews for responsiveness in a study conducted by Barnett & Godjevac.

B	70.2%	
O	75.5%	72.0%
	A	B

Figure 6: Pairwise agreement (expressed as a percent) among three reviews for responsiveness in a study conducted by Roitblat et al.

The TREC 2009 Legal Track involved seven different document review games, using seven different topics, each of which was a request for production in a mock civil proceeding.[12] The requests for production are shown in Figure 7; the coding guidelines are available online.[13]

8. *Id.* at 3, 7-8.
9. *Id.* at 3-4, 13.
10. *Id.*
11. *Id.*
12. *Id.* at 5-6.
13. *See Topic-Specific Guidelines—Topic 201*, TEXT RETRIEVAL CONFERENCE

The appeal and adjudication process resulted in a substantial number of the first-pass reviewers' assessments being reversed; that is, the Topic Authority (and also, presumably, the appealing team) often disagreed with the first-pass reviewer as to responsiveness. Figure 8 shows the number of documents coded responsive and non-responsive by the first-pass reviewer, and the number of coding decisions reversed by the Topic Authority, for each of the seven topics used at TREC 2009. Over all topics, the average agreement for documents coded responsive by the first-pass reviewer was 71.2 percent, while the average agreement for documents coded non-responsive by the first-pass reviewer was 97.4 percent.

Topic	Production Request
201	All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in structured commodity transactions known as "prepay transactions."
202	All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in transactions that the Company characterized as compliant with FAS 140 (or its predecessor FAS 125).
203	All documents or communications that describe, discuss, refer to, report on, or relate to whether the Company had met or could, would, or might meet its financial forecasts, models, projections, or plans at any time after January 1, 1999.
204	All documents or communications that describe, discuss, refer to, report on, or relate to any intentions, plans, efforts, or

(TREC) 2009 LEGAL TRACK (Oct. 31, 2009), http://plg1.cs.uwaterloo.ca/trec-assess/TopicGuidelines_201_.pdf;
Topic-Specific Guidelines—Topic 202, TEXT RETRIEVAL CONFERENCE (TREC) 2009 LEGAL TRACK (Nov. 2, 2009), http://plg1.cs.uwaterloo.ca/trec-assess/TopicGuidelines_202_.pdf;
Topic-Specific Guidelines—Topic 203, TEXT RETRIEVAL CONFERENCE (TREC) 2009 LEGAL TRACK (Nov. 2, 2009), http://plg1.cs.uwaterloo.ca/trec-assess/TopicGuidelines_203_.pdf; *Topic-Specific Guidelines—Topic 204*, TEXT RETRIEVAL CONFERENCE (TREC) 2009 LEGAL TRACK (Oct. 22, 2009), http://plg1.cs.uwaterloo.ca/trec-assess/TopicGuidelines_204.pdf; *Topic-Specific Guidelines—Topic 205*, TEXT RETRIEVAL CONFERENCE (TREC) 2009 LEGAL TRACK (Dec. 14, 2009), http://plg1.cs.uwaterloo.ca/trec-assess/TopicGuidelines_205_.pdf; *Topic-Specific Guidelines—Topic 206*, TEXT RETRIEVAL CONFERENCE (TREC) 2009 LEGAL TRACK (Nov. 2, 2009), http://plg1.cs.uwaterloo.ca/trec-assess/TopicGuidelines_206_.pdf; *Topic-Specific Guidelines—Topic 207*, TEXT RETRIEVAL CONFERENCE (TREC) 2009 LEGAL TRACK (Oct. 22, 2009), http://plg1.cs.uwaterloo.ca/trec-assess/TopicGuidelines_207_.pdf.

Topic	Production Request
	activities involving the alteration, destruction, retention, lack of retention, deletion, or shredding of documents or other evidence, whether in hard copy or electronic form.
205	All documents or communications that describe, discuss, refer to, report on, or relate to energy schedules and bids, including but not limited to, estimates, forecasts, descriptions, characterizations, analyses, evaluations, projections, plans, and reports on the volume(s) or geographic location(s) of energy loads.
206	All documents or communications that describe, discuss, refer to, report on, or relate to any discussion(s), communication(s). or contact(s) with financial analyst(s), or with the firm(s) that employ them, regarding (i) the Company's financial condition, (ii) analysts' coverage of the Company and/or its financial condition, (iii) analysts' rating of the Company's stock, or (iv) the impact of an analyst's coverage of the Company on the business relationship between the Company and the firm that employs the analyst.
207	All documents or communications that describe, discuss. refer to, report on, or relate to fantasy football, gambling on football and related activities, including but not limited to, football teams, football players, football games, football statistics, and football performance.

Figure 7: Mock production requests ("topics") composed for the TREC 2009 Legal Track Interactive Task.

Topic	First-Pass Assessment	# Documents	# Overturned	% Overturned	First-Pass/TA Agreement
201	Responsive	603	363	60.2%	39.8%
201	Non-Responsive	5,605	101	1.8%	98.2%
202	Responsive	1,743	115	6.6%	93.4%
202	Non-Responsive	5,462	469	8.6%	91.4%
203	Responsive	131	69	53.7%	47.3%
203	Non-Responsive	5,296	186	3.6%	96.4%
204	Responsive	105	50	47.6%	52.4%
204	Non-Responsive	7,024	169	2.4%	97.6%
205	Responsive	1,631	882	54.1%	45.9%
205	Non-Responsive	4,289	50	1.2%	98.8%
206	Responsive	235	50	21.3%	78.7%
206	Non-Responsive	6,860	0	0.0%	100.0%
207	Responsive	938	23	2.5%	97.5%
207	Non-Responsive	7,377	125	1.7%	98.3%
All	Responsive	5,386	1,552	28.8%	71.2%
All	Non-Responsive	41,913	1,100	2.6%	97.4%

Figure 8: Number of documents appealed and the

> success rates of appeals for TREC 2009 (expressed both as an absolute number and as a percentage), categorized by topic and first-pass assessment (responsive or non-responsive).

A key question that naturally arises is whether the inconsistency in coding determinations between the first-pass reviewer and the Topic Authority is a matter of (i) reasonable differences in opinion, (ii) an error by the first-pass assessor, or (iii) an error by the Topic Authority. The Authors set out to resolve this question by examining these documents, and others, and coding them each as "*clearly responsive*," "*clearly non-responsive*," or "*arguable*." If the documents about which the first-pass reviewer and the Topic Authority disagree are arguable, one may consider either determination to be valid; that is, the inconsistency reflects a reasonable difference of opinion. If, on the other hand, the documents are clearly responsive or clearly non-responsive, the inconsistency reflects an error on the part of either the first-pass reviewer or the Topic Authority.

The validity of any evaluation process hinges on the answer to this question. If the first-pass assessments are just as good as the final adjudicated results, one can use them instead as the gold standard of relevance.[14]

IV. Evaluating "Arguability"

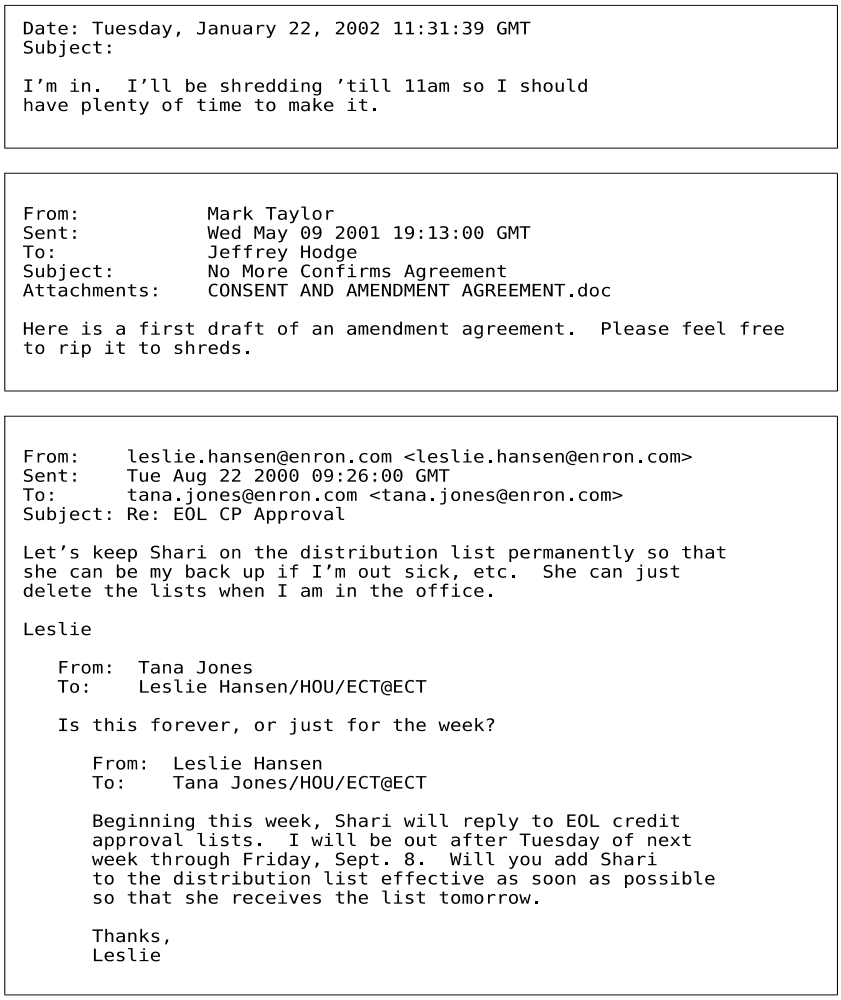
The Authors illustrate the issue of "arguability" by using six documents and two topics from TREC 2009.

Figure 9 shows three documents for which the first-pass reviewer's coding decision on responsiveness to Topic 204 (*supra* Figure 7) was reversed by the Topic Authority. In the Authors' opinion, the first document is clearly responsive; there is no reasonable doubt that it refers to document shredding, which is explicitly referenced in the request for production. In the Authors' opinion, the second document is clearly non-responsive; there is no reasonable doubt that the phrase "rip it to shreds"

14. The TREC 2009 preliminary results, which used the first-pass assessments as the gold standard, were dramatically different from the final adjudicated results. *See* Hedin et al., *supra* note 4, at 13; *see also* William Webber et al., Assessor Error in Stratified Evaluation, *in* PROCEEDINGS OF THE 19TH ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT 529 (2010), *available at* http://ww2.cs.mu.oz.au/~wew/papers/wosh10_cikm.pdf (noting a large discrepancy between the preliminary and final results).

is figurative and does not refer to document destruction. In the Authors' opinion, the third document is of arguable responsiveness. It discusses the deletion of redundant copies of "EOL credit approval lists," but it is not clear whether or not the potential deletion of redundant copies should be considered "destruction," as referenced in the request for production and the coding guidelines.

Figure 10 shows three documents for which the first-pass reviewer's coding decision for responsiveness to Topic 207 (*supra* Figure 7) was reversed by the Topic Authority. In the Authors' opinion, the first document is clearly non-responsive; the subject line makes it clear that the document pertains to *baseball*, not football, as required by the request for production. Furthermore, the guidelines explicitly state that documents that refer exclusively to sports other than football are non-responsive. In the Authors' opinion, the second document is clearly responsive; there is no reasonable doubt that it refers not only to football, but to fantasy football and to gambling on football, both of which are explicitly referenced in the request for production. In the Authors' opinion the third document is of arguable responsiveness. It contains a whimsical reference to television coverage of football. The guidelines specify that "jokes about football" are not responsive unless they refer to a specific football player, football team, or football game. Is this reference a joke about football? Does it refer to a specific football game? A reasonable argument could be constructed for either point of view.



```
Date: Tuesday, January 22, 2002 11:31:39 GMT
Subject:

I'm in.  I'll be shredding 'till 11am so I should
have plenty of time to make it.
```

```
From:           Mark Taylor
Sent:           Wed May 09 2001 19:13:00 GMT
To:             Jeffrey Hodge
Subject:        No More Confirms Agreement
Attachments:    CONSENT AND AMENDMENT AGREEMENT.doc

Here is a first draft of an amendment agreement.  Please feel free
to rip it to shreds.
```

```
From:    leslie.hansen@enron.com <leslie.hansen@enron.com>
Sent:    Tue Aug 22 2000 09:26:00 GMT
To:      tana.jones@enron.com <tana.jones@enron.com>
Subject: Re: EOL CP Approval

Let's keep Shari on the distribution list permanently so that
she can be my back up if I'm out sick, etc.  She can just
delete the lists when I am in the office.

Leslie

   From:  Tana Jones
   To:    Leslie Hansen/HOU/ECT@ECT

   Is this forever, or just for the week?

      From:  Leslie Hansen
      To:    Tana Jones/HOU/ECT@ECT

      Beginning this week, Shari will reply to EOL credit
      approval lists.  I will be out after Tuesday of next
      week through Friday, Sept. 8.  Will you add Shari
      to the distribution list effective as soon as possible
      so that she receives the list tomorrow.

      Thanks,
      Leslie
```

Figure 9: The Document Review Game: From the three documents above, identify all and only those documents concerning "the alteration, destruction, retention, lack of retention, deletion, or shredding of documents or other evidence." These documents and the request for production were taken from TREC 2009 (from top to bottom, Documents 0.7.47.1149688, 0.7.47.833163, and 0.7.6.252211, and Topic 204).

```
From:           Barry Tycholiz
Sent:           Fri Sep 14 2001 12:21:34 GMT
To:             Jessica Presas
Subject:        Baseball Tickets
Importance:     Normal
Priority:       Normal
Sensitivity:    None

Jessica, did we place the order for the playoff tickets... BT
```

```
From:   Bass, Eric
Sent:   Thursday, January 17, 2002 11:19 AM
To:     Lenhart, Matthew
Subject:        FFL Dues

You owe $80 for fantasy football.  When can you pay?
```

```
Subject:        RE: How good is Temptation Island 2

They have some cute guy lawyers this year-but I bet you
probably watch that manly Monday night Football.
```

Figure 10: The Document Review Game: From the three documents above, identify all and only those documents concerning "fantasy football, gambling on football, and related activities, including but not limited to, football teams, football players, football games, football statistics, and football performance." The documents and request for production are from TREC 2009 (from top to bottom, Documents 0.7.47.5813, 0.7.47.320807, and 0.7.6.179483, and Topic 207).

The Topic Authority was required to code each document as responsive or non-responsive. For the four documents that the Authors characterized as clearly responsive or clearly non-responsive, the Topic Authority agreed. That is, the first-pass reviewer was clearly wrong. For the two documents that the Authors characterized as arguable, the Topic Authority coded one as non-responsive (Document 0.7.6.252211 for Topic 204), and one as responsive (Document 0.7.6.179483 for Topic 207).

These six documents and two topics illustrate the crux of the problem. If, in the ultimate evaluation effort, the majority of disagreements between the first-pass review and the final gold standard are due to arguable responsiveness, there is no reasonable basis to choose either of the two as the correct answer, even if the two answer keys yield radically different results. If, on the other hand, the majority of documents about which there are disagreements are clearly responsive or clearly non-responsive (and thus, inarguable), there is a basis to choose one answer over the other as correct. That is, we can construct a valid gold standard against which to compare review efforts.

V. Experiment

The Authors' objective in this experiment was to test two competing hypotheses:

> **Hypothesis 1:** *Reviewer disagreement is largely due to ambiguity or inconsistency in applying the criteria for responsiveness to particular documents*; or
>
> **Hypothesis 2:** *Reviewer disagreement is largely due to human error*.

Hypothesis 1 and Hypothesis 2 are mutually incompatible; evidence refuting Hypothesis 1 supports Hypothesis 2, and vice versa.

To test the validity of the two hypotheses, the Authors constructed an experiment in which, prior to the experiment, the two hypotheses were used to predict the outcome. An observed result consistent with one hypothesis and inconsistent with the other would provide evidence supporting the former and refuting the latter.

In particular, Hypothesis 1 predicted that if one examined a document about whose responsiveness human reviewers disagreed, it would generally be *difficult* to determine whether or not the document was responsive; that is, it would usually be possible to construct a reasonable argument that the document was either responsive or non-responsive (i.e., arguable). On the other hand, Hypothesis 2 predicted that it would generally be *clear* whether or not the document was responsive; it would usually be possible to construct a reasonable argument that the document was responsive, or that the document was non-responsive, but not both (i.e., inarguable).

At the outset, the Authors conjectured that the results of the experiment would more likely support Hypothesis 1.

VI. TREC 2009 Adjudicated Assessments

TREC 2009 used a two-pass adjudicated review process to construct the gold standard.[15] In the first pass, law students or professional contract attorneys reviewed a stratified random sample of documents for each of seven production requests (topics), coding each document as responsive or not.[16] TREC 2009 participating teams were invited to appeal any of the first-pass reviewer coding decisions with which they disagreed, and the Topic Authority was asked to make a final determination as to whether the appealed document was responsive or not.[17] The gold standard considered a document to be *responsive* if the first-pass reviewer coded it as responsive and that decision was not appealed, if the first-pass reviewer coded it as responsive and that decision was upheld by the Topic Authority, or if the first-pass reviewer coded it as non-responsive and that decision was overturned by the Topic Authority.[18] The gold standard considered a document to be *non-responsive* if the first-pass reviewer coded it as non-responsive and that decision was not appealed, if the first-pass reviewer coded it as non-responsive and that decision was upheld by the Topic Authority, or if the first-pass reviewer coded it as responsive and the decision was overturned by the Topic Authority.[19]

A gold standard was created for each of the seven topics.[20] A total of 49,285 documents—about seven thousand per topic—were assessed during the first-pass review. A total of 2,976 documents (5 percent) were appealed and therefore adjudicated by the Topic Authority. Of those appeals, 2,652 (89 percent) were successful; that is, the Topic Authority *disagreed* with the first-pass reviewer 89 percent of the time. A breakdown of the number of documents appealed per topic, and the outcome of those appeals, is provided in Figure 8.

VII. Post-Hoc Assessment

The Authors performed a qualitative, post-hoc assessment on a sample of the successfully appealed documents from each category

15. Hedin et al., *supra* note 4, at 4-5.

16. *See id.* at 8.

17. *See id.* at 3.

18. *See id.* at 2-3.

19. *See id.*

20. TREC 2009 LEGAL TRACK, http://trec.nist.gov/data/legal09.html (last updated Feb. 23, 2011) (linking to the gold standard and evaluation tools).

represented in Figure 8; that is, documents where the TREC 2009 first-pass reviewer and Topic Authority disagreed. Where fifty or more documents were successfully appealed, the Authors selected a random sample of fifty. Where fewer than fifty documents were successfully appealed, the Authors selected all of the appealed documents.

The Authors used the plain-text version of the TREC 2009 document corpus, downloaded by one of the Authors while participating in TREC 2009,[21] and redistributed for use at TREC 2010.[22] For each topic, one of the Authors of this study examined every document, in every sample, and coded each one as "responsive," "non-responsive," or "arguable," based on the content of the document, the production request, and the written coding guidelines prepared for TREC 2009 by each Topic Authority. The Authors coded a document as "responsive" if they believed there was no reasonable argument that the document fell outside the definition of responsiveness dictated by the production request and coding guidelines. Similarly, the Authors coded a document as "non-responsive" if they believed there was no reasonable argument that the document should have been identified as responsive to the production request. Finally, the Authors coded the document as "arguable" if they believed that informed, reasonable people might disagree about whether or not the document met the criteria specified by the production request and coding guidelines.

Figure 11 shows the agreement of the Authors' post-hoc assessment with the original TREC 2009 Topic Authority's determination on appeal, categorized by topic and by the Topic Authority's assessment of responsiveness. Each row shows the Topic Authority's opinion (which is necessarily the opposite of the first-pass reviewer's), the percentage of post-hoc assessments for which the Authors believe that the only reasonable coding was the one rendered by the Topic Authority, the percentage of post-hoc assessments for which the Authors believe that either coding would be reasonable, and the percentage of post-hoc assessments for which the Authors believe that the only reasonable coding contradicts the one that was made by the Topic Authority.

---

21. Gordon V. Cormack & Mona Mojdeh, Machine Learning for Information Retrieval, *in* THE EIGHTEENTH TEXT REtrieval CONFERENCE PROCEEDINGS (E. M. Voorhees & Lori P. Buckland eds. 2010), *available at* http://trec.nist.gov/pubs/trec18/papers/uwaterloo-cormack.WEB.RF.LEGAL.pdf.

22. *Practice Topic and Assessments for TREC 2010 Legal Learning Task*, TEXT REtrieval CONFERENCE LEGAL TRACK, http://plg1.uwaterloo.ca/~gvcormac/treclegal09 (last visited Feb. 20, 2012).

Topic	First-Pass Assessment	TA Assessment	TA Correct	Arguable	TA Incorrect
201	Non-Responsive	Responsive	74%	20%	6%
201	Responsive	Non-Responsive	94%	2%	4%
202	Non-Responsive	Responsive	96%	2%	2%
202	Responsive	Non-Responsive	96%	0%	4%
203	Non-Responsive	Responsive	94%	2%	4%
203	Responsive	Non-Responsive	82%	4%	14%
204	Non-Responsive	Responsive	90%	10%	0%
204	Responsive	Non-Responsive	90%	8%	2%
205	Non-Responsive	Responsive	100%	0%	0%
205	Responsive	Non-Responsive	82%	4%	14%
206	Non-Responsive	Responsive	–	–	–
206	Responsive	Non-Responsive	96%	2%	2%
207	Non-Responsive	Responsive	73%	12%	14%
207	Responsive	Non-Responsive	70%	0%	28%
All	Non-Responsive	Responsive	88% (84-91%)	8% (5-11%)	4% (2-7%)
All	Responsive	Non-Responsive	89% (85-92%)	3% (2-6%)	8% (5-12%)

Figure 11: Post-hoc assessment of documents whose first-pass responsiveness determination was overturned by the Topic Authority in TREC 2009. The columns indicate the topic number, the Topic Authority's coding decision, the percent of documents for which the Authors believe the Topic Authority was clearly correct, the percent of documents for which the Authors believe the correct assessment is arguable, and the proportion of documents for which the Authors believe the Topic Authority was clearly incorrect. The final two rows give these proportions over all topics, with 95 percent binomial confidence intervals.

VIII. Topic Authority Reconsideration

One of the Authors (Grossman) was the original Topic Authority for Topic 204 at TREC 2009. The other Author (Cormack) conducted the post-hoc assessment for Topic 204. The post-hoc assessment clearly disagreed with the Topic Authority in only one case, and was "arguable" in nine other cases. The ten documents were presented to the Topic Authority for *de novo* reconsideration, in random order, with no indication as to how they had been previously coded. For this reconsideration effort, the Topic Authority used the same three categories as for the post-hoc assessment: "clearly responsive," "clearly non-responsive," or "arguable."[23] Figure 12 shows the results of the

23. Note that when the Topic Authority originally adjudicated the documents as part of TREC 2009, she was constrained to the categories of "responsive" and "non-responsive"; there was no category for "arguable" documents. Therefore, one cannot consider a post-hoc determination of "arguable" as necessarily contradicting the Topic

Topic Authority's blind reconsideration of the ten documents. The Topic Authority repeated her original relevance determination for five of the ten documents. She reversed her original determination for three of the documents, and rendered a determination of arguable for two more. There were no instances in which the post-hoc assessment by Cormack was "clearly responsive," while the reconsideration by Grossman was "clearly non-responsive," or vice-versa. That is, the only disagreements were with respect to documents that one of the Authors coded as "arguable" and the other did not.

Doc. Id.	First-Pass Assessment	TA Assessment	Post-Hoc Assessment	TA Reconsideration
0.7.47.1151420	Responsive	Non-Responsive	Arguable	Responsive
0.7.47.1310694	Responsive	Non-Responsive	Arguable	Responsive
0.7.47.272751	Responsive	Non-Responsive	Responsive	Arguable
0.7.6.180557	Responsive	Non-Responsive	Arguable	Non-Responsive
0.7.6.252211	Responsive	Non-Responsive	Arguable	Responsive
07.47.1082536.1	Non-Responsive	Responsive	Arguable	Responsive
0.7.47.14687.1	Non-Responsive	Responsive	Arguable	Arguable
0.7.47.758281	Non-Responsive	Responsive	Arguable	Responsive
0.7.6.707917.2	Non-Responsive	Responsive	Arguable	Responsive
0.7.6.731168	Non-Responsive	Responsive	Arguable	Responsive

Figure 12: Blind reconsideration of adjudication decisions for Topic 204 by the original TREC 2009 Topic Authority (Grossman) that were contradicted or deemed arguable by the post-hoc reviewer (Cormack). The columns represent the TREC 2009 document identifier for each of the ten documents, the opinion rendered by the Topic Authority during the TREC 2009 adjudication process, the opinion rendered by the post-hoc reviewer, and the *de novo* opinion of the same Topic Authority for purposes of this study.

IX. Discussion

The results of this study support the conclusion that responsiveness—at least as characterized by the production requests and coding guidelines used at TREC 2009—is fairly well defined, and that disagreements among reviewers are largely attributable to human error. As a threshold matter, only 5 percent of the first-pass coding determinations were appealed by participating teams. Since the teams

Authority's original adjudication at TREC 2009.

had the opportunity and incentive to appeal the coding decisions with which they disagreed,[24] one may assume that, for the most part, they agreed with the first-pass assessments of the documents they chose not to appeal. Moreover, the Authors note that 89 percent of the appeals were upheld, suggesting that the appeals had, for the most part, a reasonable basis.

This study considered only those appealed documents for which the appeals were upheld—about 89 percent of the appealed documents, or 4.5 percent of all documents reviewed. Were those documents arguably on the borderline of responsiveness, as one might suspect? At the TREC 2009 Workshop, many participants, including the Authors, voiced opinions to this effect. An earlier study by the Authors preliminarily examined this question and found that, for two topics,[25] the majority of non-responsive determinations that were overturned were the result of human error, rather than questionable responsiveness.[26] The aim of the present study was to further test this hypothesis by considering the other five TREC 2009 topics and also first-pass responsiveness determinations that were overturned (i.e., adjudicated to be non-responsive by the Topic Authority). To their surprise, the Authors found nearly 90 percent of the overturned coding decisions to be clearly responsive or clearly non-responsive, consistent with the determination of the Topic Authority. The Authors found another 5 percent or so of the documents to be clearly responsive or clearly non-responsive, contradicting the determination of the Topic Authority. *Only 5 percent of the documents were found to be arguable*. Accordingly, the Authors conclude that the vast majority of disagreements were attributable to simple human error—error that can be identified by careful reconsideration of the documents using the production requests and coding guidelines.

The results of this study also suggest that the Topic Authority's responsiveness determinations, while quite reliable, are not infallible. The Authors confirmed this directly for Topic 204 by having the original Topic Authority reconsider ten documents that she had previously assessed as part of TREC 2009. For three of the ten documents, the Topic Authority contradicted her earlier assessment; for two of the ten,

---

24. *See* Hedin et al., *supra* note 4, at 3.

25. Those were Topics 204 and 207, which were chosen because they were the least technical of the seven TREC 2009 topics. *See Topic-Specific Guidelines—Topic 204*, *supra* note 13; *Topic-Specific Guidelines—Topic 206*, *supra* note 13.

26. Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 RICH. J.L. & TECH. 11, 37-43 (2011).

the Topic Authority coded the documents as arguable. For only half of the documents did the Topic Authority unequivocally confirm her previous coding decision. While the Topic Authorities for the other six topics were not available to reconsider their coding decisions, the Authors are confident from their own analysis of the documents that some of their assessments were incorrect.

All in all, the total proportion of documents that are arguable—or for which the adjudication process yielded the wrong result—appears to be quite low. Overall, 5 percent of the assessed documents were appealed; 90 percent of those appeals were upheld; and, of those, perhaps 10 percent were borderline—that is, only about 0.45 percent of the assessed documents were "arguable." It stands to reason that there may be some borderline documents that this study did not consider. In particular, the Authors did not consider documents that the first-pass reviewer and the TREC 2009 participating teams agreed upon, and that were therefore not appealed. The Authors also did not consider documents that were appealed, but for which the Topic Authority upheld the first-pass reviewer's coding decision. The Authors have little reason to believe that the number of such arguable documents would be large in either case; however, a more extensive study would be necessary to quantify this number. In any event, the Authors were concerned here specifically with the *cause* of the reviewer disagreement that was observed, and since there was no reviewer disagreement on these particular documents, this quantity has no bearing on the hypotheses being tested.

The Authors characterize this study as qualitative rather than quantitative for several reasons. The documents that were examined were not randomly selected from the document collection; they were selected in several phases, each of which identified a disproportionate number of controversial documents:

> 1. The stratified sampling approach used by TREC 2009 to identify documents for first-pass review emphasized documents for which the participating teams had submitted contradictory results;[27]

27. *See* Hedin et al., *supra* note 4, at 3.

> 2. The appeals process selected from these documents those for which the participating teams disagreed with the first-pass review;[28]
>
> 3. For the post-hoc assessment, the Authors considered only appealed documents for which the Topic Authority disagreed with the first-pass review; and
>
> 4. For the Topic 204 Topic Authority's reconsideration, the Authors considered only 10 percent of the documents from the post-hoc assessment—those for which the post-hoc assessment disagreed with the decision rendered by the Topic 204 Topic Authority at TREC 2009.

All of these steps tended to focus on controversial documents, consistent with the Authors' purpose of determining whether disagreement arose primarily due to ambiguity concerning responsiveness, or human error. Therefore, it would be inappropriate to use these results to estimate the error rate of either the first-pass reviewer or the Topic Authority on the collection as a whole.

Finally, neither of the Authors was at arm's length from the TREC 2009 effort; their characterization of responsiveness reflects their informed analysis and, as such, may be open to debate. Accordingly, the Authors invite others in the research community to examine the documents themselves and to let the Authors know their results. Towards this end, the Authors have made publicly available the text rendering of the documents they reviewed for this study.[29]

X. Conclusion

Some have argued that it is impossible to derive accurate measures of recall and precision for the results of a document review effort because large numbers of documents in every review set are "arguable," meaning that two informed, reasonable reviewers can disagree on whether the documents are responsive or not. The results of this study suggest that the number of such arguable documents is in fact quite small. This finding is consistent with Hypothesis 2—that the vast

28. *See id.*

29. *See* Text REtrieval Conference (TREC) 2009 Legal Track Documents, UNIV. OF WATERLOO, http://plg1.cs.uwaterloo.ca/~gvcormac/maura1/ (last visited Feb. 20, 2012).

majority of cases of disagreement are a product of human error rather than documents that fall in some "gray area" of responsiveness. Accordingly, it should be possible to derive a gold standard that yields accurate measures by providing reviewers with tools—such as "rulers"—that decrease their tendency to make errors, or by incorporating quality-control processes designed to detect and correct those errors. The results also show that while Topic Authorities—like all human reviewers—make coding errors, adjudication of cases of disagreement in coding using an informed senior attorney can nonetheless yield a reasonable gold standard.

XI. Acknowledgements

The Authors wish to thank the following ten individuals (listed in alphabetical order) for their good-natured participation in The Tall T's Game: Hon. Gail A. Andler, Jason R. Baron, Hon. John M. Facciola, Hon. Paul W. Grimm, Ronald J. Hedges, Ralph C. Losey, Hon. Frank Maas, Hon. Andrew J. Peck, Prof. Mark D. Smucker, and Kenneth J. Withers. (The eleventh player was the first Author.) The Authors also wish to thank Ben Kerschberg for his helpful editorial comments on an earlier draft of this Article.

	A	B	C	D	E
1		T			
2			T		T
3			T		T
4			T		T
5			T		T

Figure 13: Answer key for The Tall T's Game. The locations of the taller T's in Figure are indicated by the T's in this figure.