EJLT European Journal of Law and Technology

CORMAS: A Computerized Tool for the Analysis of Eyewitness Memory Correspondence

Richard V. De Mulder and Kees van Noortwijk [1]

Morris Goldsmith, Ainat Pansky, Asher Koriat, Shlomit Kadouri Labin [2]

Cite as: De Mulder, R.V., van Noortwijk, K., Goldsmith, M., Pansky A., Koriat, A., Labin S.K., CORMAS: A Computerized Tool for the Analysis of Eyewitness Memory Correspondence, Issue 3, Vol. 1, European Journal of Law and Technology, 2010.

Abstract

This paper presents a new approach to the study and assessment of eyewitness memory reports. The Eyewitmem Project, an interdisciplinary research initiative financed by the European Commission [3], attempts to use psychological knowledge and computer-aided document analysis as well as magnetic (fMRI) scans of the brain to measure the reliability of eyewitness statements in legal contexts. The paper focuses mainly on the document analysis part of the project, for which the CODAS software, developed by Erasmus University, is used.

1 Introduction

Eyewitness testimony is a staple ingredient of virtually all criminal legal proceedings. [4] Indeed, with the exception perhaps of a "smoking gun," few other kinds of evidence are as compelling or have as much impact on the outcome of a trial. Yet, in many cases, the faith that is placed in such testimony is unjustified. In the past 30 years, a large amount of scientific research has accrued, demonstrating the malleability and fallibility of witness memory. At the same time, reliance on erroneous eyewitness testimony has been shown to be the most common cause of the false conviction of innocent people. [5] In a study initiated by the U.S. Department of Justice, DNA evidence was re-examined for cases in which defendants were convicted prior to the forensic use of DNA technology. [6] To date, this DNA typing led to the exoneration of 172 people who were mistakenly convicted, 14 of whom were sentenced to death. Analyses of exoneration cases revealed that the majority of these innocent people were convicted on the basis of eyewitness testimony.

The reliance on eyewitness memory for solving crimes has not been significantly diminished by the development of forensic DNA tests. [7] Thus, in view of the potentially

devastating consequences of faulty witness memory, the question arises, can scientists provide courtroom judges and law-enforcement officials with the means to tell whether a witness' testimony is accurate or not? So far, forensic scientists have focused on the issue of how to distinguish a witness who is deliberately lying from one who is testifying truthfully (i.e., the detection of deception). Such efforts, including of course the development of physiological, polygraph methods, have achieved a certain amount of success. [8] Less attention has been devoted to the far more challenging problem of assessing the reliability (likely accuracy) of the memory of a witness who is attempting to testify truthfully. The challenge here is enormous given the abundant demonstrations of false memories that are endorsed by rememberers with strong conviction. Police investigators, judges and jurors typically use subjective intuitions and naïve theories of memory to estimate the extent to which what the witness reports from memory corresponds to what actually occurred. Unfortunately, these intuitions have been shown to be largely misguided. [9]

In view of the above, there is clearly a critical need to identify and develop more objective and effective tools that can help in assessing the extent to which a particular eyewitness memory report should be relied on as evidence. This goal imposes formidable theoretical and methodological challenges that have stymied progress so far. In attempting to overcome these challenges, experts from the fields of human memory, forensic psychology, neuroscience, artificial intelligence, and law have joined forces in a project called *The Assessment of Eyewitness Memory, a Multi-Componential Correspondence-Oriented Approach (short title: Eyewitmem)* **.** [10] The project involves cooperating researchers from four universities:

- The University of Haifa, Israel - Prof. Asher Koriat, Dr. Morris Goldsmith, Dr. Ainat Pansky (project coordinators) and others;
- The Erasmus University Rotterdam, The Netherlands - Prof. Richard V. De Mulder and Dr. Kees van Noortwijk and others;
- The Royal Holloway University of London (and previously, The University of Aberdeen), United Kingdom - Prof. Amina Memon and others;
- The University of Bielefeld, Germany - Prof. Hans Markowitsch and others.

The research teams at the participating universities each have their own expertise to contribute to the project. At the University of Haifa, the Institute of Information Processing and Decision Making (IIPDM) with its team of cognitive psychologists comprised of Koriat, Goldsmith, and Pansky, is well known worldwide for its expertise in the study of human memory and *metamemory*-the processes involved in monitoring and regulating one's memory. The Centre for Computers and Law at Erasmus University has performed research in automatic document classification for over 15 years and has applied the results of this in a software package called CODAS, capable of ranking documents according to certain concepts and based on user-indicated example documents. The CODAS software is applied in this project toward the goal of assessing the overall correspondence between memory reports and the actual events that they refer to. Professor Amina Memon (initially at University of Aberdeen, currently at Royal Holloway University of London), a forensic research psychologist, is a recognized expert on eyewitness memory, and particularly on the factors that affect the quantity and accuracy of the information that can be elicited from witnesses to a crime. Finally, Prof. Hans Markowitsch, a cognitive neuropsychologist

and recognized authority on forensic neuropsychology, contributes his expertise towards the search for neural and neuropsychological correlates of accurate and inaccurate remembering. In particular, he and his team have used a brain imaging technique called functional Magnetic Resonance Imaging (fMRI) to trace patterns of brain activity while test participants perform various memory-related tasks. The imaging data can then be used to assess the correctness of what a person remembers (or believes he/she remembers).

In what follows, we give a general overview of the approach we have taken, focusing in more depth on the adaptation of CODAS to assess the quality of eyewitness reports.

2 Theories About Memory

The history of experimental research on memory has been shaped by two competing conceptions. [11] One treats memory as a storehouse, and emphasizes the phenomenon of forgetting in the sense of "information loss." This conception has generally led to a quantity-oriented approach along with a relatively passive role for the person who is doing the remembering (the "rememberer"). An alternative view treats memory as a *reconstruction* of past events. In this approach the rememberer is seen to have a more active role in determining the correspondence between what is encoded and remembered and the events that actually occurred. Here the primary interest is not on simple forgetting, but rather on false remembering-on remembering events inaccurately or even remembering events or details that never occurred. These are the main differences between the competing views:

Memory as Storehous e

- Emphasis on forgetting (omission errors).
- Quantity-oriented approach.
- Passive role of rememberer.

Memory as Reconstruction

- Emphasis on false memory (commission errors).
- Accuracy/quality-oriented approach.
- Active role of rememberer.

To illustrate the reconstructive approach, consider the following classic study. [12] A series of participants were asked to wait alone in the experimenter's office under a false pretext. After less than a minute, they were taken to another room and asked to recall every detail that they had seen while waiting in the experimenter's office. After freely describing (or drawing) what they remembered, the participants were asked questions such as: "Was there a desk in the room?"; "Were there any chairs in the room?"; "How many chairs did you see: two chairs, three chairs, or four chairs?"; "Did you see a wine bottle?"; "Did you see books?"; "Did you see a picnic basket?". In fact, there were no books in the room. Yet, in this experiment nineteen out of thirty participants reported seeing books. Very few people saw the wine bottle and only one subject saw the picnic basket. This was particularly strange, as the wine bottle and picnic basket were positioned on the same shelves where some participants remembered seeing books.

Clearly, books are a very central element of the typical office "schema", whereas wine

bottles and picnic baskets are not. This classic experiment demonstrates that what people remember depends not only on what actually occurred or was present, but also on unconscious inferences regarding what should have happened. Here, people remembered the office not only based on their actual experience of seeing the office, but also on their expectations of what should have been there, based on their relevant world knowledge and prior experience of offices (generally called "schemas"). However, these inferences are often wrong. Moreover, false inferences and erroneous memory reconstructions are often the result of false information that is explicitly provided or implied by the questioning process itself, for example when the witness is asked to answer "leading" questions. [13] In sum, a wealth of laboratory and eyewitness memory research [14] has revealed that memory is quite fallible and often inaccurate. Therefore, eyewitness testimony is not always reliable, even when the witness is making a sincere effort to tell the truth. Judges, juries, and police officers rely on subjective intuitions and naive theories in evaluating the veracity of eyewitness testimony. Unfortunately, as mentioned earlier, these intuitions are often wrong.

3 The Research Goals Of The Project

In the light of these problems, the Eyewitmem research project has three main goals:

- Enhancing scientific knowledge of the cognitive and metacognitive processes underlying memory accuracy and inaccuracy.
- Developing sophisticated, multi-componential methods of assessing memory accuracy and inaccuracy.
- Providing legal personnel with diagnostic tools for estimating the reliability of individual eyewitness memory reports.

Metacognitive processes are processes that operate on one's own cognitive processes-for example, one's ability to monitoring the accuracy of the information that comes to mind when trying to recall an event, and on that basis deciding whether to report the information (answer the question) or instead respond "don't know" Metacognitive processes play a critical role in both learning and remembering, emphasizing the active role of the rememberer in determining the quality of what is remembered. "Multi-componential" refers to the different strands of the project, all of which are being brought to bear on the problem of understanding the determinants of-and hence being able to predict or "diagnose"-accurate and inaccurate eyewitness memories:

- Developing and refining a Quantity-Accuracy Profile (QAP) methodology to isolate and assess cognitive and metacognitive components of memory performance.
- Developing a quantitative (Artificial Intelligence-based) measure of overall memory correspondence.
- Examining the brain correlates of accurate and inaccurate remembering.
- Evaluating and incorporating existing diagnostic tools and variables for assessing witness memory.
- Developing and validating integrative, multi-dimensional and multi-componential models for assessing the accuracy of eyewitness testimony.

These strands will be explained in the next part of this paper.

4 The Quantity-Accuracy Profile (Qap) Methodology

To illustrate the type of challenges we are dealing with, let us presume there is an event, a bank robbery. The question presented to certain eyewitnesses is: "What was the color of the cloth that covered the robber's face?" One eyewitness says: "I'm not sure, I think it was red". The other eyewitness says: "I'm absolutely positive that it was blue". Assuming that both witnesses are sincerely trying to tell the truth, their different memories presumably stem from differences in the way in which the original even was perceived and encoded into memory, differences in the intervening events and experiences that have also been encoded into memory and may interfere or become confused with the original memories, and differences in the way in which the memories are retrieved and reconstructed. In addition, people may differ in the ability to monitor the accuracy of their own memories and withhold information that they themselves believe cannot be trusted. All of these components may lead to differences in the correspondence between what one remembers and the original "reality." (See Figure 1)

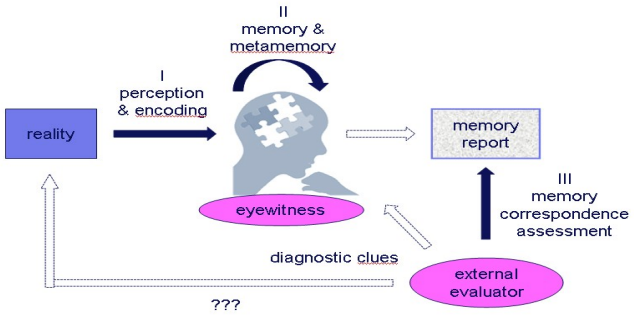


Figure 1 - The problem of evaluating the correspondence between memory reports and "reality."

A judge (or police officer) must decide which eyewitness is correct. The eyewitnesses provide their separate memory reports, and the judge must assess those reports as an external evaluator. In real-life situations, external evaluators generally do not have an independent "objective" description of the original events to which the witnesses' reports can be compared (though sometimes it may be possible to compare parts of the reports with corroborating evidence). In the laboratory, however, researchers can expose research participants to memory inputs (e.g., a film or staged crime) under controlled conditions, and then question the participants about what they remember. In this case, the original input information is known to the investigator, and hence the quality of the memory reports can be objectively assessed.

Generally speaking, memory reports can be evaluated in terms of the quantity of information they contain and the accuracy of that information. An optimal memory report provides to the external evaluator all the events and details that occurred and does not provide anything that did not occur. These two properties-quantity and accuracy-correspond roughly to the oath that witnesses are sometimes asked to take, "to tell the whole truth" (quantity) and "nothing but the truth" (accuracy). An additional property is the "grain size" of the reported information, that is, its level of precision or coarseness. [15] Some eyewitnesses may only remember the gist of what happened, whereas others may provide a very precise description of what they saw. These three properties-the

quantity, accuracy, and precision of reported information-must be taken into account and weighted in assessing the quality of eyewitness memory reports.

These are the measures of the quality of eyewitness statements that were used in this project:

Item Based Measures

- Quantity (input-bound): How many correct propositions of information were recalled and reported?
- Accuracy (output-bound): What proportion of what was reported is accurate?
- Grain size : How coarse or precise (e.g., gist vs. detail) is each item of reported information?

Overall Correspondence Measures

- Overall similarity of the report to an 'objective' description of the actual event.

With regard to each of these measures, our goal was gain a better understanding of the factors that influence the quality of the memory report, and then based on these factors, derive a set of "diagnostic clues" that can be used to predict the extent to which a given memory report is complete, accurate, and precise. These diagnostic tools can then be made available for use by judges and law-enforcement agents in real-life situations, in which the external evaluator has no direct access to the original events, and hence has no other choice but to rely on indirect clues as to the details of those events, and as to the veracity of the witnesses' description of those events.

The approach used in this project to develop such a set of diagnostic tools capitalizes on the fact that just as external evaluators lack privileged access to the accuracy of a witness' memory, and therefore must rely on indirect clues to assess its likely accuracy, the same is true for the witnesses themselves in monitoring their own memories! A great deal of research on human "metacognition" has revealed that in evaluating the veracity of their own memories, people use a variety of heuristic and analytic cues, such as how quickly and easily the information comes to mind, whether or not the memory is accompanied by vivid mental images and contextual details, whether the information is consistent with other remembered details, and so forth. [16] This metacognitive evaluation or "self-monitoring" operation is carried out "online" during the process of remembering, often without conscious awareness that one is performing such an operation. Its output is generally expressed as one's level of confidence in the veracity of the information that one remembers. Critically, people's behavior in general, and memory reporting in particular, is tightly controlled by their confidence in the remembered information. When very certain that a remembered piece of information is correct, people tend to act on it, and include it in reporting what they remember. When uncertain about its veracity, however, they tend to refrain from acting on it, and also refrain from including this information in their memory report (or in their answer to a specific question), preferring instead to respond "don't know" or "don't remember." However, the efficiency of these monitoring and control processes may vary between individuals and may also depend on the specific conditions under which the information was originally encoded and later retrieved.

In view of the above, it is critical to take into account the role of metacognitive monitoring

and control processes in determining the ultimate quality of the information that is reported from memory. It also suggests that measures based on the metacognitive processes of individual witnesses may be used by external evaluators to assess the likely quality of the information that is provided by these witnesses.

Members of the Haifa team have spent many years investigating the ways in which memory and metamemory processes jointly determine the quantity, accuracy, and precision of the information that is reported from memory. Figure 2 is a schematic illustration of their basic model [17] in the context of eyewitness testimony.

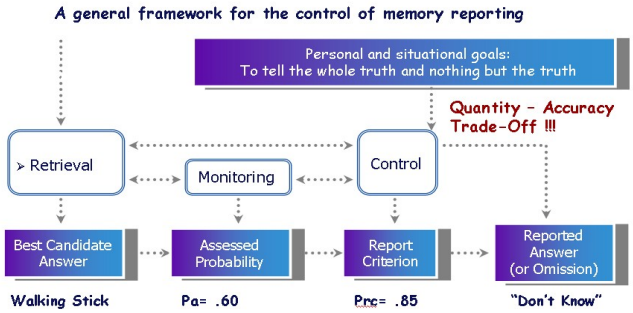


Figure 2 - A model of the strategic regulation of memory quantity and accuracy performance (adapted from Koriat & Goldsmith 1996c)

By this model, when attempting to answer a question from memory, witnesses do not simply spew out all of the information that comes to mind. Rather, once an item of information is retrieved or reconstructed from memory, a monitoring process is used to evaluate the probability that the information is correct, that this assessed probability is then compared with the report criterion that has been set in a particular reporting context: The information will be reported if its assessed probability passes the criterion; otherwise it will be withheld (responding "don't know). [18] A higher (stricter) report criterion will be set in contexts emphasizing accurate reporting (e.g., courtroom testimony), whereas a lower (more liberal) criterion will be set in contexts emphasizing the quantity of information (e.g., initial stages of an investigation; informal social interaction). To the extent that the monitoring process is effective, the use of a higher report criteria will in fact increase the accuracy of the information that is reported. However, because the monitoring process is not perfect, some incorrect information will nevertheless be reported, and some correct information will be mistakenly withheld. Thus, the general pattern is a quantity-accuracy trade-off: By utilizing one's metacognitive monitoring and control processes, memory accuracy can be increased, but this generally comes at a cost in the amount of information that is reported. .

This model, then, yields a number of cognitive and metacognitive components that potentially contribute to the quality of witness testimony:

Retention ('Memory')

- The amount and quality of the target information that can be retrieved from the eyewitness (when the "don't-know" option is denied).

Monitoring Effectiveness (Confidence Û Correctness)

- Resolution: The extent to which the person's assessed probabilities successfully

differentiate correct from incorrect candidate answers.
- Calibration bias (over/underconfidence): The extent to which the person's assessed probabilities over/under-estimate the actual probability that the answers are correct.

Report Criterion Setting

- The confidence criterion for volunteering or withholding answers, set according to competing demands for quantity and accuracy.

Control Sensitivity (Confidence Volunteering)

- The extent to which the volunteering or withholding of answers is in fact based on the monitoring output (assessed probabilities).

Each of these components can be evaluated under controlled laboratory conditions to produce a "quantity-accuracy profile" (QAP) [19] that reflects the potential memory quantity and accuracy performance that can be achieved by an individual under particular conditions, given the quality of that person's memory and metamemory processes. The scores of a particular witness on specific components of this profile can then be used to predict the quality of the memory reports that will be produced by the witness in other contexts, including those in which the actual input events are inaccessible to the external evaluator. Results from this part of the project indicate that the QAP measures based on tasks involving the recall of word lists or details of a short crime film are sufficiently reliable to provide useful predictors of memory (and meta-memory) performance in other memory contexts.

5 The Computer Aided Analysis Of Eyewitness Reports

The Centre for Computers and Law of Erasmus University, Rotterdam, is responsible for certain computational aspects of the measurement process described above. The problem this team was invited to help solve is that, at present, there is no standard method for evaluating the overall correspondence between the verbal free-narrative reports of witnesses and the actual events that are being described. Some researchers have attempted to treat the memory report as a mere "list" of individual propositions, evaluating the "truth or falsehood" of each proposition, but this method is extremely time-consuming and plagued by subjectivity with regard to such issues as what constitutes an individual proposition, how to handle propositions that are partly true and partly false, and how to take into cases in which the overall gist of the report may be accurate, despite inaccuracies in the specific details. Hence, the development of a more reliable and valid method of assessing the overall quality of free-narrative memory reports would be of great value in its own right, as it would provide a useful research tool that is currently lacking. It was also essential for achieving the goals of the present project-without a valid method of assessing the overall correspondence between free-narrative accounts and actual witnessed events, it would not be possible to identify the variables that are diagnostic of the level of that correspondence.

To achieve this goal, this strand of the project capitalized on technology that had been developed in Rotterdam to support the grading of open question exams by computer,

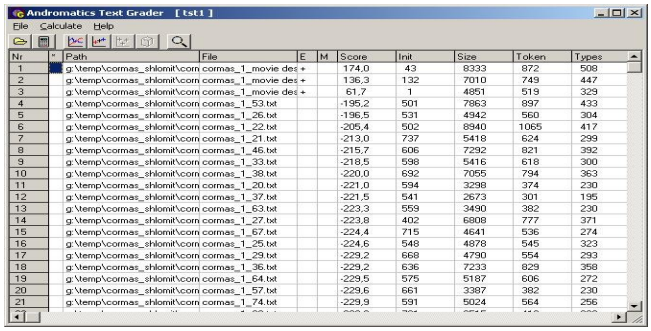
known as the Conceptual Document Analysis System ("CODAS"). The CODAS program was originally designed for the conceptual retrieval of legal texts. [20] It was later adapted for the assessment of student assignments [21]. Now, the software has been adapted again to meet the needs of the Eyewitmem project, and has become known as the "CORrespondence-oriented Memory Assessment System" (CORMAS).

The CORMAS software is used in the following way. First, eyewitnesses are shown a movie that depicts a crime. Later, these eyewitnesses have to report from their memory what they have seen in the movie. These reports are stored on a computer, as text files. The software can then be used to calculate, among other things, an overall correspondence score for each of these memory reports that reflects the likelihood that the eyewitness report is accurate, that is that the report is in accordance with what actually happened in the movie.

In order to allow the software to do so, the researcher has to indicate some examples of reports that contain an accurate description of what happened in the movie. For this, reports of 'privileged observers' can be used, observers who are given the opportunity to compose their reports while viewing the movie as often as they feel is necessary. Based on these examples (and their content), the program can calculate a score reflecting the reliability of each eyewitness's testimony. Quantity-based, accuracy-based, and gist-based criteria can be implemented, mainly by varying the type of examples that are supplied to the system. After the calculation of a set of scores, these can be plotted in charts, in several formats (see Figure 3).

The technique which is used in the programme for comparing reports and putting them in order is described in Combrink-Kuiters, De Mulder, Elffers & Van Noortwijk 1999. A central role in this process is played by the word usage in the documents. Other linguistic aspects, for example word frequency in the document, position of the word in the document, interaction effects between (pairs of) words and particularly syntactical analysis, are not used in this system. In the initial stage, the program reads the reports and produces a data matrix. In this matrix, each word type (i.e. each different word) that is found while examining all the documents is placed in a row and each document has a column. An indication is then given for each word type of the documents in which it was found.

The technique, which is then applied is reported in (Salton, 1989 p. 345-349) and can briefly be described as follows. On the basis of their appearance or non appearance in the exemplars ('good' example reports) and counter-exemplars (examples of less perfect reports), Bayesian word odds are computed for each word. Words of which the appearance or non appearance is important to the relevance of the document (the quality of the report) will have high odds. Unimportant words will have low odds. Once the word odds are calculated, document odds are computed for all reports. These document odds are obtained by multiplying all the word odds of the words that appear in the documents and then multiplying the result of this process with the odds of the non-appearance of words that are not part of the document. (Salton, 1989 p. 346) explains the condition under which the multiplication of the word odds is allowed. Although the condition (i.e. independence) is generally not fulfilled, in practice the document odds have appeared to be a good indication of the likelihood that a document is relevant.



Nr	Path	File	E	M	Score	Init	Size	Token	Types
1	g:\temp\cormas_shlomit\corm	cormas_1_movie des	+		174,0	43	8333	872	508
2	g:\temp\cormas_shlomit\corm	cormas_1_movie des	+		136,3	132	7010	749	447
3	g:\temp\cormas_shlomit\corm	cormas_1_movie des	+		61,7	1	4851	519	329
4	g:\temp\cormas_shlomit\corm	cormas_1_53.txt			-195,2	501	7863	897	433
5	g:\temp\cormas_shlomit\corm	cormas_1_26.txt			-196,5	531	4942	560	304
6	g:\temp\cormas_shlomit\corm	cormas_1_22.txt			-205,4	502	8940	1065	417
7	g:\temp\cormas_shlomit\corm	cormas_1_21.txt			-213,0	737	5418	624	299
8	g:\temp\cormas_shlomit\corm	cormas_1_46.txt			-215,7	606	7292	821	392
9	g:\temp\cormas_shlomit\corm	cormas_1_33.txt			-218,5	598	5416	618	300
10	g:\temp\cormas_shlomit\corm	cormas_1_38.txt			-220,0	692	7055	794	363
11	g:\temp\cormas_shlomit\corm	cormas_1_20.txt			-221,0	594	3298	374	230
12	g:\temp\cormas_shlomit\corm	cormas_1_37.txt			-221,5	541	2673	301	195
13	g:\temp\cormas_shlomit\corm	cormas_1_63.txt			-223,3	559	3490	382	230
14	g:\temp\cormas_shlomit\corm	cormas_1_27.txt			-223,8	402	6808	777	371
15	g:\temp\cormas_shlomit\corm	cormas_1_67.txt			-224,4	715	4641	536	274
16	g:\temp\cormas_shlomit\corm	cormas_1_25.txt			-224,6	548	4878	545	323
17	g:\temp\cormas_shlomit\corm	cormas_1_29.txt			-229,2	668	4790	554	293
18	g:\temp\cormas_shlomit\corm	cormas_1_36.txt			-229,2	636	7233	829	358
19	g:\temp\cormas_shlomit\corm	cormas_1_64.txt			-229,5	575	5187	606	272
20	g:\temp\cormas_shlomit\corm	cormas_1_57.txt			-229,6	661	3387	382	230
21	g:\temp\cormas_shlomit\corm	cormas_1_74.txt			-229,9	591	5024	564	256

Figure 3 - CODAS scores and scatter graph

The program can also produce a scatter graph of the data for these three separate measures. These are three-dimensional graphs of quantity-based, accuracy-based, and gist-based criteria. These graphs show how the measures correlate with one another (see Figure 4).

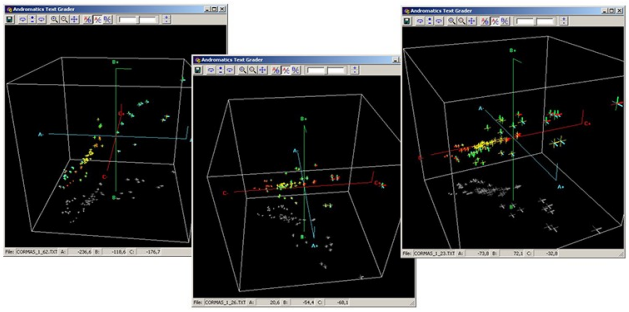


Figure 4 - CODAS 3D Scatter Graphs

Finally, the CORMAS program can compute what words were important in determining the relative scores for the testimonials. For example, the results of this analysis have shown that more specific and factual words influence the quality of the eyewitness statement in a positive sense.

Several experiments were conducted to check the reliability and validity of the CORMAS scoring procedure. Participants watched a short film or narrated slide show and were later asked to recall verbally as much as they could about what they had seen and heard. Free narrative recall was then followed by specific questions regarding particular aspects of the witnessed events (e.g., "Do you remember seeing a car parked on the street? If so, please describe it"). In some experiments, the amount of time that passed between exposure to the event and recall was manipulated (e.g., 1 week vs. 7 weeks). The participants' verbal reports were then transcribed into text files for scoring by CORMAS and two other scoring methods: (a) subjective correspondence ratings by human judges who were given privileged access to the original events (film or slide show), and (b) an item-based scoring procedure by which research assistants were trained to parse the texts into individual propositions and then rate each proposition as correct or incorrect, allowing the calculation of an item-based quantity score reflecting the number of correct propositions contained in each text.

Scores yielded by different variants of the CORMAS scoring procedure (different methods of choosing examples and counterexamples) were correlated with each other, and with the

scores yielded by the two other scoring methods. Intercorrelations between the different variants of the CORMAS scores were very high (typical $r > .90$) indicating that the CORMAS scores are quite robust. In fact, this level of inter-rater reliability was as high or higher than the inter-rater reliability of the correspondence ratings yielded by the human judges, and of the item-based quantity scores. The correlations between the CORMAS scores and the scores yielded by the two other methods were also quite high (typical $r > .8$), yielding a high level of criterion validity. More fine-grained analyses indicate that the CORMAS scores may in fact tap a "blend" of the amount of correct information, emphasized in the item-based quantity measure, and a more global correspondence judgment, captured in the ratings of the human judges. For example, in one analysis, the correlation between the CORMAS scores and the overall correspondence ratings of the human judges was $r = .88$. This correlation decreased but remained statistically reliable even when variance stemming from the simple amount of correct information, indexed by the item-based quantity scores, was partialled out (*partial* $r = .34$). This finding suggests that some unique variance in overall correspondence that is reflected in the human global judgments is also being tapped by CODAS, above and beyond the mere amount of correct information contained in the texts.

Finally, as an additional indication of construct validity, the CORMAS scores were found to differentiate between different memory conditions, yielding for example lower scores for reports provided after a longer delay between exposure and testing, and lower scores for the reports of subjects exposed to "contaminating" post-event misinformation relative to control subjects who were not exposed.

The main conclusions from the CORMAS strand of the project can be summarized as follows:

- The CORMAS procedure yields reliable and valid memory scores which capture both the overall amount of correct information (number of correct statements) scores and a more global evaluation of the overall correspondence between the contents of a free-narrative memory report, and the original events and details that actually occurred.
- The CORMAS method is standardized, in the sense that it involves a minimal amount of human intervention (in the production of examples and counterexamples), and is quite robust across minor variations in the nature of this intervention. It also has the practical advantage of requiring far less human effort than existing alternatives. It is language-independent, having been tested so far in Hebrew, English and German (only very minor adaptations are needed to handle differences between languages in the ASCII character set).
- CORMAS provides a new and convenient research tool that can be used to evaluate the overall quality of free-narrative memory reports in a standardized and reliable way.

6 The Use Of Brain Imaging

The third strand of this project is supervised by Professor Hans Markowitsch of the University of Bielefeld in Germany. His task is to identify the neural correlates of accurate

versus inaccurate remembering (and monitoring and control) using functional magnetic resonance imaging ("fMRI"). MRI scans (without the "f") are commonly performed in hospitals for medical purposes, for which patients are subjected to a "passive" scan (they just lie still while the scan takes place). The functional, fMRI procedure used in neurocognitive research, however, is rather different. While under the scanner, participants may be shown texts on a small computer screen and asked to respond to questions using a special response pad with a small number of buttons. These buttons allow the participants to provide simple responses such as 'yes', 'no', or "maybe'. More complicated manual or verbal responses are precluded because these would produce too much "noise" in the brain activation patterns that are being measured. The basic goal is to identify which parts of the person's brain are active during different types of cognitive tasks [22] (e.g., retrieving information from memory, monitoring the correctness of the information, deciding whether to venture an answer or not) and also to determine if different levels of success in these tasks are marked by particular patterns of activation, averaged across participants.

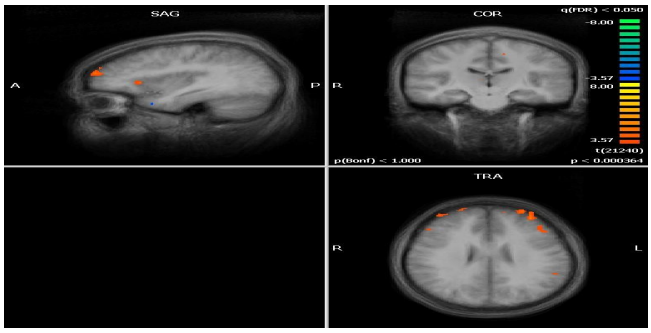


Figure 5 - Brain activity contrast: Recognizing deceptive versus truly related word pairs.

Figure 5 presents an illustrative diagram of fMRI results produced in this project. This diagram graphically depicts regions of the brain that are active when one is attempting to recognize a "deceptive" unrelated word pair (e.g., the cue NURSE - DO___R, when the studied word pair was NURSE - DOLLAR) that are not active when one is attempting to recognize a truly related word pair (e.g., the cue NURSE - DO___R, when the studied word pair was NURSE - DOCTOR). Such methods are being applied widely in the attempt to better understand where and how mental operations are carried out in the brain. In the present project strand we are examining whether patterns of brain activity can be added to the set of tools used to distinguish between correct and false remembering.

7 Evaluating And Incorporating Existing Diagnostic Tools For Assessing Witness Memory

The fourth strand of our program is headed by the forensic expert of the project, Prof. Amina Memon, who is currently at the Royal Holloway University of London. The goal of this strand is to identify, evaluate, and potentially adapt methods and variables in the existing literature that might be of use in assessing the accuracy of verbal witness reports. First, a systematic review and meta-analysis of the eyewitness memory literature was undertaken with the specific purpose of exploring how various variables interact to mediate the accuracy of eyewitness memory reports, as reported in (Memon, Meissner, &

Fraser, in press). In particular, interactions between two types of variables were examined. [23] The first class of variables, called "system variables," are variables over which the legal system has control, such as the methods that are used to elicit information from witnesses, in particular, witness questioning and suspect identification (lineup or parade) procedures. The second class, called "estimator variables," are variables which the legal system cannot control (change) but can take into account in deciding how much weight to give to the testimony of an eyewitness in a particular case. These include both situational factors such as viewing conditions, exposure duration, the amount of time that has passed since the event, and attributes of the witness, such as age and emotional state at the time of the event.

A second objective of this strand was to examine the potential usefulness of existing "interpersonal reality monitoring" content analysis tools, developed and used previously to distinguish truth-tellers from liars (i.e., deception detection) and to distinguish accounts based on memory from those based on imagination. Two such tools were examined in the project. [24] Based on the assumption that reports of experienced events differ in quality from reports of imagined or invented events, these tools specify several diagnostic criteria, such as the presence or absence of sensory, temporal, and spatial information, affective and cognitive details) that are extracted from the contents of the report by trained evaluators. The question we asked is whether these tools might also be used to distinguish accurate from inaccurate memory reports. The results of our study involving witness reports from a live (staged) scenario suggest that the usefulness of these tools for this purpose may be quite limited.

8 An Integrated Assessment Model

As stated earlier, the ultimate goal of the Eyewitmem project is to provide additional tools to help legal practitioners better appraise the accuracy of eyewitness memory reports. Each of the individual strands of the project presented so far addresses this problem from a particular perspective, utilizing partially overlapping tools and research strategies. The objective of the final phase of the project is to assemble the various measures and potential predictors of memory accuracy developed in the various threads, and combine them in an optimal manner into one or more multi-componential assessment instruments. The potential predictors include the following: QAP measures of individual differences in memory retrieval, monitoring and control, measures of brain activity patterns, system and estimator variables identified in the research literature, and measures derived from standard personality and neuropsychological test batteries. The type of model that provided the framework for this phase of the project is depicted schematically in Figure 6.

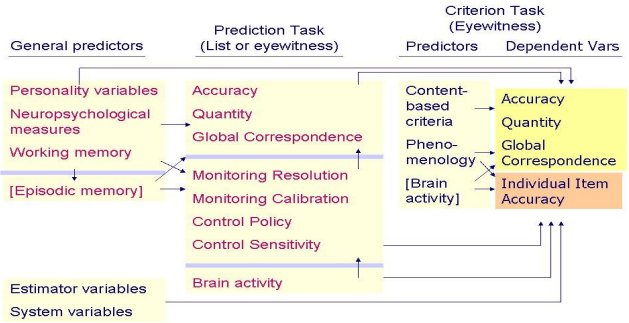


Figure 6 - Schematic depiction of a hierarchical multi-componential assessment model that could be used to predict the accuracy of memory reports, both at a global level (entire report) and at the level of individual statements.

The idea behind this scheme is that various measures of individual differences in personality, neurospsychological functioning, memory and metamemory functioning can be used to predict the general ability or tendency of a person (witness) to provide complete and accurate memory reports. These measures can be derived using parts of standard test batteries as well as using new measurement techniques (e.g., the QAP methodology described earlier) developed in the other strands of this project. These individual-difference measures, together with measures of selected system or estimator variables (e.g., the amount of time that has passed since the witnessed event, the age of the witness, the type of questioning technique, and so forth) can be used to predict the overall quality of a memory report produced by a particular witness under specific conditions. At the same time, indices tied to individual answers or specific statements made by the witness (e.g., confidence or "vividness" ratings elicited from the witness) can be added to the more global individual and situational variables to allow prediction of the accuracy or inaccuracy of these specific statements. Thus, the assessment scheme is both multi-componential and hierarchical, including predictive variables that relate to qualities of the witness, to aspects of the overall memory context, and to specific pieces of information provided by the witness in that context.

To examine the actual predictive ability of the measures and variables included in the scheme and derive the best predictive models, a final validation study was run. Each participant in this study was tested in a series of four experimental sessions across a 3-week period to obtain all of the needed measures. [25] The criterion-task performance that we attempted to predict was the accuracy of memory-based testimony regarding the events and details contained in a short (12-minute) crime film of a gas-station robbery and murder. The participants' ("witnesses'") free-narrative verbal accounts of the crime were analyzed and scored by CORMAS. In addition, item-based quantity and accuracy scores were calculated for the answers to a set of specific questions relating to details from the film. Some of these details had been deliberately "contaminated" by misinformation conveyed in an intervening questioning episode. We found that several of the QAP measures that were based on a particular participant's memory and metamemory performance on a word-list study task and on a task involving memory of a different crime film could in fact be used to predict the quality of that same participant's free-narrative account of the criterion crime episode, indexed by the CORMAS score assigned to that text. Several different predictive models were identified, including these and other variables from the tested set, which successfully accounted for a substantial amount of the between-individual variance in free-narrative memory quality. Additional models were derived that could predict, with a fair amount of success, the accuracy or inaccuracy of individual answers to specific questions, taking into account differences in the participants' rated confidence for different answers.

9 Conclusion

In this paper we described a new approach to the study and assessment of eyewitness

memory, with potentially important applications to the problem of evaluating the veracity of eyewitness testimony. What sets this work apart from previous approaches is the focus on verbal memory reports (rather than, for example, suspect lineup identification) and the multi-componential, multi-pronged attack- including the focus on both cognitive and metacognitive processes and the search for brain-activity indices related to the use of these processes. Until now, the study and assessment of verbal eyewitness memory reports has been stymied by the lack of a standard, reliable and valid method of scoring the overall accuracy of such reports. The development of CORMAS, based on the CODAS software program, constitutes an important advance in this direction, opening up many new potential avenues for both research and applications.

An interesting aspect of the Eyewitmem project is the interaction between the various research groups involved. This interaction made it possible to combine the methods developed by each of the groups, which has resulted in new and useful outcomes. For instance, the CORMAS software described in this paper was used to fill in a gap in the existing QAP methodology for measuring memory accuracy, namely the evaluation of the overall correspondence between verbal free-narrative reports of witnesses and the actual events that are being described in these reports. A second example is the use of fMRI technology in order to increase insight in the functioning of the brain when a person is remembering what happened and is reporting about this.

Some conclusions about the use of CORMAS in methods for assessing eyewitness memory reports can already be drawn. In the experiments carried out at the University of Haifa, CORMAS has yielded reliable and valid memory scores which capture both the overall amount of correct information (number of correct statements) and a more global evaluation of the overall correspondence between the contents of a free-narrative memory report and the events and details that actually occurred. Furthermore, the CORMAS model is standardized, in the sense that it involves a minimal amount of human intervention (namely, only for the production of examples and counterexamples), and is quite robust across minor variations in the nature of this intervention. It also requires less human effort than existing alternatives. The CORMAS model is language-independent with respect to the languages that have been tested so far; Hebrew, English and German. For these languages, only minor adaptations are needed to handle differences between them in the ASCII character set. The overall conclusion, therefore, is that the CORMAS software constitutes a new and effective research tool that can be used to evaluate the overall quality of free-narrative memory reports in a standardized and reliable way.

References

Brewer, W. F. & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*, *13*, 149-305.

Colwell, K., Hiscock-Anisman, C. K., Memon, A., Taylor, L., & Prewett, J. (2007). Assessment criteria indicative of deception (ACID): An integrated system of investigative interviewing and detecting deception. *Journal of Investigative Psychology and Offender Profiling*, *4*, 167-180.

Combrink-Kuiters, C. J. M., De Mulder, R. V., Elffers, H., & Van Noortwijk, C. (1999).

Comparing Student Assignments by Computer. *Cyberspace, Proceedings of the 14th BILETA conference*, York: BILETA.

Goldsmith, M., Koriat, A., & Pansky, A. (2005). Strategic regulation of grain size in memory reporting over time. *Journal of Memory and Language, 52,* 505-525.

Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). The strategic regulation of grain size in memory reporting. *Journal of Experimental Psychology: General, 131,* 73-95.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*, 67-85.

Koriat, A. & Goldsmith, M. (1996a). Memory as something that can be counted versus memory as something that can be counted on. In D. Herrmann, C. McEvoy, C. Hertzog, P. Hertel, & M. Johnson (Eds.), *Basic and applied memory research: Practical applications, Vol. 2 (pp. 3-18).* Hillsdale, NJ: Erlbaum.

Koriat, A. & Goldsmith, M. (1996b). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490-517.

Koriat, A. & Goldsmith, M. (1996c). Memory metaphors and the real-life/laboratory controversy: Correspondence versus storehouse conceptions of memory. *Behavioral and Brain Sciences, 19,* 167-188.

Koriat, A. Goldsmith, M., & Halamish, V. (2008). Control processes in voluntary remembering. In H. L. Roediger, III (Ed.),*Cognitive psychology of memory. Vol. 2 of Learning and memory: A comprehensive reference, 4 vols. (J. Byrne, Editor) (pp. 307-324) .* Oxford, UK: Elsevier.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology, 7,* 560-572.

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning and Memory, 12*, 361-366.

Markowitsch, H. J. & Kalbe, E. (2006). Neuroimaging and crime. In S. A. Christianson (Ed.), *Offender's memory of violent crime*. Chichester, UK: John Wiley & Sons.

Memon, A., Meissner, C. A., & Fraser, J. (2010). The cognitive interview: A meta-analytic review and study space analysis of the past 25 years. *Psychology, Public Policy, & Law,* in press.

Memon, A., Vrij, A. & Bull, R. (2003). *Psychology & Law: Truthfulness, accuracy and credibility of victims, witnesses and suspects* (2nd edition). Chichester, Wiley.

Mulder, R.V., De & Noortwijk, C., van (2005). *Concepts in computer aided essay assessment: Improving consistency by monitoring the assessors.* Paper presented at the Proceedings of the 20th Bileta Conference, Belfast.

Noortwijk, C., van & Mulder, R. V., De (1997). The similarities of text documents. *Journal of Information, Law and Technology*, *2*, 1-10.

Pansky, A., Koriat, A. & Goldsmith, M. (2005). Eyewitness recall and testimony. In N. Brewer & K. D. Williams (Eds.), *Psychology and law: An empirical perspective (pp. 93-150).* New York: Guilford Publications.

Reinhold, N., Kühnel, S., Brand , M., & Markowitch, H. J. (2006). Functional neuroimaging in memory and memory disturbances. *Current Medical Imaging Reviews*, *2*, 35-57.

Salton, G. (1989). *Automatic text processing; The transformation, analysis, and retrieval of information by computer*. Reading, Massachusetts: Addison-Wesley Publishing Company.

Vrij, A., Mann, S., Kristen, S., & Fisher, R. P. (2007). Cues to deception and ability to detect lies as a function of police interview styles. *Law and Human Behavior*, *31*, 499-518.

Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology*, *36*, 1546-1557.

---

[1] Centre for Computers and Law, Erasmus University, Rotterdam, the Netherlands

[2] Institute of Information Processing and Decision Making, *University of Haifa, Israel*

[3] This work was supported by a grant from the European Commission [FP6- New and Emerging Science and Technology ("Measuring the Impossible"): EYEWITMEM; 43460].

[4] Overbeck 2005.

[5] Huff, Rattner, & Sagarin 1996.

[6] Connors, Lundregan, Miller, & McEwan 1996; Wells, Malpass, Lindsay, Fisher, Turtle, & Fulero 2000.

[7] Wells & Olson 2003.

[8] See review in Granhag & Vrij 2005.

[9] Leippe 1994; Loftus 2002.

[10] See Note 3.

[11] Koriat & Goldsmith, 1996a,b; Koriat, Goldsmith, & Pansky, 2000.

[12] Brewer & Treyens 1981.

[13] Loftus 1975, 2005.

[14] Memon, Vrij & Bull 2003; Pansky et al. 2005

[15] Goldsmith, Koriat & Pansky 2005.

[16] Koriat, Goldsmith, & Halamish 2008.

[17] Koriat & Goldsmith 1996c.

[18] In a later extension of the model, one may also decide to report a relatively coarse answer when one is insufficiently confident in the correctness of a more precise answer (Goldsmith et al. 2005 ; Goldsmith, Koriat, & Wienberg-Eliezer 2002).

[19] Koriat & Goldsmith, 1996b,c; Goldsmith & Koriat, 2008.

[20] The basic characteristics of the technique involved are described in Van Noortwijk & De Mulder 1997.

[21] See De Mulder & Van Noortwijk 2005 and Combrink-Kuiters, De Mulder, Elffers & Van Noortwijk 1999.

[22] Markowitsch & Kalbe 2006; Reinhold, Kühnel, Brand & Markowitch 2006.

[23] Wells 1978.

[24] Colwell, Hiscock-Anisman, Memon, Taylor, & Prewett 2007; Vrij, Mann, Kirsten, & Fisher 2007.

[25] fMRI measures of brain activity patterns were not included in this study.