Integrated Data Representation and Formal Analysis for Human and In Silico Experimentation in Cognitive Psychology

Antonio Cerone[1(✉)] and Graham Pluck[2]

[1] Department of Computer Science, School of Engineering and Digital Sciences, Nazarbayev University, Astana, Kazakhstan
antonio.cerone@nu.edu.kz

[2] Faculty of Psychology, Chulalongkorn University, Bangkok, Thailand
graham.ch@chula.ac.th

Abstract. This paper presents a novel framework aimed at enhancing the experimental methodology within cognitive psychology, integrating both informal and formal components to accommodate varied reasoning approaches. The framework supports the design and performing of experiments, and the structured storage of the information on outcomes in a way that is independent of specific interpretations and findings. This facilitates the replication and reproduction of the experiments and their interpretation under alternative theories. Moreover, the formal components of the information stored can be used for in silico simulation as well as within rigorous analytical approaches, such as formal methods and process mining.

Keywords: Cognitive psychology · Experiment builder · In silico simulation · Data storage · Experimental variables · Formal methods · BRDL · Real-Time Maude · Process mining

1 Introduction

Cognitive psychology research is mostly based on experimental methods to investigate the intricacies of human thought and reasoning and how these affect human behavior. In fact, around 97% of published research in cognitive psychology reports on experiments [28]. Key points of these research methods are the *interpretation*, normally based on a specific, commonly-accepted theory, and the *reproducibility* of the experiments. Reproduction refers to repeating analyses on the original data set to see if the same interpretation is reached. This is in

Work partly funded by Project "K-RSL4P—Kazakh-Russian Sign Language Processing: Data, Tools and Interaction", Nazarbayev University, Kazakhstan (Award number: 11022021FD2902).

G. Broccia and A. Cerone (Eds.): DataMod 2023, LNCS 14618, pp. 3–23, 2025.
https://doi.org/10.1007/978-3-031-87217-4_1

contrast to replication, which involves using the same experimental methods to collect a new data set and see if the same interpretation is reached [22]. The theory on which the experiments are based provides the assumptions on which the researcher builds the interpretation. However, such assumptions also tend to drive the way data are preserved for possible reproduction of the experiments. This may results in a biased selection of the 'relevant' data, which potentially excludes some aspects of the data that could be helpful for exploring alternative interpretations of the experiment outcomes.

In the literature, we can find a number of frameworks and computer applications that aim at facilitating the design and performing of experiments in behavioral science, including psychology, neuroscience and linguistics [4,19,21]. However, some of these tools (e.g., PEBL [19] and PsychoPy2 [21]) may be applied to many psychological or even behavioural science disciplines, and such a generality conflicts with the possibility of defining a standard set of controls, stimuli and reactions to be preserved in the data storage process. Other tools, on the other hand, have a very restricted focus, such as the comparison of computer applications in a human-computer interaction context (e.g., CogTool [4]). Moreover, these tools often fall short in offering a cohesive structure that marries informal reasoning with formal, executable models.

In this paper, we define a framework that focuses on experiments conducted within the area of cognitive psychology and define a general experimental design model that integrates informal and formal aspects as well as a way to provide a structured description of experimental data.

In Sect. 2, we informally describe how experiments with human subjects are conducted in cognitive psychology and we introduce some terminology. In particular, we focus on experiments aiming to understand how human memory works and how its logical components interact with each other and support various aspects of human reasoning and behaviour. Two categories of classical experiments are introduced in order to provide illustrative examples through the paper. In Sect. 3, we define our experimental design model. In Sect. 4, we show how our framework can be used to perform in silico experiments and apply formal methods, specifically model checking, in order to compare and validate alternative cognitive models. We also suggest how to carry out process mining to explore the outcomes of experiments with human subjects in order to discover anomalous behaviour, enhance the in silico model and analyse its conformance with the reality. In Sect. 5, we draw conclusions and explore possible future work.

2 Experiments in Cognitive Psychology

A typical experiment in cognitive psychology consists of two main parts:

presentation in which some *stimuli* are presented to the subject aiming at a certain *target*;
reaction in which the subject has to react to the stimuli by providing a *response*, which may or may not be required to equal the target.

These two parts may occur sequentially as two separate phases, whereby all the stimuli are provided first and then the subject provides the responses, or be interleaved, whereby the subject provides a response after each stimulus.

A *stimulus* may be of several kinds:

numerical such as a digit of a number or any number or anything with a numerical semantics (for example, a number of dots denoting a number);
character-based such as any utterable character in a written language or any sequence of such characters, in any case with no explicit semantics associated with it (e.g. meaningless words);
semantics-based such as a meaningful word, statement or question;
symbolic such as a simple symbol or icon, which the subject can associate with a precise semantics either conventionally or intuitively;
visual such as a detailed drawing or photo or other forms of images, whose richness and multiplicity of information stimulates complex analysis rather than immediate associations.

Such a categorisation of stimuli covers all possible variants we normally find in experiments conducted within the area of cognitive psychology. However, in this paper, we will limit our illustrative examples to meaningless but utterable sequences of letters (*character-based stimuli*), and words and sentences (*semantics-based stimuli*).

A *target* may be one of the following:

- identical to the stimulus,
  - either to be recalled by the subject (*recall experiment*),
  - or to be recognised among a number of presented alternatives (*recognition experiment*);
- matching the stimulus, with respect to some relation between the two (*matching experiment*), where
  - either a *unique matching* is correct,
  - or *alternative matchings* are possible and any of them may be the single outcome of the response,
  - or *multiple matchings* have all to be captured as the outcome of the response.

In a recognition experiment, the wrong alternatives associated with the target are called *foils* or *distractors*. If the stimulus is semantics-based, various kinds of matching experiment are possible. For example, if the stimulus is a meaningful word, various forms of association may be investigated (the target may be another word or even an image); if the stimulus is a statement, the subject may be asked to assess its validity (the target is either 'true' or 'false' and the foils have the opposite truth value); if the stimulus is a question, the subject may be asked to answer (the target is the correct answer).

The purpose of an experiment is expressed in terms of research statements, called *hypotheses*, that make a prediction on its outcome. The experiment either *confirms* or *rejects* the hypotheses. It is important to note that a hypothesis can

only be confirmed but never be proved to be true, whereas rejecting a hypothesis means to prove that it is false.

In general, a stimulus may have a large set of possible values, even an infinite set. For example, meaningful words of a language are normally of the order of hundreds of thousands, and the number of meaningful sentences is potentially infinite. However, the hypotheses normally identify aspects of stimuli that are apt to manipulation as well as the effect of such a manipulation. Such aspects are called *independent variables* (IVs). The manipulation consists in assigning different values to the IV, which we call *levels*. For example, we might consider the number of syllables of a meaningful or meaningless word as an IV; the levels are the possible values of this number.

The effect of the manipulation is captured by *dependent variables* (DVs), which are usually directly identified by the hypotheses. In fact, a hypothesis is normally stated in terms of the difference in the DV that is supposed to be caused by distinct levels of the IVs. A hypothesis is normally conceived as a general statement in natural language, but is then made more formal once the IVs and DVs are identified. For example, if our general hypothesis is that it is easier to remember meaningful rather than meaningless words, then we need to use two kinds of stimuli, character-based for meaningless words and semantics-based for meaningful words. Then the IV could just be a boolean, with two possible levels, meaningful and meaningless, and the DV would be the percentage of words remembered. Therefore, the hypothesis can be stated more formally as 'the percentage of meaningful words remembered is higher than the percentage of meaningless words remembered.'

Distinct ways of manipulating IVs are called *conditions*. When there is only one IV with only two levels, which describe the presence and absence of a condition, then we talk about *experimental condition* and *control condition*, respectively. With reference to the example above, the condition is the presence of meaning in the word, thus meaningful words are the experimental condition and meaningless words are the control condition. When we have more than one IV and/or one or more IVs that can have more than two values, then we may have alternative experimental conditions

At the other end of the spectrum, an extreme case in cognitive psychology experiments is when we have a single condition aiming at defining the profile of an experimental subject. For example, we might want to determine the level of knowledge of word meanings of a subject who is going to take part in the experiment described previously, since the same word might be meaningful for a subject with a high level of knowledge and meaningless for a subject with a low level of knowledge.

Research hypotheses are normally based on an existing theory and the corresponding experiments aim at validating or extending such a theory. Therefore, the theory or some of its aspects define the assumptions under which the experiments are carried out. Some of these assumptions are *essential* while others are just *preferred*. For example, in the memory experiment discussed above, an essential assumption is that memorisation is facilitated by linking what we try to

memorise (e.g. a specific word) with what we already know (the meaning of the word). Moreover, a preferred assumption could be that, whenever possible, we try to associate the word with some form of visualisation driven by its meaning.

In order to illustrate this terminology, in the following two sections we consider two classical categories of experiments, which will be our ongoing case studies throughout the paper. In general, an essential assumption defines a theoretical framework underlying the hypothesis that the experiment aims to validate, while preferred assumptions are compatible with such a framework, but cannot be confirmed by the experiment.

2.1 Collins and Quillian's Experiments on Semantic Memory (SM)

In the 1960s Ross Quillian produced a computer program to model a semantic network of about 850 words [24]. Within the data representation, all properties were linked to each other within interconnected hierarchies. This model used multiple kinds of inter-connections between properties to produce an in silico representation that sat somewhere between natural language and symbolic logic, and aimed to efficiently represent meaning in terms of English-language dictionary definitions. Initially the efficiency of the semantic network was tested by comparing how well intersections of meaning could be calculated. However, as the network was biologically-inspired, based on a model of how a human semantic network within long-term memory might be organised, later studies examined experimentally how people responded to requests, using response time as indicator of processing efficiency. These experimental studies on human cognition (SM) aimed to test whether human performance was consistent with the semantic network previously modelled in silico, or on the other hand, it was consistent with a model in which attributes of concepts were stored repetitively locally. For example, whether the concept of living thing was associated with each entry for different types of animal, or whether it was associated with the higher-order concept of animal, and inherited by all subtypes of animal.

We consider experiments belonging to the category of unique matching that were conducted by Collins and Quillian [14]. The general hypothesis was that if the previously modelled semantic network were true for human long-term memory stores, response times would increase as greater distance across the network was needed to verify statements. A typical semantics-based experiment involved presentation of statements in text on a computer screen [14]. For the presentation phase each sentence was shown for two seconds, and the experimental subject, sat in front of the screen, gave their reaction by pressing a switch with one index finger to indicate true, or with their other index finger to indicate false. This was repeated over multiple trials to build up a set of response times for different statement types. In the classical SM experiments by Collins and Quillian [14] stimuli are statements in the English language. At the most basic level, there were two types of statements, those in which the correct response would be 'true' (e.g., 'An elm is a plant') or those in which the correct response would be 'false' (e.g., 'A poplar has thorns'). In fact, the target can be either value 'true' (e.g., for 'An elm is a plant') or value 'false' (e.g., for 'A poplar has thorns'). The

former of these, verifications of the truth of statements, are of primary interest. Obviously, it is necessary to include an equal number of trials with statements which are not true, randomly assorted with the trials containing true statements, otherwise the subject would be able to accurately respond 'true' on each trial without processing the meaning.

At a finer level, statements could be either property (P-statement) or superset (S-statement). A P-statement refers to a statement that can be verified by moving with n steps through the hierarchy to reach a property that affirms the statement. For example, given the hierarchy 'canary → bird → animal', the 'is yellow' attribute could be a property of 'canary'. In contrast, when n steps through the hierarchy are required to verify a statement that occurs by encountering the superset, this is called an S-statement. As an example, to decide that a canary is a bird, one has to move one step up the hierarchy from the word 'canary' to the word 'bird'. In this notation, we can also add the number of steps taken, thus an Sn-statement is the superset statement and a Pn-statement is a property statement, for which we have to move n levels up in the hierarchy to verify truth. For example, 'a canary is yellow' is a P0-statement, 'a canary is a canary' is an S0-statement, 'a canary can fly' is a P1-statement, assuming that 'can fly' is an attribute of 'bird', and 'a canary is an animal' is an S2-statement.

Collins and Quillian [14] in their experiments assumed that

1. both directly retrieving a property at a specific level and moving up a level in a hierarchy take time;
2. the times for these two processes are additive in the two cases of
   - moving up more than one level, and
   - retrieving a property after moving up
     (in accordance with Donders' assumption of additivity [26]);
3. the time to retrieve a property is independent of the level of the hierarchy;
4. searching for properties at a specific level and moving up a level my occur in parallel.

They considered assumptions 1–3 essential and assumption 4 preferred. They investigated human performance of classifying true and false for P0, P1, and P2 statements, as well as S0, S1 and S2-statements. In the context of their experimental setting, if RT denotes the reaction time, their general hypothesis can be refined into the following two hypotheses:

SM-H1 If $n > m$ then the RT of an Sn-statement is greater than the RT of an Sm-statement
SM-H2 If $n > m$ then the RT of a Pn-statement is greater than the RT of a Pm-statement

Such hypotheses were confirmed by the results of the experiments, consistent with their theory, previously developed in silico, that human semantic memory may consist of hierarchies of connected items and properties.

The hierarchical nature of the semantic network was later deemphasised in favour of a spreading activation mechanism [13]. Indeed, many aspects of the network originally proposed have been questioned. Nevertheless, as a classic study

in cognitive science, Quillian's work modelling semantic networks remains pertinent to understanding the organisation of concepts and provides the foundations for more recent semantic network theories [16]. In particular, the overall effect of time taken to respond reflecting organisation of the semantic system is still well recognised in experimental cognitive psychology [3]. Furthermore, the sentence verification task has been adapted directly from the original network models of Collins and Quillian, and is still used in applied tools to gauge efficiency of semantic memory in clinical and psychoeducational contexts [17].

2.2 Experiments on Memory Decay and Rehearsal (DR)

A fundamental distinction is made in cognitive psychology between memory that is stored for immediate use, over a matter of seconds, usually called short-term memory (STM), and for memory traces that are robust for periods over several minutes to several years, usually called long-term memory (LTM) [1].

The evidence for dual, relativity independent stores comes from classic psychology experiments involving free recall (DR). As an example, Glanzer and Cunitz [15] used a slide projector to display a sequence of 15 words to research participants. All the words were semantic-based stimuli in the forms of simple English nouns consisting of the same number of letters. Each stimulus was presented for 1 000 milliseconds, with a 2 000 milliseconds inter-stimulus interval (presentation phase). After all stimuli were presented, the subject was asked to free recall all of the words that they are able to (reaction phase). Each subject performed these presentation-then-reaction phases several times, each time with new stimulus words. The experimenters found that free recall accuracy depended on the serial position within the list. There was a bimodal distribution producing a U shape graph, such that words early in the list were well recalled (*primacy effect*), and words late in the list were also well recalled (*recency effect*), words towards the middle were recalled relatively less frequently [15].

The experimenters, Glanzer and Cunitz, argued that this data is best explained by recall from distinct memory stores, early words from LTM, and later words from STM [15]. This distinction is still very widely accepted today and forms a cornerstone of human memory research, in both cognitive psychology and neuroscience. It is, for example, integral for the recent versions of the Multistore Working Memory Model [2], the most widely accepted cognitive interpretation of human memory. In fact, the STM store, particularly that which is targeted in the semantic-based experiments described here, is functionally equivalent to the phonological loop component of the Multistore Working Memory Model [9].

Glanzer and Cunitz's classic experiments also showed that different experimental manipulations could selectively affect the primacy and recency effects, providing further support for their functional differentiation. Firstly, the effect of lack of opportunity for rehearsal of information in the period between the presentation and reaction phases differently affects the STM-based recency effect, leaving the LTM-based primacy effect intact. Rehearsal is the process of looping the target words within STM storage (i.e., silently repeating them to yourself).

That strategy potentially holds information as long as it can be rehearsed for, negating the rapid decay which usually limits duration of STM traces to a few seconds. This is why STM is often nowadays interpreted as a phonological loop [2]. In fact, if rehearsal is blocked, phonological memory traces in STM appear to have a duration of no more than 10 s, and probably somewhat less [18].

In the original Glanzer and Cunitz experiments, they had some of the free recalls occur immediately after the last stimuli word was presented. As expected, there was a pronounced recency effect. However, in some trials there was a completion delay inserted between stimuli presentation of either 10 or 30 s, in which the subject had to wait until attempting the free recall. This effectively removed the recency effect, leaving the primacy effect unattenuated. Importantly, attenuation of the recency effect only occurred if the delay was filled with an activity that requires STM storage, such as counting out loud [15]. This is because the additional activity prevents rehearsal looping which could otherwise be used to maintain the STM trace across the completion delay. Experimental manipulation of completion delays is therefore an important factor that can be used to isolate different components of human memory storage. The second notable experimental manipulation used by Glanzer and Cunitz was presentation rate. In that way, they were able to show that slowing presentation rate enhances the LTM-based primacy effect, but not the STM-based recency effect. This is opposite to the influence of a completion delay, and likely, occurs because slow presentation of stimuli allows for greater rehearsal of items during the actual presentation phase. The ability to control the rate that stimuli are presented is therefore also an important experimental manipulation in cognitive psychology.

2.3 What Historical Data to Preserve

It is fundamental to keep a history of all instances of the experiment, including the apparatus and technology used for the setting, the interpretations used, the analysis carried out and the presentation of the findings. Such a history can be used to evaluate the impact that the apparatus and technology have on findings and to compare the outcomes of replications with those of previous experiments. This may contribute to identify pitfalls in previous interpretations and define more appropriate interpretations, replicate the experiment procedure and/or the analysis process with the support of advanced technologies which were unavailable at the time of a previous instance of the experiment, and improve the representation and presentation of the findings.

For the SM experiments, Fig. 1(a) illustrates some original outcomes. For each knowledge domain the three level of statements, Sn for supersets and Pn for properties, in the hierarchy (1, 2 and 3) are mapped onto the corresponding mean reaction time RT. According to the two hypotheses, **SM-H1** and **SM-H2**, and the essential assumptions 1–3, the resultant RT curves for the Sn and Pn should be parallel straight lines. Considering the hierarchy 'canary → bird → animal', there is an anomaly for S0 point, which corresponds to statement 'A canary is a canary' in the knowledge domain of animals. Collins and Quillian's interpretation of this anomaly is that the subjects use pattern matching on the

two occurrences of the word 'canary' rather than thinking about the meaning of the statement, as documented by spontaneous reports from several subjects, thus reducing the RT with respect to the expected retrieval time. We can observe that the use of pattern matching was possible because true S0 statements, such as 'A bird is a bird' were coupled, with false S0 statement, such as 'A canary is a fish', thus enabling the subject's strategy to classify S0 statements as true by using pattern matching and as false based on their semantics. We could then replicate the experiment by replacing 'A canary is a fish' by 'A fish is not a fish', thus preventing the subject from discriminating between the two S0 statements using pattern matching, provided that we also include true S0 statements, such as 'An insect is not a mammal', to prevent pattern matching on the word 'not'. It would be interesting to see whether this change would remove the anomaly.

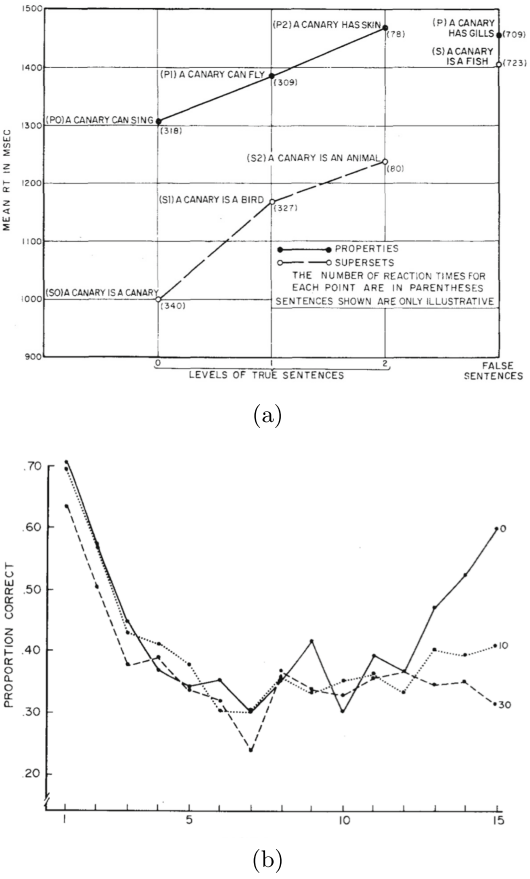

Fig. 1. (a) Average reaction times for different types of sentences in Collins and Quillian experiments (SM) [14]. (b) Serial position curve for 0, 10 and 30 sec. delay in Glanzer and Cunitz' experiments (DR) [15].

For the DR experiments, the serial position curve in Fig. 1(b) highlights the primacy and recency effects. In terms of experimental setting, it would be interesting to see how the frequency of words affects the curve. The experiment could be replicated by selecting words with a frequency within a given range, thus reducing the effect of frequency, or by using easily utterable but meaningless sequences of letters, thus ruling out the issue of frequency. In terms of interpretation and analysis, it would be interesting to understand whether the order of recall effects the outcome. In order to enable this sort of reinterpretation and alternative analysis, it is important that the order of recall is not lost while storing the dataset, although we are considering a free-recall experiment. In fact, we could speculate that if the subject tried to first recall early presented words, the time involved in the process might result in the decay of the late presented words, which have not been rehearsed, and, as a consequence, in a reduction of the recency effect. On the contrary, if the subject tried to first recall late presented words, the recency effect would be maximised and there would be no substantial impact on the primacy effect, since early words have been rehearsed.

Returning to discussion of the SM experiments, rather than using statements the experiment might have been conducted using questions (e.g., either 'Are elms plants?' and 'Do poplars have thorns?' or 'Is an elm a plant?' and 'Does a poplar have thorns?') as stimuli with a 'yes'/'no' answer as a response. Although Collins and Quillian do not discuss these possible alternatives, we can observe that the use of statements instead of questions as stimuli is more appropriate for the purpose of the experiment, that is, measuring the response time. In fact, deciding the validity of a statement is likely to better reflect the mental process of information retrieval from semantic memory and is more immediate than first decoding a question and then accessing the corresponding information stored in semantic memory.

Linguists distinguish between the structures of sentences that we speak and hear, called *surface structure* or *s-structure*, and their basic structure, called *deep structure* or *d-structure*, which is believed to be more similar to our mental representation of their meanings. When we speak and hear we apply transformational rules respectively from d-structure to s-structure and vice versa [12,23]. As a result, the additional time required for transforming the *s-structure* of a question into the *d-structure* of the corresponding statement might be greater than the actual retrieval time from semantic memory. Consequently, this could produce substantial and variable differences in response times for different questions, obscuring the actual difference between distinct retrieval times. This is especially true for the English language, in which the transformation process is quite complex due to the use of either inversion (e.g., 'Are elm plants?' and 'Is an elm a plant?') or the interrogative auxiliaries 'do', 'does' and 'did' (e.g., 'Do poplars have thorns?' and 'Does a poplar have thorns?'). Therefore, it is important that all original stimuli and responses, as well as their time- and context-related aspects, are preserved in the storage of the experiment data. This is essential not only for reproducing the experiments but also for interpreting its results.

3 Design and Dataset Description Framework

Outcomes of experiments are normally recorded in an unstructured way or, when structured, the imposed structure and information to store are usually driven by the goals of the experiments, especially in terms of the kind of planned analysis, or they might even be effected by subjective biases. Such a recording and organisation of data often determines a loss of information and, after an interpretation of the results is chosen, further information may be removed because it is considered irrelevant. If in the DR experiments the order of recall is not recorded, since it is irrelevant in a free-recall experiment, then some of the hypotheses discussed in Sect. 2.3 cannot be validated. Therefore, in this section, we provide a methodology for structuring experiment outcomes in a general way that is not affected by the goals of the experiments and is neither driven by the planned analysis nor affected by the chosen interpretations. Moreover, we provide definitions that can be used for both human and in silico experimentation.

We start considering the kind of stimulus we need to use. We can observe that a numerical or character-based stimulus is a *basic stimulus*, whose properties are *internal*, in the sense that they can always be inferred from its syntactical representation, though sometimes through complex algorithms. For example, possible properties of an integral number are being odd, even or prime, and possible properties of a meaningless, but utterable sequence of characters are its length or the number of syllables comprising it. Instead, a meaningful word has a number of *external* properties that are related to its usage and semantics rather than its syntax. It is the case of the word frequency and the word meaning. Therefore, this sort of stimulus, which we call *complex stimulus*, has to be stored in a database that associates a set of external properties with it. In the case of external properties of words or sentences, there are a number of linguistic databases to refer to. For the English language we refer to the web-based interface to the British National Corpus[1] (BNCweb). A corpus is a large and structured set of texts from speech transcription and/or written language usage, which includes annotations that identify the roles of words within sentences and extracts properties of words such as their frequencies. In addition, dictionaries, such as the Cambridge English Dictionary,[2] can be sources for the word semantics.

Therefore, complex stimuli are defined together with their properties within a database. A stimulus entry in the database is formally defined as follow.

Definition 1. *A* **stimulus entry** *consists of*

a key *for accessing the stimulus;*
a type *which can be a* text *or a* reference*;*
a representation *which is an ascii sequence if the type is text or a link to another database if the type is reference;*
external properties *whose number and quality depends on the nature of the stimulus and may be restricted depending on the nature of the experiment.*

[1] http://bncweb.lancs.ac.uk.
[2] https://dictionary.cambridge.org/dictionary/english.

A word would be of type text and would normally have its key identical to its representation. External properties associated with the nature of the word are syntactic category (e.g., noun, verb, etc.), grammatical category (e.g., for nouns: gender, number, etc.), semantic category (e.g., for nouns: concrete or abstract.), meaning definitions, frequency and dispersion.[3] Internal properties, such as length and number of syllables, do not need to be included in the entry since they are explicit in the representation.

A statement would also be of type text and its key might or might not be the same as its representation, depending on the requirements of the experiment. A meaningful code associated with the level of the statement may be more useful an access key than the statement representation. For the SM experiments, a key could encode statement type (S = superset or P = property), level in the hierarchy (0,1,2), source point in the hierarchy, truth value (0 or 1) and target point in the hierarchy. For example, statement 'An elm is a plant' would have 'S2elm1plant' as the key (true 2-level superset hierarchy: elm $\subseteq$ tree $\subseteq$ plant).

For the DR experiments, relevant external properties of words are syntactic, grammatical and semantic categories, frequency, frequency rank and dispersion, whereas a meaning definition is too fine-grain a property to be relevant. In fact, it is important to use words that are familiar (high frequency and dispersion close to 1) and can be easily visualised (concrete nouns with singular gender).

There is no rule of thumb to decide which restrictions should be imposed by the nature of the experiment. Common sense should be used here. For example, for the DR experiments, meaning definitions may not be considered relevant, since the human subject is not expected to retrieve a formal definition of the word from semantic memory, but just to read the word. Thus the only relevant semantic aspect of the word is the fact of being a singular concrete noun.

3.1 Experimental Design Model

Information on the experimental design needs to be stored in a detailed way in order to enable reproducibility and support the interpretation process. Experimental design is a very creative process and involves several interconnected aspects, which cannot be captured within a formal process but require a rigorous, though informal description, normally in natural language. As we discussed in Sect. 2.3 it is also important to keep a history of all instances of the experiment as a reference for future replication and variants of the experiment.

Definition 2. *An* **experimental design** *consists of the following components:*

experiment identifier *which is unique for each experiment;*
design version *which is a sequential number greater than 0;*
condition identifier *which is a sequential number greater than or equal to 0, with 0 denoting a* control condition*, and a positive number denoting an* experimental condition*;*

[3] Dispersion measures the distribution of a word over different parts of the corpus. A value close to 1.00 indicates that a word is perfectly spread all over the corpus.

experiment type *which can be one of the following:* recall, recognition, unique matching, alternative matchings *and* multiple matching*;*

experiment instructions *in textual form, to appear at the beginning of the presentation;*

research hypotheses *as a list in which each research hypothesis consists of an* identifier *and a* textual description*;*

set of independent variables (IVs) *with each IV consisting of an* identifier *and a* textual description*;*

set of dependent variables (DVs) *with each DV consisting of an* identifier *and a* textual description*;*

set of stimulus presentations *with each stimulus presentation associated with*

- *a* stimulus key*, which is the key in the stimulus entry,*
- *a* sequential position *which is a sequential number greater than 0,*
- *a* presentation duration *in ms (milliseconds),*
- *a* presentation location *which may be* centre, top, bottom, left *or* right*,*
- *a* presentation size *which may be* small, normal *or* large*,*
- *an* interstimulus interval*, which is the time in ms between the end of the current stimulus and the start of the next stimulus in the sequence, randomly selected between* min *and* max *time or fixed if* min = max*,*
- *a* pre-response delay*, which is the minimum time in ms after the start of the stimulus when a response can be recorded, with 0 denoting that the subject's response may already occur at the start of the stimulus,*
- *a* completion delay*, which is the maximum time in ms after the start of the stimulus when a response can be recorded, with 0 denoting that there is no limit to the delay of the response;*
- *an* IV mapping*, which maps each IV identifier into a value;*
- *a set of* targets*;*
- *a possibly empty set of* foils*;*

assumptions *as a list in which each assumption consists of an* identifier*, a* significance*, which may be either* essential *or* preferred*, and a* textual description*;*

The experiment identifier is common to all instances of the experiments, which may be carried out with different design versions.

Collins and Quillian [14] ran three different variants of the SM experiment. In our formal definition, the experiment identifier would be the same for all three variants, say SM. Each of the three variants would have a distinct sequential number as design version, that is, 1, 2 and 3. In the first experiment (design version 1) each subject reads 128 two-level sentences followed by 96 three-level-sentences. The statements are grouped in runs of 32, with a rest period of 1 min between runs. The statement appears for 2 s in normal format in the centre of the screen followed by a blank screen for 2 s before the next statement appears. The subject's response can occur any time within the 4 s between the start of the consecutive stimuli. Therefore:

- the presentation duration is 2 000 ms;
- the presentation location *centre*,
- the presentation size *normal*;
- the interstimulus interval is min = max = 2 000 ms, except for the last statement of each run for which it is min = max = 60 000 ms.
- the pre-response delay is 0 ms.
- the completion delay is 4 000 ms.

The two research hypotheses and the four assumptions have been presented in Sect. 2.1. Each presented statement is characterised by the following three IVs:

1. $k \in \{S, P\}$, the kind of statement, either S-statement or P-statement;
2. $h \in \{0, 1, 2\}$, the number of level in the hierarchy we have to move up to decide the truth value of the statement, which may be 0, 1 or 2;
3. $v \in \{0, 1\}$, the truth value of the statement, with 0 representing $false$ and 1 representing $true$.

The set of DV is $\{RT\}$. The hypotheses predict that RT is different for different levels of h. Thus, as an example, for the statement 'An elm is a plant':

- the IV mapping to values is $k = S$, $h = 2$ and $v = 1$, which are reflected in the stimulus key 'S2elm1plant'
- the set of targets is $\{1\}$, which represent the truth value $true$;
- the set of *foils* is $\{0\}$, which represent the truth value $false$.

3.2 Structured Description of Experiment Data

We see an instance of a given experiment as a sequence of structured *experimental events*. Each of them is a record of information, which includes a number of fixed components but can be extended with further components, whenever needed.

Definition 3. *An* **experimental event** *(or simply* **event***) consists of*

event identifier *to identify the event;*
experiment identifier *to identify the experiment;*
design version *to identify the design version;*
case *to specify the specific instance of an experiment with a specific subject;*
instance type *to denote the purpose of the instance of the experiment among:*
exploration *to discover information aiming at defining the human subject's profile,*
profiling *to extract subject's information aiming at defining the human subject's profile,*
learning *to deliberately carry out transfer of learning to the subject,*
practice *in which the subject gets familiar with the experiment,*
trial *in which the experiment is used to test its suitability,*
performance *in which the experiment is used with human subjects for its planned purpose,*

simulation *in which a model simulates the human behaviour,*
event type *which can be either* presentation *or* response*;*
subject identifier *to identify a specific human or model as the subject;*
value *which is either the key of the presented stimulus or the response value,*
timestamp *which is the start time of the event.*

The access to the experimental design through the experiment identifier and the design version equips an experimental event of all relevant information.

The timestamp, together with the temporal parameters defined in the experiment design, fully provides the temporal characterisation of the event and its sequential position within the case. A further benefit of storing events with timestamps, is that it allows for greater detail of analysis of the data. Traditionally, cognitive science has emphasised the importance of temporally dense data on human performance in experiments [25]. Such temporally dense data potentially allows for a microscopic analysis of processes underlying performance, and how they may be active at different time points.

When an instance of an experiment is created, the experiment identifier and the design version are selected among the available designs of experiments defined in the database, by the experimenter, who also chooses an instance type. For example, the experimenter may select design version 1 among SM experiments (*SM* identifier). The *profiling* instance type can be used to evaluate the response time of the subject to a purely visual stimulus separately for the right and left hand. This test can be implemented by asking the subject to push specifically one of the two buttons as soon as any sentence appears on the screen. The *learning* instance type can be used to analyse how learning occurs in using pattern matching rather than semantic meaning. The *practice* instance type allows subject to practice with the experimental procedure and apparatus before taking the actual experiment. The *trial* instance type is for testing the experiment. Obviously, the *performance* instance type should not be used with the same human subjects who serve in the same experiment with an incompatible instance type. For example, those who test the experiment cannot take part in the actual experiment. In fact, perfomance is incompatible with exploration, learning and trial, but it is compatible with profiling and practice. The *simulation* instance type is for in silico experiments, in which subject identifiers identify models of human behaviour rather than real humans.

The instance type is chosen by the experimenter and, depending on it, the subject identifier is either assigned by the experimenter or automatically assigned by the system. Table 1 shows the first eight events for the first SM experiment. For example, the first event (with identifier 1) is the presentation of the statement 'An elm is a plant' to subject with identifier 'H001'. The timestamp is the seconds since the epoch provided by any Python interpreter and rounded to the milliseconds. The second event (with identifier 2) is the response of the subject. The timestamp shows that it occurs 146 milliseconds ($146 = 907 - 761$) after the start of the presentation. This is the value of the dependent variable *RT*.

Table 1. Example of event log of the first eight events for the first SM experiment: 'An elm is a plant', 'A canary can eat', 'A pine is barley', 'A birch has seeds'.

Case	Ev	Exp	Vers	Instance Type	Subj	Ev Type	Value	Timestamp
1	1	*SM*	1	performance	H001	presentation	S2elm1plant	1716378651.761
1	2	*SM*	1	performance	H001	response	*true*	1716378651.907
1	3	*SM*	1	performance	H001	presentation	P2pine0barley	1716378655.761
1	4	*SM*	1	performance	H001	response	*false*	1716378656.099
1	5	*SM*	1	performance	H001	presentation	P2canary1eat	1716378659.761
1	5	*SM*	1	performance	H001	response	*true*	1716378660.103
1	7	*SM*	1	performance	H001	presentation	P2birch1seeds	1716378663.761
1	8	*SM*	1	performance	H001	response	*true*	1716378664.081

4 Human Versus In Silico Experiments

Experiments with human subjects and in silico experiments may be compared according to three perspectives:

global perspective in which the global result of the experiment (for example, the serial position curve in DR) for human subjects is compared against in silico models;
local perspective in which the result of the experiment (for example, the serial position curve in DR) for each category of human subjects with a given profile is compared with the result for the in silico model for that category.
unitary perspective in which each result of the experiment on a single human subject (for example, the serial positions of the recalled words in DR) is compared with the result for the in silico model of that subject.

Although the global perspective already gives a rough idea of how the model reflects the reality, it does not directly provide an explanation about why this happens. For example, the global serial position curve in DR does not show that the primacy effect is due to rehearsal. This interpretation is the result of the analyst's intuition and reasoning in linking the outcomes of the experiments to an existing model or theory from cognitive psychology. In the case of DR, the outcomes are linked to the Multistore Working Memory Model. However, the link may be the actual result of the analyst's belief or even bias and cannot be validated using a conceptual, nonexecutable model.

If the conceptual model is implemented into an executable model, as in our previous work [8,9,11], validation is instead possible. The local perspective has the purpose to facilitate such a form of validation. Categories of subjects may be defined using various forms of data collections: interviews/questionnaires, observations or experiments. These three forms of data collection can be captured by our notion of experimental event with profiling instance type. In fact, questionnaires can be designed using a matching experiment with some empty components (e.g., research hypotheses and sets of variables) and closed questions

defined as stimulus presentations. Observations may be emulated using interactive tasks augmented with instrumented code that records reaction times, habits and other patterns of behaviour. Profiling data is then stored in the database, where it is accessed by experimental events through the subject identifier field.

For example, in the case of DR, we can validate the hypothesis that the number of words in the primacy effect may be affected by a number of parameters: the actual STM capacity, which limits the number of words that may be rehearsed simultaneously, the language used in the mental phonological rehearsal process (e.g., Welsh number words are slower to pronounce than English words, leading to an average digit span of 5.8 versus 6.6 items), which affect the measured STM capacity, the chunking ability, which increases the amount of information stored in STM as a single item, and the STM decay time, which affects the number of words we are able to rehearse. These characteristics of the human subject can be measured through preliminary profiling experiments and associated with the human subject entry in the database. Then, for each identified category of human subjects, a distinct in silico experiment is carried out with a different combination of the considered parameters: STM capacity, word processing time, chunking factor and decay time. This allows us to analyse each of the parameters in isolation as well as in various combination by running multiple instances of the in silico experiment. The comparison of each in silico experiment with an experiment with human subjects belonging to a compatible category highlights which parameters affect the primacy effect and to which extent they do so.

The profiling data can also be used singularly to define a specific model for each human subject. The resultant population of human models can be used to emulate an experimentation on the entire population of the human subjects in the database, independently of whether or not those human subjects participated in the actual experiment. The result of the emulation may be compared with actual experiments, limited to the specific models that correspond to experiment participants, using the three perspectives above. In particular, the unitary perspective allows us to carry out a fine-grained validation of the model and identify anomalous outcomes which may need further investigation. Such an investigation may lead to refinement or change of the implemented cognitive model. This approach also provides a continuous validation by predicting the outcome of future experiments with the subject in the database and by running a new single model emulation for each new human subject added to the database and a full population emulation every time the model is modified.

Our approach in modelling in silico experiments [8,9,11] also features model-checking analysis using Real-Time Maude [20]. Using the local perspective, model checking allows us to prove interesting properties concerning a specific category of human subjects. For example, we could limit the parameters considered above in the subject's profile and prove that the number of words in the primacy effect cannot exceed a given threshold that has been observed in reality. This would verify the in silico model by proving that it accurately reflects the reality.

A final usage of our framework is through process mining [27]. Process mining combines model-oriented and data-oriented approaches to analyse processes.

The central idea is to discover, monitor and improve real processes by extracting knowledge from event logs generated by observed instances of such processes. This leads to three kinds of analysis: *discovery* of a process model, its visualisation and, possibly, the identification of pitfalls; *conformance*, that is, the comparison of an existing process model with an event log; *enhancement*, by extending or improving an existing process model using information recorded in some event log. In our case, the existing process model is the one produced through in silico experiments. Event logs are a structured representation of data for which one of the fields, called *case*, characterises a specific instance of the process. As seen in Def. 3, in our experimental events, the case identifies a specific instance of an experiment with a specific subject.

The *discovery* use of process mining on experiments with human subjects can be used to analyse the dynamic of the experiment and discover and analyse pitfalls aiming at improvements in the way the experiment is designed and administered. In the DR experiments, we could find that a specific word has an anomalous behaviour, for example being recalled less often than the others, independently of its position in the sequence. We then check the word frequency and find out that it is considerably lower than the frequencies of other words.

The *enhancement* use of process mining can identify some characteristics of the human subject that are not addressed by the in silico model. For example, in the DR experiments, if the in silico model generates recalls in a random order, among the words still present in STM at the moment of recall, then event logs of experiments with human being would show a different order. Now, if this order shows a pattern, then it is necessary to enhance the in silico model to somehow consider aspects of the words stored in STM that could reflect the order observed in reality. The characteristics of the human subjects associated with a specific ordering pattern could provide important parameters to be incorporated in the in silico model to generate the order of recall.

Finally, *conformance* analysis is based on defined metrics that measure the kind and severity of the deviations in the in silico process with respect to the event logs from experiments with human beings. This measures how the behaviour of the in silico model deviates from the reality.

5 Conclusion and Future Work

In this paper we have proposed a framework to design experiments in cognitive psychology and record data on the design itself as well as the performing and outcomes of the experiments. Data is stored in a hybrid informal and formal format in order to support both informal reasoning and rigorous analysis using in silico simulation, model checking and process mining.

We are currently implementing the framework in a toolset that supports the forms of analysis described in Sect. 4. The toolset will feature the translation of event logs from experiments with human subjects into the representation of human perceptions and human actions in the Behaviour and Reasoning Description Language (BRDL) [5,6] thus supporting in silico simulation and model checking through its implementation in Real-Time Maude [8,9,11].

The toolset will build on a previous web-based tool and portal (ColMASC) for the collaborative modelling and analysis of human cognition, behaviour, and interactive systems [7]. The new toolset will also be based on a web portal and will target researchers in human-computer interaction and cognitive science by supporting collaboration in the design, performance and replication of experiments, both with human subjects and in silico, and will be equipped with tools for their comparison according to the three perspectives introduced in Sect. 4.

The toolset will include a simulation tool to be used not only for in silico experiments, but also for emulating the interaction between computer/physical components and human components. BRDL will be the modelling language for the human components while the use of various modelling languages will be possible for the computer/physical components: variants and extensions of labelled transition systems and Petri nets as well as Real-Time Maude. Additionally, domain oriented modelling and visualisation interfaces will ease the modelling process, making it accessible to cognitive scientists, psychologists and usability experts. The output of the simulation will be converted into various presentation formats, including domain-oriented visual representations, possibly animated, and natural language description according to different linguistic codes to address a variety of experts from computer science, interaction design, usability analysis, psychology, sociology, linguistics and neuroscience. Moreover, the conversion to appropriate exchange formats will provide the interface to external tools, such as process mining tools.

The toolset will also include an analysis tool, which will exploit the Real-Time Maude model-checker to carry out formal verification of interactive systems and to validate theories in cognitive psychology and neuroscience. This tool will also aim at addressing domain experts by making use of visual interfaces that support high-level definition of properties and visual representations of the outcome of the analysis process.

The collaborative aspects of the toolset will be provided through the web portal and will include:

- collaborative design of experiments and interactive systems;
- synergetic collaboration between computer system designers on one side and domain expert involved in experiments aiming at studying various aspect of the human-system interaction;
- comparison of experiments conducted by distinct teams;
- replication and reproduction of the experiments;
- reinterpretation of previous experiments;
- online testing of system protoypes;
- online access to a large number of experimental subjects and users for an extensive period of time and creation of rich subject and user profiles;
- sharing, comparison and reuse of data from subjects/user profiles and experiment outcomes.

Obviously, the toolset will be built incrementally, by incorporating existing modules developed in previous work [7–10] as well as developing the new modules. Finally, the development will proceed through three phases:

1. functionalities for design, simulation and analysis and their use through a command-line interface;
2. definition of the domain-oriented interfaces for design and the visual representations, natural language description and animations for simulation;
3. definition of the domain-oriented interfaces for property specification and the visual representations for the analysis and comparison processes.

Orthogonally to these three phases, the toolset will be applied to a number of case studies of interactive systems and experiments in cognitive psychology, neuroscience and linguistics and will be deployed on a public website.

References

1. Atkinson, R.C., Shiffrin, R.M.: Human memory: a proposed system and its control processes. In: Spense, K.W. (ed.) The psychology of learning and motivation: Advances in research and theory II, pp. 89–195. Academic Press (1968)
2. Baddeley, A.: The episodic buffer: a new component of working memory? Trends Cogn. Sci. **4**(11), 417–423 (2000)
3. Beckage, N.M., Colunga, E.: Language networks as models of cognition: understanding cognition through language. In: Mehler, A., Lücking, A., Banisch, S., Blanchard, P., Job, B. (eds.) Towards a Theoretical Framework for Analyzing Complex Linguistic Networks. UCS, pp. 3–28. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-47238-5_1
4. Bellamy, R., John, B., Richards, J., Thomas, J.: Share on using CogTool to model programming tasks. In: Proceedings of PLATEAU 2010. ACM (2010)
5. Cerone, A.: Behaviour and reasoning description language (BRDL). In: SEFM 2019 Collocated Workshops (CIFMA), LNCS, vol. 12226, pp. 137–153. Springer (2020). https://doi.org/10.1007/978-3-030-57506-9_11
6. Cerone, A.: Modelling and analysing cognition and interaction. In: Formal Methods for an Informal World — ICTAC 2021 Summer School, LNCS, vol. 13490, pp. 30–72. Springer (2023). https://doi.org/10.1007/978-3-031-43678-9_2
7. Cerone, A., Mengdigali, A., Nabiyeva, N., Nurbay, T.: A web-based tool for collaborative modelling and analysis in human-computer interaction and cognitive science. In: Proceedings of DataMod 2021. LNCS, Springer (2022). https://doi.org/10.1007/978-3-031-16011-0_12
8. Cerone, A., Murzagaliyeva, D.: Information retrieval from semantic memory: BRDL-based knowledge representation and maude-based computer emulation. In: Cleophas, L., Massink, M. (eds.) SEFM 2020. LNCS, vol. 12524, pp. 159–175. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67220-1_13
9. Cerone, A., Murzagaliyeva, D., Tyler, B., Pluck, G.: In silico simulations and analysis of human phonological working memory maintenance and learning mechanisms with behavior and reasoning description language (BRDL). In: SEFM 2021 Collocated Workshops (CIFMA). LNCS, Springer (2022). https://doi.org/10.1007/978-3-031-12429-7_3
10. Cerone, A., Ölveczky, P.C.: Modelling human reasoning in practical behavioural contexts using real-time maude. In: Sekerinski, E., et al. (eds.) FM 2019. LNCS, vol. 12232, pp. 424–442. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-54994-7_32

11. Cerone, A., Pluck, G.: A formal model for emulating the generation of human knowledge in semantic memory. In: Bowles, J., Broccia, G., Nanni, M. (eds.) DataMod 2020. LNCS, vol. 12611, pp. 104–122. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70650-0_7
12. Chomsky, N.: Language and Mind. Cambridge University Press (2006)
13. Collins, A.M., Loftus, E.F.: A spreading-activation theory of semantic processing. Psychol. Rev. **82**, 407–428 (1975)
14. Collins, A.M., Quillian, M.R.: Retrieval time from semantic memory. J. Verbal Learn. Verbal Behav. **8**, 240–247 (1969)
15. Glanzer, M., Cunitz, A.R.: Two storage mechanisms in free recall. J. Verbal Learn. Verbal Behav. **5**(4), 351–360 (1966)
16. Jones, M.N., Willits, J., Dennis, S.: Models of semantic memory. In: Oxford Handbook of Mathematical and Computational Psychology, vol. 1, pp. 232–254. Oxford University Press (2015)
17. May, J., Alcock, K.J., Robinson, L., Mwita, C.: A computerized test of speed of language comprehension unconfounded by literacy. Appl. Cogn. Psychol. **15**(4), 433–443 (2001)
18. Mueller, S.T., Krawitz, A.: Reconsidering the two-second decay hypothesis in verbal working memory. J. Math. Psychol. **53**(1), 14–25 (2009)
19. Muellera, S.T., Piper, B.J.: The psychology experiment building language (PEBL) and PEBL test battery. J. Neurosci. Methods **222**, 250–259 (2014)
20. Ölveczky, P.C.: Real-Time Maude and its applications. In: Escobar, S. (ed.) WRLA 2014. LNCS, vol. 8663, pp. 42–79. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12904-4_3
21. Peirce, J., et al.: Psychopy2: experiments in behavior made easy. Behav. Res. Methods **51**, 195–203 (2019)
22. Peng, R.D.: Reproducible research in computational science. Science **334**(6060), 1226–1227 (2011)
23. Pinker, S.: The Language Instinct. William Morrow (1994)
24. Quillian, M.R.: Word concepts: a theory and simulation of some basic semantic capabilities. Behav. Sci. **12**, 410–430 (1967)
25. Simon, H.A.: Information processing models of cognition. Annu. Rev. Psychol. **30**(1), 363–396 (1979). https://doi.org/10.1146/annurev.ps.30.020179.002051
26. Smith, E.E.: Choice reaction time: an analysis of the major theoretical positions. Psychol. Bull. **69**, 77–110 (1968)
27. Aalst, W.: Data science in action. In: Process Mining, pp. 3–23. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-49851-4_1
28. Youyou, W., Yang, Y., Uzzi, B.: A discipline-wide investigation of the replicability of psychology papers over the past two decades. Proc. Natl. Acad. Sci. **120**(6) (2023). https://doi.org/10.1073/pnas.2208863120