Csilla Rákosi
MTA-DE-SZTE Research Group for Theoretical Linguistics

Replication of psycholinguistic experiments and the resolution of inconsistencies

1. Introduction

Replications are regarded as inevitable means of securing the reliability of scientific experiments. Despite this, most replications in cognitive metaphor research are not exact repetitions, but modified or refined versions of another experiment.

Non-exact replications are often conducted in order to rule out possible systematic errors, or to test a more differentiated research hypothesis. If there is harmony between the results of the original experiment and its non-exact replication(s), then the results are mostly evaluated as *reinforcing* the original research hypothesis. If, however, the revised version of the original experiment is carried out by adherents of a rival theory, then the experimental data gained are usually found to *conflict* with the original results and prompt the *rejection* of the research hypothesis. This seems to lead to a *paradox*:

(P) Replications are
   (a) *effective tools of problem solving* because they lead to more plausible experimental results; and they are also
   (b) *ineffective tools of problem solving* because they trigger cumulative contradictions among different replications of an experiment.

This paper intends to propose a possible resolution to this paradox. It will present three case studies on replication attempts conducted within cognitive metaphor research. Section 2 will offer a first concise description of the original experiments and the replication attempts. In Section 3, a metascientific model will be presented with the help of which the relationship between original experiments and their repetitions can be described. In Section 4, this model will be applied to the replication attempts delineated in Section 2. On the basis of our findings, Section 5 will try to generalise the results and provide a solution to (P).

2. Psycholinguistic experiments on metaphor processing and their replications

2.1. Keysar, Shen, Glucksberg & Horton (2000) and its replications

2.1.1. The original experiment: Keysar, Shen, Glucksberg & Horton (2000)

Experiment 1: Experiment 1 was intended to test different predictions of Cognitive Metaphor Theory. Participants were presented with 4 kinds of scenarios:

1. *implicit-mapping scenario*: contains conventionalised expressions supposed to belong to the same conceptual metaphor as the target expression (which was always the final sentence of the scenario);[1]

[1] For example:

2. *no-mapping scenario*: conventional instantiations of the supposed mapping are replaced by expressions not related to the given mapping;[2]
3. *explicit-mapping scenario*: in addition to the implicit-mapping scenario, the supposed mapping has been made explicit by being mentioned at the beginning of the text;[3]
4. *literal-meaning scenario*: renders the target expression as literal.[4]

From Lakoff and Johnson's theory it would follow that, first, the target sentences containing novel instantiations of the given metaphor family were readily accessible and easier to understand in the case of the implicit-mapping scenario than in the case of the no-mapping scenario; second, explicit mention of the mapping should further facilitate the creation of the given metaphorical mapping. To find out whether this is the case, reading times of the final sentences were measured and compared. Literal-meaning scenarios had a control function. The authors also applied totally irrelevant filler items, quiz questions, and practice scenarios.

Experiment 2: Since the experimental data indicate that conventional metaphors are not capable of facilitating the comprehension of metaphorical expressions that belong to the same metaphorical mapping according to CMT, regardless of whether they are explicit or implicit, in Experiment 2, explicit mapping scenarios were changed for scenarios containing novel, non-conventional metaphorical expressions. The novel condition turned out to be significantly faster than the implicit or the no-mapping conditions.

Experiment 3: The authors expressed the concern that fluency and conceptual homogeneity of the literal and novel-mapping scenarios may, in comparison to implicit-mapping and no-mapping scenarios, give rise to semantic priming. This experiment tried to rule out this possible source of error. A target word in the last sentence of the novel-mapping contexts was selected on the basis of the votes of 8 participants; following this, another group of participants had to decide whether these words were English words after having read the text of different types of scenarios. Since there was no significant difference between the reaction times given in the scenarios in this lexical decision task, Keysar et al. concluded that there were no priming effects.

2.1.2. Replication: Thibodeau & Durgin (2008)

Experiment 1: Experiment 1 was an exact repetition of Experiment 2 in Keysar et al. (2000). Although the results showed a similar pattern, the authors did not draw the conclusion that the experiment is reliable, but did point out a possible systematic error source. Namely, they raised the concern that conventionality might have been confounded with the fit between contexts and targets, since novel scenarios were judged to have a better fit than conventional ones by participants.

As a scientist, Tina thinks of her theories as her contribution. She is a *prolific* researcher, *conceiving* an enormous number of new findings each year. ***Tina is currently weaning her latest child***.

[2] For example:
As a scientist, Tina thinks of her theories as her contribution. She is a dedicated researcher, initiating an enormous number of new findings each year. ***Tina is currently weaning her latest child***.

[3] For example:
As a scientist, Tina thinks of her theories as her children. She is a *prolific* researcher, *conceiving* an enormous number of new findings each year. ***Tina is currently weaning her latest child***.

[4] For example:
As a scientist, Tina thinks of her theories as children. She makes certain that she nurtures them all. But she does not neglect her real children. She monitors their development carefully. ***Tina is currently weaning her latest child***.

Experiment 2: After a thorough analysis and criticism of Keysar et al.'s (2000) Experiment 2, Thibodeau and Durgin conducted the same experiment by making use of new, improved stimulus materials. In this case, the results were inconsistent with the earlier findings: there was no significant difference between novel, conventional and literal scenarios.

Experiment 3: In a reading times experiment, there were 3 types of scenarios. In the related metaphor scenarios, the target sentence contained a novel metaphor instantiating the same metaphor family as the conventional metaphors in the previous text. In the unrelated metaphor scenarios, the target sentence and the previous text made use of different metaphor families. Non-metaphor scenarios used literal sentences. The authors found that in the related metaphor scenarios, the final sentence read significantly faster than the final sentences of the unrelated scenarios, or in the non-metaphor scenarios. This also means that the experiments resulted in a shift in the judgement concerning what data should be regarded as relevant: instead of novelty/conventionality, the key factor seemed to be matchedness/unmatchedness.

2.1.3. Commentary

The most interesting point is, of course, the evaluation of the exact replication attempt by Thibodeau and Durgin. Instead of interpreting the similar results as a sign of reliability, they rejected the original experiment as an unusable data source and conducted non-exact replications which produced contradictory results. *Thus, the positive outcome of an exact replication did not lead to a higher degree of plausibility but to the emergence of inconsistencies.*

2.2. Glucksberg, McGlone & Manfredi's (1997) experiment and its replications

2.2.1. The original experiment: Glucksberg et al. (1997), Experiment 1

The authors intended to provide empirical evidence for the claim that metaphors are, in harmony with the Attributive Categorisation View, nonreversible. The stimulus material consisted of 24 metaphors, their corresponding similes and 12 literal similarity statements, each of them in original-order, in noun-reversed and noun-phrase reversed versions.[5] Participants had to evaluate the meaningfulness of the sentences on a 0-7 scale,[6] and, in the case of ratings 1-7, they were asked to write a paraphrase of the sentence as well. The paraphrases were analysed by two independent judges. The authors found that both reversed metaphors and metaphoric comparisons obtained significantly lower meaningfulness ratings than their original counterparts, while with literal comparisons, there was no such difference. Only a few reversed metaphoric statements were equivalent in meaning with the original-order statement; most reversed metaphoric statements were explicitly or implicitly re-reversed, and some were interpreted with new grounds.

2.2.2. First replication: Chiappe, Kennedy & Smykowsky (2003), Experiment 1

The first modification to Glucksberg et al's (1997) first experiment pertains to the stimulus material: the set of the target metaphors and similes was extended from 24 to 52, and literal similes were omitted. The authors also modified the research hypothesis as follows: (a) if the

[5] For example: Original-order metaphor: *my marriage was an icebox*; noun-reversed: *my icebox was a marriage*; noun-phrase-reversed: *an icebox was my marriage*.

[6] 0 = makes no sense; 7 = makes perfect sense.

traditional comparison theory of metaphors holds, then metaphors are converted into similes and interpreted as comparisons; thus, reversing topics and vehicles should decrease the comprehensibility of metaphors and similes to a slight but equal degree; (b) if Glucksberg's ACV is correct, then non-literal similes are interpreted, similarly to metaphors, as category statements; thus, both metaphors and similes should be irreversible; (c) if the authors' "distinct statements" view holds, then metaphors function like category claims and similes like similarity claims; thus, reversal should affect metaphors more strongly than similes. The analysis of the paraphrases was conducted in two steps. First, a judge examined the original order items and identified the most frequent interpretations. As the second step, the reversed order paraphrases were classified by two further judges in such a way that they compared the reversals to the most frequent original versions, without knowing whether they were presented as metaphors or similes. In contrast to Glucksberg et al., who found that both metaphors and (metaphorical) similes received significantly lower values when reversed, Chiappe et al. came to the conclusion that reversion affected metaphors to a greater extent than similes. The results of the paraphrase analyses were considerably different from the earlier findings, too. Namely, reversed similes were accepted to a greater extent than metaphors, and most reversed items (metaphors and similes alike) were equivalent in meaning to their original counterparts. Further, re-reversal was more frequently applied for metaphors than for similes.

2.2.3. Second replication: Campbell & Katz (2006)

In Experiment 1, the authors applied the same stimulus material, tasks and scoring scheme as in Glucksberg et al.'s (1997) Experiment 1. In addition, in two booklets of four, items were presented not in isolation but in a discourse context. These contexts were written so as to invite use of the salient characteristics of the vehicle to interpret the metaphor, as identified by the two authors on the basis of the canonical order of the given metaphor. The coding of the received paraphrases (the identification of the ground of participants' interpretations) was initiated with the help of codes stipulated by the two authors, but the list of the grounds of metaphors was extended by items found in the paraphrases which were different from the grounds previously determined by the authors. One of the scorers was blind to the aim of the experiment. The results differed substantially from those obtained in Glucksberg et al. (1997) and by Chiappe et al. (2003) alike, and there were big differences between the versions with context and without context as well.

Experiment 2 aimed to test the hypothesis of Glucksberg's ACV which states that metaphors are irreversible with the help of the same stimulus material but using a different method. From this hypothesis the prediction was made that when topic and vehicle are reversed, there should be great problems finding an appropriate interpretation, and, as a consequence, reading times should be slower. The stimulus material consisted of the same 24 metaphors used in context in the previous experiment and filler passages. The items were presented in a one-word-at-a-time self-paced moving windows format. Reading latencies for each word were recorded. In the statistical analyses, reading times over five regions with canonical and with reversed order were compared: for the word before the metaphor, for the NP-topic, for the verb, for the NP-vehicle and for the word following the metaphor. Since no significant differences were found between the values of canonical and reversed metaphors, the authors came to the conclusion that this experiment does not provide support for Glucksberg's ACV.

2.2.4. Commentary

Although none of the replication attempts was an exact repetition of the original experiment, the results, and especially, the diversity of the values gained, is really perplexing. Neither the extension of the stimulus material, nor the addition of a second type of stimuli (target sentences in a discourse context), nor the methodological changes should lead to such huge differences. *However, criteria on the basis of which one could decide which version of the experiment should be accepted, are missing.*

2.3. Bowdle & Gentner (2005) and its replications

2.3.1. The original experiment: Bowdle & Gentner (2005)

Experiment 1: Participants had to indicate on a 10-point scale whether a certain idiom sounds more natural or sensible in metaphor form or in simile form. On the basis of pre-tests, the stimulus material consisted of 64 items: 32 figurative statements in both the comparison (simile) form and the categorization (metaphor) form,[7] 16 literal comparison statements[8] and 16 literal categorization statements.[9] Half of the figuratives were conventional, the other half were novel; similarly, the figuratives were either abstract or concrete. According to Gentner's career of metaphor hypothesis, novel metaphors are processed as comparisons, while conventionality results in a shift to another mode of processing, namely, categorisation. The experimental data were found to be in harmony with the predictions of the career of metaphor hypothesis, as conventional figurative statements were more acceptable in categorization form than novel figuratives. No main effect of concreteness was found, but there was an unpredicted interaction between concreteness and conventionality.

Experiment 2: In order to find out whether the grammatical form preferences mirror processing differences, the online version of Experiment 1 was conducted. That is, the same stimulus material was applied but each sentence was seen in only one form. The 32 participants read the prime sentences on the computer screen, and had to press a key when they understood the sentence and type in an interpretation of the statement. Response time was measured from the appearance of the sentence until the first key press. Moreover, aptness ratings were collected from 32 further participants with the help of a 10-point scale. The results corresponded to the predictions. First, conventional items were quicker than novel items, independently of whether they were presented as metaphors or similes. Second, novel similes were quicker than novel metaphors, and conventional metaphors were quicker than conventional similes – that is, processing times were found to be shorter whenever the processing mode according to the career of metaphor theory and grammatical form were in harmony. Furthermore, post hoc tests yielded the result that conventionality is a decisive factor in the choice of simile/metaphor form, while aptness is not.

Experiment 3: Experiments 1 and 2 do not touch upon the claim of Gentner's career of metaphor hypothesis that the shift in the processing mode of metaphors occurs gradually, as a by-product of the repetitions of the comparison process. That is, during the repeated derivation or activation of the same abstract, domain-general meaning of the vehicle term, this meaning becomes lexicalised and added as a secondary sense to the vehicle term. To test this part of Gentner's theory, the authors developed a two-stage experimental design. In the first,

[7] For example: *Friendship is like a wine* vs. *Friendship is a wine*.
[8] For example: *An encyclopedia is like a dictionary*.
[9] For example: *Pepper is a spice*.

study stage, participants saw pairs of novel similes using the same base term and they had to fill in a target term in a third example of the same structure.[10] The authors' hypothesis was that this kind of priming "would promote conventionalization of the novel base terms". In this way, the authors aimed to "speed up the process of conventionalization from years to minutes" (Bowdle & Gentner 2005: 206). The material also involved similar tasks with literal comparisons. In the second, test stage, subjects received a list of novel and conventional figuratives and had to decide whether they prefer them in simile (comparison) or metaphor (categorisation) form with the help of a 10-point scale. The base term of some figuratives was presented in the novel similes from the study stage, while others were borrowed from the literal comparisons; a third group of base terms was not present in the materials of the study stage. The prediction was that conventional figuratives should be clearly preferred in metaphor form and, accordingly, receive the highest values, while the occurrence of the base term in novel similes in the study phase should lead to significantly higher preference numbers than in the case of figuratives with no prior exposure, but the same should not hold with items in which the prime had been seen in literal comparisons. The experimental data corresponded to these predictions.

2.3.2. Replication: Jones & Estes (2006)

Experiment 1: The participants' task was to indicate on a 7-point scale whether they prefer a certain idiom in metaphor form or in simile form. On the basis of pre-tests, the stimulus material consisted of 64 pairs of high and low apt statements; 32 of these sentences had a conventional vehicle, while 32 had a novel vehicle. According to the authors, Gentner's SMT yields the prediction that the metaphor form should be preferred with conventional vehicles, and the simile form should be chosen with novel vehicles. In contrast, on the basis of Glucksberg's ACV, aptness should be the decisive factor. The experimental data provide evidence against Gentner's SMT, because categorical preference was lower with conventional vehicles than with novel items. In contrast, the data support Gluckberg's ACV, because metaphor form preference was higher with more apt items, although aptness was only marginally significant in the item analysis.

Experiment 2: This experiment was a replication of Experiment 2 by Bowdle & Gentner (2005), with two modifications. The authors applied the same stimulus material as in the previous experiment. Participants were asked to read figurative statements (either in metaphor or in simile form) on the screen and press the spacebar when they had an interpretation ready. The authors also added a second task: after typing in the interpretation in a textbox, participants had to rate on a 7-point scale the ease of the thinking which led to that interpretation. The length of the sentences was taken into consideration by the statistical analysis. The results were completely different from Bowdle & Gentner's findings: Jones & Estes found a significant main effect of aptness both in the comprehension times and in the easiness ratings.

Experiment 3: Since this experiment makes use of the same stimulus material, but used a different method from the previous two experiments by Jones and Estes, it cannot be regarded as a refined version of the original experiment by Bowdle and Gentner, or of Experiments 1 and 2.

[10] For example:
An acrobat is like a butterfly.
A figure skater is like a butterfly.
_____________ is like a butterfly.

2.3.3. Commentary

We are faced with a situation where *pairs of experiments lead to conflicting results*. That is, on the basis of three experiments which rely on the same stimulus material but apply different methods of data production, we obtain results that are in harmony with each other – but in conflict with two further experiments replicating the first two experiments. Therefore, *the second (and further) experiment(s) by the researcher who conducted the original experiment increases the original experiment's plausibility by applying a different method, but the replications of a rival researcher decrease it.*

3. Metatheoretical background

3.1. Experimental complexes

Contemporary philosophy of science rejects the idea of providing general, uniform norms for scientific theorising such as verifiability, falsifiability, etc. Instead, only tentative hypotheses with more or less restricted scope are formulated by philosophers of science on the basis of detailed case studies focusing on different aspects of research practice in special fields of scientific inquiry and from diverse historical periods.[11] This approach fits into this tendency. Its main motivation was to grasp a specific characteristic of psycholinguistic experiments. Namely, in this research field, most papers publishing experimental results involve – in contrast to other branches of science such as physics, medicine, or chemistry – not only one experiment but 3-4 similar experiments, the relationship of which, however, is not clear. They are usually not complementary but rather seem to be improved versions of one another. Despite this, their results are often interpreted in such a way that they reinforce each other, and provide converging evidence. If they were regarded in fact as improved versions of each other, then only the last member of such a chain of experiments should be taken into account and made public.

In order to grasp the relationship between (non)-exact replications and original experiments, we have to transgress the boundaries of single experiments and identify more complex structures. This motivates the elaboration of the concept of 'experimental complex':[12]

[11] "In the late 1950s, philosophers too began to pay more attention to actual episodes in science, and began to use actual historical and contemporary case studies as data for their philosophizing. Often, they used these cases to point to flaws in the idealized positivistic models. These models, they said, did not capture the real nature of science, in its ever-changing complexity. The observation language, they argued, could not be meaningfully independent of the theoretical language since the terms of the observation language were taken from the scientific theory they were used to test. All observation was theory-laden. Yet, again, trying to model all scientific theories as axiomatic systems was not a worthwhile goal. Obviously, scientific theories, even in physics, did their job of explaining long before these axiomatizations existed. In fact, classical mechanics was not axiomatized until 1949, but surely it was a viable theory for centuries before that. Further, it was not clear that explanation relied on deduction, or even on statistical inductive inferences. […] All the major theses of positivism came under critical attack. But the story was always the same – science was much more complex than the sketches drawn by the positivists, and so the concepts of science – explanation, confirmation, discovery – were equally complex and needed to be rethought in ways that did justice to real science, both historical and contemporary. Philosophers of science began to borrow much from, or to practice themselves, the history of science in order to gain an understanding of science and to try to show the different forms of explanation that occurred in different time periods and in different disciplines. Debates began to spring up about the theory ladeness of observation, about the continuity of scientific change, about shifts in meaning of key scientific concepts, and about the changing nature of scientific method. These were both fed by and fed into philosophically new areas of interest, areas that had existed before but which had been little attended to by philosophers." (Machamer 2002: 6f.)

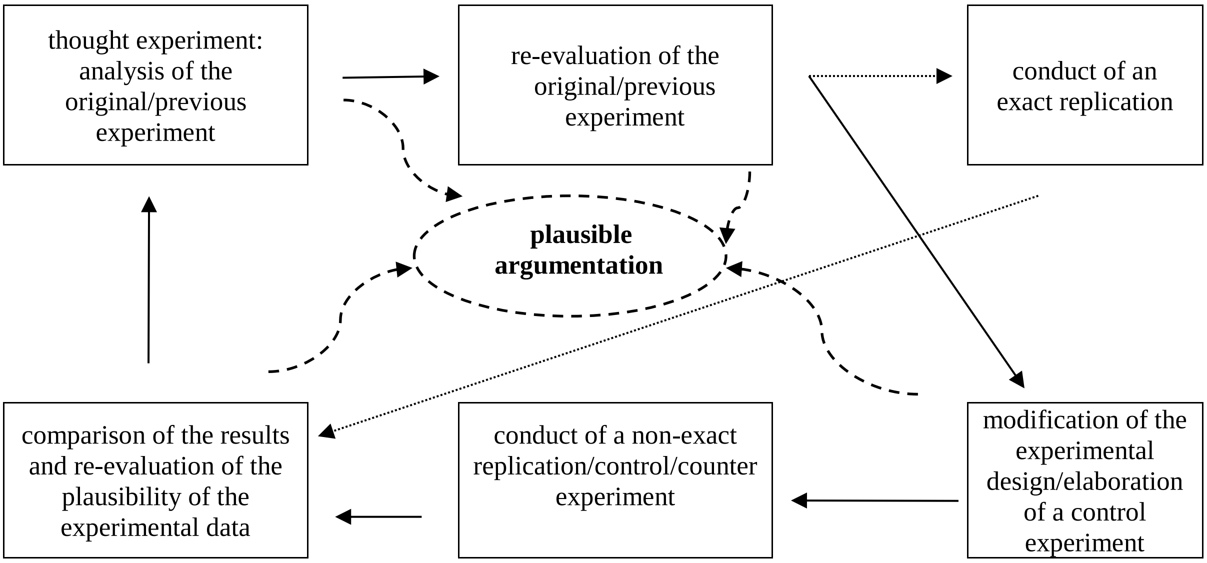
[12] For a more detailed elaboration of this concept, see Rákosi (manuscript).

(D1) An *experimental complex* consists of chains of closely related experiments which re-evaluate some part of the original experiment such as its reliability, experimental design, research hypothesis, applied methods, statistical tools, etc.

Each member of the experimental complex also re-evaluates the plausibility (acceptability) of the results obtained in the original experiment, and makes them more plausible, less plausible or shows them to be implausible.[13] Such experimental complexes are considerably more complex than single experiments, because they may involve, among other things,

- *modified (improved) versions* of the original experiment,
- *exact replications* of the original experiment or one of its non-exact replications,
- *control experiments* intended to rule out possible systematic errors in the original experiment or in one of its modifications,
- *counter-experiments*, which make the most radical revision to the original experiment by applying a different method (experimental paradigm) to the same stimulus material in order to provide evidence against the research hypothesis at issue,
- a *wider set of perceptual and experimental data*,
- *diverse perspectives* by adherents of different theories,
- *different versions of the research hypothesis*, but also
- *conflicts* emerging from different evaluations of the outcome of the original experiment,
- different kinds of *problems*, as well as *attempted solutions*;
- *a process of plausible argumentation* that re-evaluates the earlier experimental results in the light of the newer experiments in the experimental complex and tries to resolve the inconsistencies between them.[14]

Experimental complexes have a basically cyclic structure. See Figure 1.[15]


thought experiment: analysis of the original/previous experiment
re-evaluation of the original/previous experiment
conduct of an exact replication
plausible argumentation
comparison of the results and re-evaluation of the plausibility of the experimental data
conduct of a non-exact replication/control/counter experiment
modification of the experimental design/elaboration of a control experiment

[13] The notion of ‘plausibility’ is the central concept of the p-model of linguistic theorising and argumentation as presented in Kertész & Rákosi (2012, 2014) and applied to diverse fields of linguistic research.

[14] For a more thorough analysis of the argumentative aspects of psycholinguistic experiments, see Rákosi (2012, 2014), Kertész & Rákosi (2012).

[15] Simple and dotted arrows indicate successive (alternative) stages of the re-evaluation process; dashed arrows signify the argumentation process which organises the re-evaluation process.

Figure 1. The structure of experimental complexes

The aim of these *cyclic re-evaluations* is the elaboration of an experiment that is, at least temporarily, stable and generally accepted by the members of the given research field:

(D2) An experiment is a *limit* of an experimental complex, if

- (a) it evolved from the original experiment through a series of non-exact replications (that is, it results from the gradual modifications of the original experiment),
- (b) it has at least one successful exact replication (that is, it is reliable), and
- (c) it does not contain unsolved problems, so that the elaboration of further non-exact replications seems to be unmotivated (that is, it can be regarded as valid in the given informational state).

If an experimental complex has one limit, then it is called *convergent*. It is always the limit that provides the *most plausible* experimental data within the given experimental complex, because limits are free of known problems and are also reliable. Nevertheless, we should not forget that *convergence is mostly only a temporary characteristic of experimental complexes, and it is always relative to a certain informational state and research community*. That is, an experimental complex can arrive at a limit and come to a stop only pro tem and not permanently.[16]

[16] This problem is closely related to a well-known issue pertaining to the evaluation of experiments in science. Namely, most adherents of current philosophy of science share the view that there are no general criteria that would incontestably decide on the acceptability of the outcome of an experiment. Collins (1985: 84) calls this problem the *experimenter's regress*. That is, in order to ensure that a scientific experiment's outcome is correct, one has to check whether all instruments functioned perfectly and the measurements were correct. For this end, however, further instruments, measurements and the investigation of the theoretical background of their application are needed, and so on. The experimenter's regress is mostly broken by referring to *socially accepted norms*. As Kuhn has pointed out, explicit or even only implicitly accepted but in praxis often applied methodological norms determine to a considerable extent what happens in "normal science": paradigms guide the research by prescribing, among other things, how to validate perceptual data. This strategy has, of course, not only advantages but also risks because it may lead to *circularity*:

> "Scientific communities tend to reject data that conflict with group commitments and, obversely, to adjust their experimental techniques to tune in on phenomena consistent with those commitments." (Pickering 1981: 236)

To reduce this danger, Franklin (2002: 3ff., 2009) proposes a series of strategies such as experimental checks and calibration, in which the experimental apparatus reproduces known phenomena, or elimination of plausible sources of error and alternative explanations of the result (the Sherlock Holmes strategy), etc. Nevertheless, as he remarks, "[n]o single one of them, or fixed combination of them, guarantees the validity of an experimental result". This also means that *the acceptance of experimental results unavoidably contains subjective elements as well*, since the comprehensiveness of the validating process of the results cannot be achieved. At certain points, one has to make decisions that remain necessarily *arbitrary* to some extent:

> "Of course, the application of these methods is not algorithmic. They require judgment and thus leave room for disagreement." (Arabatzis 2008: 164)

Despite this, it is vital to reduce the arbitrariness of the evaluations as much as possible. The most important method for this is the application of generally accepted criteria, and, most importantly, the publication and discussion of such analyses by the research community. Of course, the criteria proposed by Kaiser (2013: 139, 141, 143), Haberlandt (1994: 9, 18), Hasson & Giora (2007: 305, 311, 316), and Keenan et al. (1990: 384), for instance, do not make it possible to make a completely objective and final decision about experiments in cognitive linguistics, either. Nonetheless, their use can lead to a well-founded re-evaluation of the results, which

These considerations provide ample support for (P)(a). Specifically, a limit of an experimental complex can be reached or at least approached with the help of increasingly elaborated non-exact replications of the original experiment. The effectiveness of this process may result from the requirement that every non-exact replication has to solve at least one unsolved problem of the original experiment or the previous members of the chain of experiments. That is, non-exact replications have to be progressive:

(D3) A non-exact replication is *progressive* if it eliminates at least one systematic error or other problem of its predecessors and/or refines the research hypothesis by taking into consideration more relevant factors. If a non-exact replication is not progressive, then it is *stagnating*.

Nevertheless, it is not the case that every progressive replication produces more plausible experimental data. The reason for this lies in the circumstance that any modification may not only rule out possible (systematic) errors but can also lead to the emergence of new ones, which, in addition, may be more serious than the resolved problem was, or may even turn out to be fatal. Thus, a progressive replication may solve a problem but also induce a dead end at the same time. Moreover, it is not always the case that non-exact replications provide increasingly similar results in the long run: quite often the opposite of this happens and the conflicts deepen and/or multiply. Therefore, the second part of (P)(b), that is, the statement that non-exact replications trigger cumulative contradictions between non-exact replications, seems to be correct, too. From this, however, it would be premature to conclude that replications are ineffective tools of problem solving. The point is that *effectiveness – in contrast to progressivity – can be judged only in the long run*. This means that we need a methodological tool which makes it possible to *describe and evaluate different strategies of inconsistency resolution*.

3.2. Strategies of inconsistency resolution

The above definition of experimental complexes does not exclude cases in which within an experimental complex, two chains of non-exact replications (or non-exact replications and counter-experiments) lead to conflicting results. These contradictions cannot be resolved simply by a mechanical comparison of the plausibility value of the results of the last member of the chains of experiments. Most frequently, it is not the current state of the cyclic process of re-evaluation that is decisive but *the assessment of future prospects*.

Therefore, the first thing to do is to *reconstruct the structure of the experimental complex*, that is, to identify the limit-candidates as well as the chains of non-exact replications, control- and counter-experiments which produce them. The second step consists of *re-evaluating the problem solving process* within the chains of experiments, and comparing them. If the inconsistencies cannot be resolved on the basis of the information at hand, then the third step should be the *determination of the directions of the continuation of the cyclic process of re-evaluation*. Basically, two strategies are possible in such situations.

The **first strategy** consists of a separate continuation of the chains of experiments by conducting further non-exact replications, counter- or control experiments, comparing the results and taking a decision. An analogue of this method was called the "*contrastive strategy*" in Kertész & Rákosi (2012). There are three basic situations:

might be acceptable to most members of the research community temporarily, on the basis of our recent knowledge regarding metaphor processing. New developments, however, may overrule them in future.

- If the elaboration of further non-exact replications of one of the chains terminates and leads to a limit of the experimental complex in the sense of (D2), while the other chain comes to a dead-end, then the conflict can be resolved in such a way that the limit is kept, while the rival chain is rejected. Clearly, the elaboration of the first chain of experiments was an effective problem solving process, while the second one was ineffective.
- If no limit can be achieved by continuing all chains, then the experimental complex is not capable of reaching a limit and the problem solving process is ineffective.
- It may also occur that both chains of experiments evolving from the same original experiment lead to a limit. In such cases, it would not be reasonable to give up either of them. Thus, this inconsistency has to be (at least temporarily, in the given informational state) tolerated by the application of the second strategy.

A **second strategy** is based on the elaboration and conduct of further experiments involving *a refinement of the research hypothesis and experimental design in such a way that all factors found relevant so far are taken into consideration*. The analogue of this method was called the "*combinative strategy*" in Kertész & Rákosi (2012). This method may make it possible to resolve contradictions between experiments conducted by researchers committed to rival approaches by *integrating* their results with the help of paraconsistent logic.

In the next section, we will apply this model to the experiments briefly presented in Section 2.

4. Reconstruction of the experimental complexes and evaluation of the problem solving process

4.1. The experimental complex evolving from Keysar et al. (2000)

4.1.1. The structure of the experimental complex

The experimental complex evolving from Experiment 1 in Keysar et al. (2000) involves one exact and three non-exact replications, and a control experiment. See Figure 2.[17]

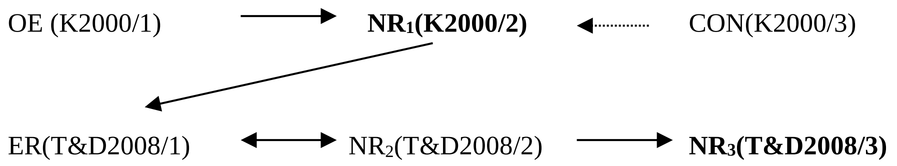


Figure 2. The structure of the experimental complex evolving from Keysar et al. (2000)

Two chains of experiments can be identified:

- NR$_1$ is an improved version of the original experiment, while CON is a related control experiment;
- ER is an exact replication of NR$_1$, while NR$_2$ and NR$_3$ are non-exact replications of ER, leading to a conflicting result.

There are two limit-candidates: NR$_1$ (Section 2.1.1), and NR$_3$ (Section 2.1.2). Thus, we have to reconstruct and re-evaluate two chains of experiments.

[17] Simple arrows lead from experiments to their non-exact replications when these are regarded as improved versions of the former. Double arrows indicate that a non-exact replication produced a conflicting result. Dotted arrows signify the relationship between experiments and control experiments. Dashed arrows are used between experiments and counter-experiments.

4.1.2. Re-evaluation of the limit-candidate by Keysar, Shen, Glucksberg & Horton (2000)

OE (cf. Section 2.1.1): The first task is the identification of the problematic points of the original experiment:[18]

Problem 1: The stimulus material is missing in the experimental report. Therefore, its correctness cannot be checked.

Problem 2: The analysis of the excerpts of the texts presented in Keysar et al. (2000: 582) and in Thibodeau & Durgin (2008: 525) indicate that metaphorical expressions in the text of the scenarios and in the related target sentence cannot be always regarded as instantiations of the same conceptual metaphor in the sense of Lakoff & Johnson (1980).

Problem 3: The stimulus material contained solely conventional metaphors. This clearly reduces the generality of the investigations.

Problem 4: Explicit-mapping scenarios start with an explicit mention of the alleged conceptual metaphor. This may have eased the comprehension of the target expression in contrast to no-mapping or implicit-mapping scenarios due to a semantic priming effect.

NR$_1$ (cf. Section 2.1.1): Experiment 2 in Keysar et al. is a progressive non-exact replication because it deals with Problem 3 insofar as it extends the stimulus material with novel metaphors. Nevertheless, it leaves Problems 1, 2 and 4 open, and leads to the emergence of some new problems, too:

Problem 5: The text of novel-mapping scenarios is (at least in some cases) more fluent, securing a better fit between the text of the scenario and the target sentence, than that of the implicit-mapping contexts.

Problem 6: The authentication of the perceptual data is controversial because there is a huge difference between the mean reading times of implicit-mapping scenarios in the two experiments, while with no-mapping scenarios, the difference is rather insignificant, and in the literal-mapping condition, the values are almost identical.

Problem 7: Novel-mapping scenarios start – similarly to explicit-mapping ones – with an explicit mention of the alleged conceptual metaphor. This may have eased the comprehension of the target expression in contrast to no-mapping or implicit-mapping scenarios due to a semantic priming effect.

Problem 8: The conceptual homogeneity of novel-mapping scenarios in comparison to implicit-mapping and no-mapping scenarios might have led to semantic priming.

CON (cf. Section 2.1.1): Experiment 3 is a control experiment aimed at providing a solution to Problem 8, whose efficiency, however, can be questioned:

Problem 9: At least in the excerpts presented in the experimental report, the target words were semantically clearly less related to the text of the scenario than other expressions of the target sentence. For example, in the case of the "ideas are

[18] See also Rákosi (2012, Part IV).

people" scenario, the target word was "weaning", while the target sentence also contained the word "child", which was semantically related to "fertile", "giving birth".

Table 1 summarises the current state of the re-evaluation process.[19]

	P1	P2	P3	P4/P7	P5	P6	P8	P9
OE	**E**	**E**	**E**	**E**				
NR_1	**O**	**O**	**P**	**O**	**E**	**E**	**E**	
CON				**O**			**O**	**E**

Table 1. Overview of the re-evaluation of the limit-candidate by Keysar et al. (2000)

4.1.3. Re-evaluation of the limit-candidate by Thibodeau & Durgin (2008)

ER (cf. Section 2.1.2): The exact replication of NR_1 leads to a similar pattern of results. Of course, Problems 1-8 emerge here, too.

NR_2 (cf. Section 2.1.2): Experiment 2 by Thibodeau and Durgin can be considered a progressive non-exact repetition because it provides a solution to Problems 1-5, and 7. Despite this, the results seem to be burdened by the following systematic errors:

Problem 10: Problem 8 has become even more severe than it was with Keysar et al.'s (2000) experiments.[20] That is, there were semantically related words in the target sentences and the texts of the scenarios with the novel metaphor, the conventional metaphor and the literal target scenarios, while this was not the case with the non-metaphoric scenarios.[21]

Problem 11: The filler scenarios were chosen on the basis of other considerations than was the case with the original experiment. Namely, Keysar et al's main motivation was to make sure that "participants would not anticipate or notice a particular pattern" (Keysar et al. 2000: 583), and in this spirit, their fillers contained neither metaphorical final sentences nor metaphors belonging to the same conceptual domains. With the new version by Thibodeau & Durgin, however, 2 in every 3 filler scenarios did contain metaphorical expressions; moreover, the fillers were intended to "avoid reading strategies that would cause people to skim over metaphors" (Thibodeau & Durgin 2008: 523). Thus, 4 of 10 questions following the fillers asked explicitly about metaphors. Therefore, participants might have discovered relatively easily that the experiment focused on the use of metaphorical expressions.

Problem 12: As the authors diagnosed, the similar reading speed of the target sentences in the novel and conventional scenarios might be due to the circumstance that

[19] In Tables 1-4, 'E' indicates that a problem has emerged, 'S' means that a solution has been put forward to the problem at issue, 'P' stands for cases when a partial solution has been offered for a problem, while 'O' signifies that the problem remains open.

[20] For a more detailed analysis, see Rákosi (2012).

[21] For example:

IDEAS ARE FOOD

Target sentence: *Otherwise, they give him* ***indigestion****.*

Novel: *David has a hard time* ***ingesting*** *new ideas. He has to* ***gnaw*** *on them for days.*

Conventional: *David has a hard time* ***swallowing*** *new ideas. He has to* ***stew*** *them over for days.*

Non-metaphor: *David takes a while to fully understand new ideas. He has to think about them for days.*

Literal-reading: *David has weak* ***stomach****. He has to take his time when* ***eating meals****.*

participants expected a metaphorical sentence after a text which also contained metaphors – independently of whether or not these metaphors belong to the same metaphorical mapping.

NR_3 (cf. Section 2.1.2): Experiment 3 aimed to solve Problem 12 and make it possible to reject an alternative explanation of the results. The practice and filler tasks were modified, too, so that they no longer asked about metaphors explicitly. Thus, NE_3 is a progressive non-exact replication. It is, however, not a limit, because a variant of Problem 10 emerges again:

Problem 13: It cannot be ruled out that the significantly shorter reading time of the consistent metaphorical scenarios was the result of semantic (lexical) priming.[22]

Table 2 shows the reconstruction of this chain of experiments.

	P1	P2	P3	P4/P7	P5	P6	P8/P10/P13	P11	P12
OE	**E**	**E**	**E**	**E**					
NR_1	**O**	**O**	**P**	**O**	**E**	**E**	**E**		
NR_2	**S**	**S**	**S**	**S**	**S**		**O**	**E**	**E**
NR_3	S	S	S	S	S		**O**	**S**	**S**

Table 2. Overview of the re-evaluation of the limit-candidate by Thibodeau & Durgin (2008)

4.1.4. Comparison of the problem solving processes

On the basis of our reconstruction, the decision of Thibodeau and Durgin regarding the rejection of NR_1 despite the successful exact replication, has become completely reasonable. Namely, both NR_1 and ER are burdened with problems which could not be eliminated. Therefore, NR_1 cannot be regarded as the limit of this experimental complex. Our analyses motivate a similar verdict with the limit-candidate NR_3. From this it follows that the conflict between Keysar et al's and Thibodeau and Durgin's results cannot be resolved on the basis of the information at our disposal at this point of the process of re-evaluation. Although Thibodeau and Durgin's results are more plausible, it would be erroneous to terminate the problem solving process at this point. Moreover, there are experiments by Gentner and her colleagues (see, above all, Gentner & Boronat 1992) whose results are in harmony with Keysar et al's findings.

4.1.5. Determination of the direction of the continuation of the cyclic process of re-evaluation

The next question is, of course, how the problem solving process should proceed. Since there were two factors (conventionality, matchedness) which seemed to be relevant, the first choice could be the application of the Combinative Strategy insofar as an experimental design should be elaborated that takes both of them into account and helps us to compare their contribution to metaphor processing. The persistent emergence of semantic priming effects, however, seriously questions the viability of this endeavour.

[22] Cf. "When David hears new ideas, he takes his time **digesting** them completely. He likes to **chew** them over slowly.
Related target sentence: They are exquisite **gourmet meals** for him. (IDEAS ARE FOOD)
Unrelated target sentence: They are exotic tropical plants for him. (IDEAS ARE PLANTS)" (Thibodeau & Durgin 2008: 537).

4.2. The experimental complex evolving from Glucksberg, McGlone & Manfredi (1997)

4.2.1. The structure of the experimental complex

This experimental complex consists of an original experiment, two non-exact replications, and a counter-experiment. See Figure 3.

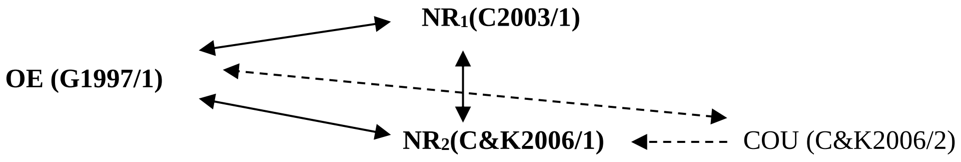


Figure 3. The structure of the experimental complex evolving from Glucksberg et al. (1997)

There is a conflict not only between the original experiment and its non-exact replications, but also between the two non-exact replications; and further, between OE and the counter-experiment COU. Thus, we have three limit-candidates: OE, NR_1, and NR_2.

4.2.1. Re-evaluation of the limit-candidate by Glucksberg et al. (1997)

OE (cf. Section 2.2.1): The first task to be undertaken is the identification of the problematic points of the original experiment.

Problem 1: Providing interpretations might require a reliance on a different representational system and skills than sentence processing. Thus, participants' performance in finding and formulating an appropriate interpretation might be misleading when judging their processing behaviour.

Problem 2: Irreversibility should mean that native speakers could not find the reversed version sensible in any context. Therefore, the inability to formulate a suitable interpretation or to find a sense of a reversed metaphor does not necessarily mean that in an appropriate context, participants could not understand the reversed metaphor.

Problem 3: Although the fillers made it less likely that participants discovered the aim of the experiment, one cannot rule out that they made use of strategic considerations, and, for example, rejected reversed versions of conventional metaphors quickly because they perceived them as strange or unnatural, and did not seek possible contexts in which they could be meaningful. As a consequence, it is questionable whether the experiment is capable of eliciting peoples' natural linguistic behaviour.

Problem 4: The same people coded the original order sentences and classified the reversed versions. As the authors also remark, "the judges could not be blind to experimental condition" (Glucksberg et al. 1997: 55).

Problem 5: The analysis and coding of the paraphrases have not been made public, although this would be vital in the evaluation of the experiment.

Problem 6: A further concern pertains to the statistical analysis of the perceptual data, because the experimental report does not contain the whole set of the experimental data, and there seem to be errors in the values provided.

Problem 7: It is debatable whether the results are capable of differentiating among rival approaches to metaphor processing. For instance, Glucksberg's Attributive Categorisation View and Gentner's Structure Mapping Theory both assign different roles to the topic and vehicle; therefore, both of them seem to be in harmony with the results and the research hypothesis.

4.2.2. Re-evaluation of the limit-candidate by Chiappe, Kennedy & Smykowsky (2003)

NR$_1$ (cf. Section 2.2.2): The progressivity of this non-exact replication is due to three factors: the extension of the set of metaphors in the stimulus material, suggesting a more elaborated research hypothesis (and providing a partial solution to Problem 7), as well as the solution of Problem 4 by applying independent scorers blind to the aim and structure of the experiment. Problems 1, 2, 3, and 5, in contrast, remained open, and also new problems emerged:

Problem 8: The reduction of the stimulus material to idiomatic expressions is a potential error source, because the aim of the experiment is less masked.

Problem 9: The number of items in a task sheet was very high. This might have led to boredom effects or to the use of conscious strategic considerations.

Problem 10: NR$_1$ seems to make use of rather novel metaphors, while OE contained both conventional and novel metaphors. Thus, the role of conventionality is not reflected upon.

4.2.3. Re-evaluation of the limit-candidate by Campbell & Katz (2006)

NR$_2$ (cf. Section 2.2.3): This non-exact replication is clearly progressive, because at several points the experimental design was re-thought and modifications were made, such as the addition of the contextually embedded versions and the refinement of the coding system. Thus, NR$_2$ provides at least a partial solution to Problems 2, 3, 4, 5, 8, 9 and 10. The author's attempt to resolve Problems 2 and 3, however, has also lead to a new problem:

Problem 11: The significant differences between the with-context and without context conditions question the usability of the latter, and prompt a clarification of the role of the context.

COU: Experiment 2 is a counter-experiment to OE because it is intended to provide evidence against the thesis of the irreversibility of metaphors by applying a similar stimulus material (i.e., an extended set) but using a different method. It offers a solution to a wide range of problems pertaining to OE. Thus, due to the application of a different method, Problems 1-5 did not emerge in this case, but two new problems came up:

Problem 12: Since participants had to press a button after reading each word, this might have distorted their normal reading habits.

Problem 13: The negative outcome of the experiment (no reliable differences were found) motivates a control experiment in order to check whether this method is sensitive enough to detect relevant differences.

4.2.4. Comparison of the problem solving processes

Table 3 shows the emergence and solution of problems in this experimental complex.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
OE	**E**	**E**	**E**	**E**	**E**	**E**	**E**						
NR_1	**O**	**O**	**O**	**S**	**O**		**P**	**E**	**E**	**E**			
NR_2	O	**P**	**P**	**S**	**P**		**O**	**S**		**S**	**E**		
COU	**S**	**S**	**S**				**O**					**E**	**E**

Table 3. Overview of the re-evaluation of the experimental complex evolving from Glucksberg et al. (1997)

Since the original experiment, as well as its non-exact replications are burdened with several problems, none of them can be regarded as a limit of this experimental complex. Thus, forcing a decision would be untimely. A further important upshot of our analyses is that the number and variety of problems related to the four experiments make the continuation of this line of research quite prospective and reasonable. The elaboration of newer versions seems to be possible, and more refined designs give good grounds for expecting more plausible experimental data.

4.2.5. Determination of the direction of the continuation of the cyclic process of re-evaluation

Since NR_2 proved to be the most refined version of the original experiment, the most promising decision might be to improve it further, i.e., to use the Contrastive Strategy. The following points should receive special emphasis during the elaboration of a new non-exact replication:

- Conventionality should be taken into consideration as a potentially relevant factor during the elaboration of the experimental design.
- The task should be formulated in such a way that the difference between those metaphors which are strange or unfamiliar but conceivable in special contexts, and those that are incomprehensible in every situation, is made clear. By the same token, context-free and contextually embedded versions should be applied as well.
- Adding an online version of the experiment (similar to COU) and relying on the results of a pair of different experiments seems to be well-motivated.
- Predictions should be formulated in such a way that they can be squarely confronted with different approaches to metaphor processing.

4.3. The experimental complex evolving from Bowdle & Gentner (2005)

4.3.1. The structure of the experimental complex

This experimental complex involves an original experiment and a related control experiment, as well as two non-exact replications of the original experiment and a non-exact replication of the control experiment. Experiment 3 by Jones and Estes is not included because it belongs to another experimental complex. That is, it is neither the non-exact replication of NR_2 or CON_2, nor a counter- or control experiment to any of the experiments. See Figure 4.

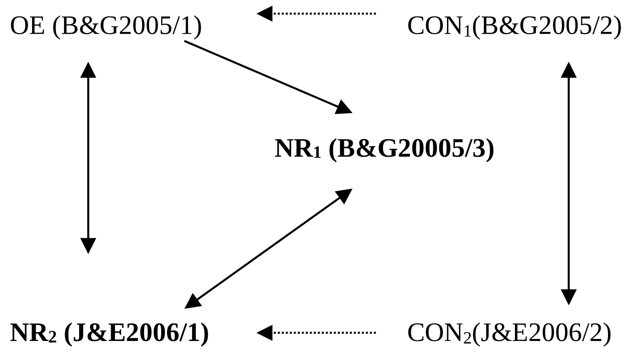


Figure 4. The structure of the experimental complex evolving from Bowdle & Gentner (2005)

In this case, there are two limit-candidates: NR$_1$ and NR$_2$.

4.3.2. Re-evaluation of the limit-candidate by Bowdle & Gentner (2005)

OE (cf. Section 2.3.1): Regarding the *identification of the problematic points of the original experiment*, the following list can be compiled:

Problem 1: The number of participants was very low, since there were only 16 subjects.
Problem 2: In the pre-tests, a small group of subjects had to evaluate the conventionality and abstractness of a huge number of items, i.e. 100 figurative statements.
Problem 3: The high number of items and the invariance in the task might have led to unnatural linguistic behaviour and the use of conscious strategies.
Problem 4: Neither the stimulus material nor the results of the pre-tests can be found in the experimental report.
Problem 5: Grammatical form preferences do not necessarily mirror processing differences. It might be the case that conventional figuratives are preferred as metaphors, due to the higher frequency and familiarity of these forms.
Problem 6: There was an unpredicted interaction between conventionality and concreteness. Thus, the research hypothesis and the predictions seem to be incomplete because they leave the role of concreteness/abstractness unclarified.

CON$_1$ (cf. Section 2.3.1): Experiment 2 in Bowdle & Gentner (2005) is a control experiment. Although it provides a solution to Problem 5 by the application of an online method, it also raises new problems:

Problem 7: Since participants knew they had to provide an interpretation, response times might have been not pure comprehension times but might have been lengthened if a participant had already tried to formulate an interpretation. Therefore, the ease of formulation of an interpretation might have influenced the comprehension times.
Problem 8: The role of aptness, as raised, for example, in Chiappe & Kennedy (2003) and Jones & Estes (2005), was only investigated in a (very thorough) post-hoc test.

NR$_1$ (cf. Section 2.3.1): Experiment 3 in Bowdle & Gentner (2005) is a non-exact replication of Experiment 1. Its progressivity is due to the involvement of further elements of the theory into the tested hypothesis and experimental design. Problem 1 was solved by recruiting a higher number of participants, but Problems 2-5 emerge in this case, again. There were two further problems, as well:

Problem 9: The key point with this experiment is, whether there is a strong enough analogy between this “in vitro” conventionalisation and “real” conventionalisation. It might be the case that the task in the first phase of the experiment utilizes short time memory and the resulting data provide information about this rather than about the mental representation of language.

Problem 10: Problem 3 has become more serious due to the high number of items both in the study phase (32 triads) and in the test phase (48 figuratives).

4.3.3. Re-evaluation of the limit-candidate by Jones & Estes (2006)

NR$_2$ (cf. Section 2.3.2): Experiment 1 by Jones & Estes (2006) is a progressive non-exact replication of OE due to the addition of a new, potentially relevant factor (aptness) to the tested hypothesis, as well as the solution of Problems 1, 4 and 8, and a partial solution to Problem 2. Two new problems have arisen:

Problem 11: Although there was a significant difference between the ratings of the conventional and novel vehicles (M = 5.14 vs. M = 3.42) in the pre-test, and similarly, the high-apt items were scored as significantly more apt than low-apt items (M = 4.85 vs. M = 3.09), the choice of the stimulus material can be questioned. Namely, the conventionality ratings made up a continuous set of numbers, which means that several experimental sentences had average conventionality. This could have been avoided if the authors had chosen metaphors with ratings from the highest third and the lowest third of the values. The aptness ratings raise a similar problem: as the list in the Appendix of Jones & Estes (2006) reveals, there were pairs which were not high-low dyads, but rather low-low (2.76-1.90, 2.64-1.79) or high-high pairs (6.48-5.69, 5.52-4.79).

Problem 12: Metaphor form preference was 3.57 and 3.27 for high apt items and for low apt items, respectively. Both values are rather inconclusive, being close to the scalar midpoint.

Problem 13: In only two cases were the results significant in the participant analysis.

Problem 14: There was a marginally significant interaction between aptness and conventionality (but only by participants, again).

CON$_2$ (cf. Section 2.3.2): Experiment 2 was a control experiment for NE$_2$, and, at the same time, a non-exact replication of CON$_1$ by Bowdle & Gentner (2005). Its progressivity is mainly due to the same factors as was the case with NR$_2$. Nonetheless, it also inherited problems from CON$_1$ and NR$_2$, and a new problem emerged, too:

Problem 15: A main effect of conventionality was found, although it was significant only in the participant analysis. More specifically, conventional similes were comprehended quicker than novel similes. This result provides weak partial support to Gentner’s theory.

Table 4 visualises the problem solving process in this experimental complex.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15

OE	**E**	**E**	**E**	**E**	**E**	**E**									
CON$_1$	**O**		**O**	**O**	**S**		**E**	**E**							
NR$_1$	**S**	**O**	**O**	**O**	**O**				**E**	**E**					
NR$_2$	S	**P**	O	**S**	O			**S**			**E**	**E**	**E**	**E**	
CON$_2$	S	**P**	O	S	S		**O**	**S**			**O**				**E**

Table 4. Overview of the re-evaluation of the experimental complex evolving from Bowdle & Gentner (2005)

4.3.4. Comparison of the problem solving processes

As Table 4's visualisation of the upshot of the re-evaluation process we conducted in Subsections 4.3.1-3 shows, both limit-candidates are burdened with problems. Thus, although the pair NR$_2$ and CON$_2$ provides more plausible experimental data, similarly to the other two experimental complexes, no well-founded decision can be made regarding the conflict between the two series of experiments. This motivates again the extension of this experimental complex with more refined versions of the experiments and control experiments.

4.3.5. Determination of the direction of the continuation of the cyclic process of re-evaluation

Despite the ineffectiveness of the problem solving process, the two chains of experiments provide us valuable starting points for starting a new cycle of non-exact replications. Namely, Problems 6, 8, 14 and 15 motivate the application of the Combinative Strategy in order to reveal the role of the three potentially relevant factors: concreteness, conventionality, and aptness.

5. Summary

In Section 1, we raised the following paradox in relation to psycholinguistic experiments:

(P) Replications are
  (a) *effective tools of problem solving* because they lead to more plausible experimental results; and they are also
  (b) *ineffective tools of problem solving* because they trigger cumulative contradictions among different replications of an experiment.

On the basis of our considerations in Section 3, as well as the case studies presented in Section 4, the following solution to (P) presents itself:
(S) Non-exact replications
  (a) are *effective tools of problem solving* if
    – they are progressive,
    – a limit of the experimental complex can be reached (temporarily, and relative to the informational state), and
    – the experimental complex has only one limit, or conflicts with other limits can be resolved;
  (b) are *ineffective tools of problem solving* if
    – they are not progressive, or

   - the chain of non-exact replications is not capable of reaching a limit of the experimental complex, or
   - conflicts among different limits of the same experimental complex cannot be resolved;

(c) *provide us valuable starting points for the elaboration of new, more refined non-exact replications* which might lead to a limit of the experimental complex;

(d) guide the choice and application of the *problem solving strategies*.

As (S) shows, while progressivity is a *local characteristic* of non-exact replications, effectiveness is a *global feature*. This means that progressivity is relative to an experiment and its non-exact replication, while effectiveness can be judged only relative to an experimental complex. Nonetheless, there are still two caveats. First, new pieces of information can overrule earlier decisions. Thus, a non-exact replication can turn out to be problematic and lose its limit-status. From this it follows that effectiveness can be judged only in the long run, and decisions are not final but only provisional. Second, a limit of an experimental complex may be inconsistent with a limit of another experimental complex. Therefore, besides intra-complex relations, inter-complex relations have to be reconstructed and evaluated.

To sum up, the proposed model supposes that experiments and experimental complexes alike are *open processes* in the sense that, in possession of new pieces of information, they may be continued, modified, or even discarded. Therefore, there are no experiments whose results were unquestionable (both practically and theoretically), nor immune to any improvement, refinement, or criticism. A second key feature of the proposed model is that experimental complexes are supposed to be *not linear but cyclic*. This means that a given step of the re-evaluation process does not necessarily lead to better results.[23] It may turn back to earlier stages and continue the revisions with an experiment for which a non-exact replication has already been conducted. Thirdly, conflicts among experiments cannot be resolved in a simple way, for example, by a mechanical comparison of the plausibility of their results. Instead, strategies of inconsistency resolution as described in Section 3.2 have to be applied.

Acknowledgements

Literature

Arabatzis, T. (2008): Experiment. In: Psillos, S., Curd, M. (Eds.), *The Routledge companion to philosophy of science*. Routledge, London & New York, 159-170.

Bowdle, B.F. & Gentner, D. (2005): The career of metaphor. *Psychological Review* 112 (1), 193-216.

Campbell, J.D. & Katz, A.N. (2006): On reversing the topics and vehicles of metaphors. *Metaphor and Symbol* 21(1), 1-22.

Chiappe, D., Kennedy, J.M. & Smykowski, T. (2003): Reversibility, aptness, and the conventionality of metaphors and similes. *Metaphor and Symbol* 18(2), 85-105.

Collins, H.M. (1985): *Changing order: Replication and induction in scientific practice*. Sage, Beverly Hills & London.

[23] This motivated the distinction between the *progressivity* and *effectiveness* of non-exact replications in Section 3.1.

Franklin, A. (2002): *Selectivity and discord. Two problems of experiment*. University of Pittsburgh Press, Pittsburgh.

Gentner, D., & Boronat, C. B. (1992). *Metaphor as mapping*. Paper presented at the Workshop on Metaphor, Tel Aviv, 1992.

Glucksberg, S., McGlone, M.S. & Manfredi, D. (1997): Property attribution in metaphor comprehension. *Journal of Memory and Language* 36, 50-67.

Haberlandt, K. (1994): Methods in reading research. In: Gernsbacher, M.A. (ed.): *Handbook of psycholinguistics*. Madison, Wisconsin: Academic Press, 1-31.

Hasson, U. & Giora, R. (2007): Experimental methods for studying the mental representation of language. In: Gonzalez-Marquez, M., Mittelberg, I., Coulson, S. & Spivey, M. J. (eds.): *Methods in Cognitive Linguistics*. Benjamins, 304-324.

Jones, L.L. & Estes, Z. (2005): Metaphor comprehension as attributive categorization. *Journal of Memory and Language* 53, 110-124.

Jones, L.L. & Estes, Z. (2006): Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language* 55, 18-32.

Kaiser, E. (2013): Experimental paradigms in psycholinguistics. In: Podesva, R.J. & Sharma, D. (eds.): *Research Methods in Linguistics*. Cambridge: Cambridge University Press, 135-168.

Keenan, J.M., Potts, G.R., Golding, J.M. & Jennings, T.M. (1990): Which elaborative inferences are drawn during reading? A question of methodologies. In: Balota, D.A., Flores d' Archais, G.B. & Rayner, K. (eds.): *Comprehension processes in reading*. Hillsdale: Erlbaum, 377-402.

Kertész, A. & Rákosi, Cs. (2012): *Data and Evidence in Linguistics: A Plausible Argumentation Model*. Cambridge: Cambridge University Press.

Kertész, A. & Rákosi, Cs. (2014): The p-model of data and evidence in linguistics. In: Kertész, A. & Rákosi, Cs. (eds.): *The Evidential Basis of Linguistic Argumentation.* Amsterdam & Philadelphia: John Benjamins, 15-48.

Keysar, B., Shen, Y., Glucksberg, S., Horton, W.S., 2000. Conventional language: How metaphorical is it? *Journal of Memory and Language* 43, 576-593.

Lakoff, G. & Johnson, M. (1980): *Metaphors we live by*. Chicago: Chicago University Press.

Machamer, P. (2002): A brief historical introduction to the philosophy of science. In: Machamer, P. & Silberstein, M. (eds.): *The Blackwell guide to the philosophy of science*. Blackwell, Malden & Oxford, 1-17.

Pickering, A. (1981): The hunting of the quark. *Isis* 72, 216-236.

Rákosi, Cs. (2012): The fabulous engine: strengths and flaws of psycholinguistic experiments. *Language Sciences* 34, 682-701.

Rákosi, Cs. (2014): On the rhetoricity of psycholinguistic experiments. *Argumentum* 10, 533-547. http://argumentum.unideb.hu/2014-anyagok/angol_kotet/rakosicsi.pdf.

Thibodeau, P. & Durgin, F.H. (2008): Productive figurative communication: Conventional metaphors facilitate the comprehension of related novel metaphors. *Journal of Memory and Language* 58, 521-540.