Csilla Rákosi: On the evaluation of psycholinguistic experiments on metaphor. Part II
Argumentum 12 (2016), 288-302
Debreceni Egyetemi Kiadó

Tanulmány

Csilla Rákosi

On the evaluation of psycholinguistic experiments on metaphor

Part II: Case studies

Abstract

Psycholinguistic research into metaphor is characterised by contradictory and often controversial experimental results. Thus, the question is, how to decide when an experiment yields plausible experimental data and when it is unreliable as a data source. On the basis of a model of psycholinguistic experiments, it is proposed that experiments should be viewed as cyclic open processes. This means that the plausibility of the statements related to different stages of the experimental process is revised again and again during the elaboration and conduct of the experiment, as well as during its evaluation. Accordingly, the analysis and evaluation of experiments is nothing else than the continuation of the experimental process by new plausible argumentation cycles, and, if possible, the elaboration of proposals for its resumption by new experimental cycles.
Keywords: psycholinguistic experiments, experiments on metaphor processing, philosophy of science, evaluation of experiments

4 Analysis of psycholinguistic experiments on metaphor processing

In the present Section, several psycholinguistic experiments will be analysed and evaluated with the help of the system of criteria presented in Part I. The aim of the analyses is not to provide a comprehensive overview of the current stand or the history of psycholinguistic experiments on metaphor processing but to illustrate the metatheoretical views advanced in Part I of this paper. Further, the analyses do not intend to be complete. Rather, they will focus on certain aspects of experiments whose closer examination seems to be especially instructive. First, we will present a short description of the experiment. Then, we will highlight some problematic points which seem to be illuminating by making use of the guidelines for the evaluation of experiments. As a third step, it has to be decided whether further developments are possible which make it possible to avoid the systematic errors revealed and/or to increase the reliability of the experiment at issue. Section 5 will summarise the results.

It is important to emphasise that the aim of these analyses is not denunciation of a research field or the researchers working in it. Instead, they are intended to exemplify the work to be done in this field of research: *rigorous and in-depth analyses and strict and determined revisions whenever there is a possible systematic error* – even though the experiment at issue was regarded as a well-founded and reliable one for decades. In many cases, the outcome could be a refined, more elaborated new version.

4.1 Keysar (1989)

Experiment 1

Description: After reading a short story, participants had to decide whether a sentence that could be interpreted both literally and metaphorically was literally true. The stories were constructed in such a way that their first part rendered the target sentence literally true (L+) or literally false (L–), while their second part rendered the target sentence metaphorically true (M+) or metaphorically false (M–).[1] According to the author's hypothesis, if metaphorical interpretation is constructed – in contrast to the traditional view on metaphors, but in harmony with Gluckberg et al's view – in an obligatory, involuntary manner, then decision in incongruent contexts (L+/M– or L–/M+) should take longer than in congruent ones. Therefore, participants' decision times were captured and compared across the different story types.

Evaluation: The stimulus material raises two concerns. First, the structure of the stories presented was unvaried: literal part – metaphorical part – target sentence, which in all cases also had a metaphorical meaning. Although filler items were also used, there were 13 practice items to complete before the experimental ones. Thus, participants might have been able to identify the structure of the items, and make use of strategic considerations instead of providing instinctive answers.[2] The second possible error source was the wording of the task. Participants were instructed to determine whether the target sentence "is literally true or strongly implied (as such) given the preceding paragraph". Since after some practice items, participants might have easily found out that the final sentences always had a metaphorical meaning, this might have resulted in a transformation of the original task to another one, requiring participants to decide whether the sentence at issue is true, and if so, whether it is (also) literally true and not (only) metaphorically. Thus, the experiment seems not to be capable of eliciting participants' natural linguistic behaviour precisely at the decisive point, since it might have been the instruction itself that triggered the metaphoric interpretation and not the stimuli.

Proposals: Since it is doubtful that the formulation of the instructions can be altered in such a way that does not lead to the problems described above, the elaboration of an improved version seems to be unworkable.

Experiment 2

Description: Participants read the context stories of Experiment 1 on the screen and they had to push a key after having read a line. There were also some quiz questions in order to engage participants' attention. Keysar put forward the prediction that if the standard model of metaphor comprehension is correct and metaphorical interpretations are only generated when the literal interpretation fails, then literally false but metaphorically true sentences should be read more slowly than literally true sentences; secondly, reading times after L+/M+ contexts should not be faster than after L+/M– contexts. In contrast, according to Keysar's rival hy-

---

[1] For example, a L+/M– story was the following one:
"Bob Jones is an expert at such stunts as sawing a woman in half and pulling rabbits out of hats. He earns his living travelling around the world with an expensive entourage of equipment and assistants. Although Bob tries to budget carefully, it seems to him that money just disappears into thin air. With such huge audiences, why doesn't he ever break even?
Target sentence: Bob Jones is a magician."

[2] Indeed, as Keysar writes, there were subjects who "suspected the goal of the experiment".

pothesis stating the automaticity of metaphorical interpretation, L+/M+ reading times should be the fastest because two interpretations are available in such cases; further, L+ should be faster than L–, and M+ should be faster than M–.

Evaluation: The stimulus material is missing in the experimental report, therefore it cannot be analysed. The problem of the unvaried item structure emerges here, too, but to a somewhat reduced extent due to the quiz questions, which might have made the aim and the structure of the experiment less transparent. Despite this, there were subjects again who seemed to have realized the object of the experiment.

Proposals: The experimental design is in need of revision in order to ensure that participants cannot find out the aim of the experiment. For this reason, the order of the literal and metaphorical contexts should vary; purely metaphorical and purely literal stories could also be included; not all final sentences should have a metaphorical meaning, and fillers should be applied so that the items' structures do not follow the same pattern.

Summing-up: Keysar's paper contains two closely related experiments with the help of which the difference between improvable and non-improvable experiments can be exemplified. As we have seen, in both cases, it is the experimental design that is burdened with problems, but the errors revealed only seem to be fatal with the first experiment in the sense that the experiment is not capable of providing plausible experimental data and it cannot be improved. With the second experiment, it is possible to elaborate and conduct a revised version.

4.2 Nayak & Gibbs (1990)

Experiment 5

Description: This experiment was intended to test the research hypothesis that "people consciously recognize the conceptual metaphors that underlie the meanings of idioms" (Nayak & Gibbs 1990: 324). Participants were given the task of matching idioms with their connected conceptual metaphors from an eight-member list. There were 2 pairs of conceptual metaphors related to anger, fear, success and failure; thus, each idiom had to be matched with one item from a closed, 8-member list.

Evaluation: The first problem is that this experiment might be suitable for checking the first part of the research hypothesis (i.e., that people's decisions are in harmony with the predictions of CMT) but only with reservations can the same be said for its second part (that these alleged conceptual metaphors in fact underlie the mental representation of the idioms). Therefore, interpretation of the experimental data is defective insofar as the analysed data are not directly related to mental representations of the participants, but pertain to their conscious behaviour. The second problem relates to the experimental design. Namely, a decision that, for example, the idiom *jump down your throat* has to be matched with the "conceptual metaphor" ANGER IS LIKE A FEROCIOUS ANIMAL from the list ANGER IS LIKE PRESSURE IN A HOT CONTAINER, ANGER IS LIKE A FEROCIOUS ANIMAL, SUCCESS IS UP, SUCCESS IS LIKE A COMPLETED JOURNEY, FAILURE IS LIKE AN INCOMPLETE JOURNEY, FAILURE IS LIKE CONSUMING SOMETHING INEDIBLE, FEAR IS LIKE ESCAPE, FEAR IS A PHYSICAL CHANGE is uninformative about the interpretation or processing of metaphors, since there is always one item from the list that is semantically related to the given idiom. The authors' caveat that "The instructions

emphasized that subjects were not to make their judgements on the basis of the literal similarity between the words in the idioms and the linguistic descriptions of the conceptual metaphors" (Nayak & Gibbs 1990: 324) cannot be regarded as an appropriate control for this problem. As a consequence, this experiment is not suitable for the investigation of people's conscious interpretation of metaphorical expressions, either. The stimulus material is missing in the experimental report.

Proposals: Due to the problems mentioned (offline method, choosing from a closed list, semantic relatedness), the experiment's basic idea is inherently flawed.

Experiment 6

Description: Participants were presented with scenarios containing idioms which were supposed to belong to the same "conceptual metaphor" (metaphorical mapping), and had to decide which of two idioms is more appropriate as the final sentence of the given story. While one idiom belonged to the same mapping, the other one did not.

Evaluation: The experimental design is above all burdened with the problem that the results may be due to participants' strategic considerations based on the semantic and stylistic relatedness of the priming text and one of the target expressions. Moreover, no filler tasks were applied, thus participants could have easily realised that the experiment is about metaphors and this might have led to the use of their own naïve theories about this topic. Third, appropriateness ratings do not necessarily reflect processing difficulty. Finally, the stimulus material is missing in the experimental report.

Proposals: The problems are similar to those with the previous experiment; no improved version seems to be available.

Summing-up: The problems related to Experiments 5 and 6 by Nayak and Gibbs are typical of the early stage of psycholinguistic experiments in favour of Conceptual Metaphor Theory. Namely, they fall short of construct validity and cannot be regarded as reliable data sources.[3] Moreover, in this case, the thought experiment checking the experimental design suffices to reject the idea of possible developments, too.

4.3 McGlone (1996)

Experiment 1

Description: Participants were presented with 16 metaphorical sentences (listed in the appendix). They were instructed to rate the comprehensibility of the sentences and write a paraphrase in their own words. Two additional participants coded the paraphrases in such a way that '0' meant that the paraphrase did not contain words referring to the supposed source domain according to Lakoff and Johnson's conceptual metaphor theory (CM-inconsistency), '1' indicated ambiguity in this respect, while '2' indicated that there is clear reference to the source domain (CM-consistency). In content analysis, 76% of the paraphrases received the code 0, 14% the code 1, and only 10% the code 2. From this and additional analyses of the

[3] Thus, for example, the experiments in Gibbs (1992) are burdened with similar errors.

perceptual data McGlone inferred that participants did not rely on conceptual metaphors as a knowledge source when they interpreted metaphorical expressions.
Evaluation: Similarly to other off-line methods, this experiment provides information about people's conscious behaviour instead of the spontaneous mental processes of metaphor interpretation. Therefore, this experiment may provide evidence against Nayak & Gibbs's (1990: 324) hypothesis that "people consciously recognize the conceptual metaphors that underlie the meanings of idioms" as already quoted in Section 4.2; but there is only a very weak and indirect link between the experimental data gained and the research hypothesis, which interpreted conceptual metaphors as possible knowledge sources. It is also debatable whether the elicited interpretations result from participants' normal, natural linguistic behaviour. As McGlone (1996: 552) remarks, a concern with this experiment is that participants interpreted the instructions in such a way that they should avoid idioms and provide literal paraphrases. A further problem results from the circumstance that the coding of the metaphor interpretations cannot be operationalized – although the application of two non-linguist participants and the procedure for achieving agreement on the judgement of the interpretations considerably reduce the resulting uncertainty. Nevertheless, their decisions may be controversial. Therefore, the whole set of paraphrases and their evaluations should have been presented in the experimental report.

Experiment 2

Description: Experiment 2 used a similar design and the same stimulus material; the difference was that participants had to provide paraphrases with the help of other metaphors.

Evaluation: All the weak points of Experiment 1 emerge in this case again. A further problem is that not only metaphor interpretation but also metaphor production was involved. This may lead to two kinds of issues. First, the impact of the two processes cannot be separated from each other. Second, it is doubtful that participants' natural linguistic behaviour was elicited because "some participants may have approached the task of generating metaphors as a test of creative ability. As a result, they may have felt pressure to employ an unconventional interpretation strategy to come up with novel metaphors" (McGlone 1996: 554).

Experiment 3

Description: Participants were asked to rate the similarity between the metaphors used as stimulus material in the previous experiments on the one hand, and metaphors provided as idiomatic paraphrases by the subjects of Experiment 2 (excerpts can be found in Appendix B in McGlone's paper) on a 7-point scale. With each target metaphor (such as *Dr. Moreland's lecture was a three-course meal for the mind*), 9 possible alternative metaphors were provided; 3 were Conceptual Metaphor Theory-consistent (*Dr. Moreland's lecture was a smorgasbord for the mind* – same CM source domain), 3 CMT-inconsistent but attributively similar (*Dr. Moreland's lecture was a full tank of gas for the mind* – different CM source domain), and 3 unrelated (i.e., *Dr. Moreland's lecture was a ceiling fan for the mind*). McGlone made the prediction that high similarity values with CMT-consistent metaphors would indicate that participants' ratings had been based on the underlying conceptual metaphors, while the choice of metaphors with a similar vehicle – similar in the sense that they belong to the same attributive category – would provide evidence for the Attributive Category View.

Evaluation: First, the indirectness of this experimental method is greater than it was with Experiments 1 and 2.[4] Second, the stimulus material is missing in the experimental report. Third, the use of strategic considerations was not prevented and, more importantly, cannot be ruled out. Fourth, the interpretation of the perceptual data and the confrontation of the experimental data with the predictions are deficient. Namely, no significant difference has been found between CMT-consistent and CMT-inconsistent, i.e. ACV-consistent, metaphors; therefore, both the predictions gained from the Cognitive Metaphor Theory (preference of CMT-consistent metaphors) and the predictions obtained by the Attributive Categorisation View are in conflict with the experimental data. In contrast to this, McGlone draws the consequence that the results are in conflict with the Cognitive Metaphor Theory but they are consistent with the Attributive Categorisation View.

Proposals: After correction of the interpretation and statistical analysis of the experimental data, this experiment may provide experimental data useable solely as evidence for or against hypotheses about conscious strategies of metaphor interpretation.

***Experiment** 4*

Description: A cued recall paradigm was applied. Participants had to write down any sentences heard from a tape recorder that seemed to be related to a given cue in a booklet. CMT-clues were related to the source domain of the assumed conceptual metaphor (*Lisa is the brain of the family* – SOCIAL GROUPS ARE BODIES – *body part*), while ACV-clues were related to the attributive category associated with the vehicle concept (*intelligent*). All 16 metaphorical sentences had a counterpart containing expressions related to the cues in their literal meaning. There were some fillers as well. It was investigated whether clues relating to Conceptual Metaphor Theory or clues based on Glucksberg's Attributive Categorisation View are more effective. The instructions did not contain any reference to the following recall task. Two additional participants coded the answers independently, but in a second turn, they had to come to an agreement about the evaluations.

Evaluation: The first problem is that participants heard the 16 sentences only twice; therefore, error rates were high. Secondly, it is not clear whether literal sentences provide an appropriate control in this case. Thirdly, there was semantic relatedness between CMT-cues and the sentences, but not between ACV-cues and the sentences;[5] a control experiment only checked the relationship between the CMT- and ACV-cues. A fourth problem is that the stimulus material, in contrast to Experiments 1-3, cannot be found in McGlone (1996).

Proposals: Repetition and two control experiments could increase the reliability of this experiment. Namely, the outcome of the repeated experiment should be compared with the results of an experiment that differs from Experiment 4 only insofar as no cues are applied, and with

---

[4] Cf. "[…] the reflective, deliberate nature of paraphrase and ratings tasks may not be generalizable to situations in which a metaphor is encountered in ongoing text or discourse. The knowledge base that people use to reflectively interpret and appreciate metaphors may be broader than that which is required for immediate comprehension […]." (McGlone 1996: 556)

[5] For example: The faculty meeting was a *battle* – Many men took part in the *battle* – *war* (CMT-cue) – *dispute* (ACV-cue); Lisa is the *brain* of the family – *body part* (CMT-cue) – *intelligent* (ACV-cue); The lecture was a three-course *meal* – She prepared a three-course *meal* – *food* (CMT-cue) – large quantity (ACV-cue).

the results of an experiment in which the sentences are not presented but participants are asked to write down as many metaphors as possible related to the cue words.

Summing-up: McGlone (1996) intends to provide experimental evidence against Conceptual Metaphor Theory by challenging Gibbs's results. These experiments are manifestly more elaborate insofar as they take more factors into consideration and are built on each other cyclically in order to rule out possible systematic errors. Despite this, only the last experiment can be developed into a reliable data source on metaphor processing, because the others provide evidence against CMT as a part of native speakers' conscious strategies of metaphor interpretation.

4.4 Bowdle & Gentner (1999)

Description: In order to test Gentner's career of metaphor hypothesis, the authors developed a two-stage experimental design. In the first, 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.[6] The authors' hypothesis was that priming with novel similes using the same base term makes subjects "derive an abstract schema and associate it with the base term". In this way, the authors "aimed to speed up the process of conventionalization from years to minutes" (Bowdle & Gentner 1999: 93). The material also involved similar tasks with literal comparisons. According to the career of metaphor hypothesis, there is a shift in metaphor processing insofar as novel metaphors are processed as comparisons, while conventional metaphors are processed as categorizations. Therefore, 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 in novel similes should lead to significantly higher preference numbers than figuratives with no prior exposure, but the same should not hold with items in which the prime had been seen in literal comparisons.

Evaluation: The key point with this experiment is whether and to what extent "in vitro" conventionalisation corresponds to "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 it rather than about the mental representation of language. A further problem is the high number of items, both in the study phase (32 triads) and in the test phase (48 figuratives), and the invariance in the task – these factors might have led to unnatural linguistic behaviour and the use of conscious strategies.

Proposals: This experiment makes use of an offline method, and the link between the experimental data and the theory is rather weak. Therefore, the search for alternative interpretations

---

[6] For example:
An acrobat is like a butterfly.
A figure skater is like a butterfly.
_____ is like a butterfly.

and control experiments for their elimination, as well as a repetition of the experiment, seem to be vital and could increase the plausibility of the results and their supportive force considerably.

Summing-up: Bowdle and Gentner put forward a highly original experimental design, whose evaluation, however, requires further experiments and repetitions. Therefore, this experiment should be treated rather as the starting point of a longer and promising experimental complex and not as a (single) full-fledged experiment.

4.5 Wolff & Gentner (2000)

Experiments 1-2

Description: Experiment 1 aimed to provide relevant data about the question of the asymmetry or symmetry of the initial stage of metaphor processing. The former hypothesis follows from Gluckberg's ACV, while the latter from Gentner's SMT. Participants had to decide whether the presented statements are literally true or false by pressing the left or the right arrow keys. In the 180-item list, there were four kinds of literally false statements: ordinary false (*Some birds are apples*), high directionality forward metaphors (*Some jobs are jails*), scrambled metaphors (*Some rumours are jails*), and reversed metaphors (*Some jails are jobs*). The literally true statements were either high-typicality statements (*Some birds are robins*), or low-typicality statements (*Some birds are penguins*) and they served as manipulation checks. Experiment 2 relied on a similar experimental design with two modifications. With the help of control experiments presented in Gentner & Wolff (1997), only metaphors of high-conventionality were selected, and the forward and reversed metaphors were divided into two subgroups: high-similarity and low-similarity metaphors.

Evaluation: The first concern is that the high number of items, the identical syntactic structure of the sentences, the task and the feedback after errors in both the practice and the test phases might have led to monotony and unnatural linguistic behaviour, considerably different from normal reading strategies. The second problem pertains to the experimental design, too. It is not clear what should invite the reader to seek an analogy between the two terms in the case of reversed metaphors but not with scrambled metaphors or ordinary false statements. A third weak point seems to be that error rates are not parallel with reaction times, although both should indicate difficulties in processing. Fourth, it is only supposed that the reaction times are related to the early phase of metaphor processing. As Wolff & Gentner (2000: 535) also remark, "it is conceivable that the results instead reflect late processes". If so, then there is a danger that they mirror rather conscious strategies of participants instead of their unconscious, natural linguistic behaviour. The authors claim that this concern is unfounded because "in metaphor comprehension studies, the mean RTs typically lie between 1800 and 4000 ms" (Wolff & Gentner 2000: 535). This explanation is, however, not satisfactory and provides only weak evidence, because the experiments they refer to involve more complex tasks such as providing or creating an interpretation, or giving meaningfulness ratings, and the average duration of the conduct of the different sub-processes is unknown.

Proposals: Without correction of the revealed errors, these experiments cannot be regarded as reliable data sources.

Experiment 3

Description: Participants were presented with metaphoric sentences (forward, reversed, scrambled) and they had to decide whether they are comprehensible or not. The stimulus material was selected from that of the previous experiment. There were 64 practice items and 72 test items. In this case, subjects did not receive feedback during the test session.

Evaluation: The first problem is the formulation of the instructions: participants were told that they would see either metaphorical statements or anomalous statements. This might lead to the application of conscious strategies instead of reliance on natural linguistic behaviour. The second problem is the huge number of items and the monotony of the tasks, as in the previous experiments. Third, if we take a closer look at the stimulus material, we can see that many low-similarity items can also be easily interpreted as low-constraint topics, and high-similarity metaphors are often also high-constraint metaphors.[7] Therefore, the experimental data are not capable of discriminating between predictions based on Gentner's SMT and Glucksberg's ACV.[8] Thus, the interpretation of the experimental data is debatable. A fourth issue is that decision times were used only to compare comprehensibility decisions with truth or falsity decisions in the previous experiment, but they were neglected in the comparison and analyses of the different conditions in this experiment.

Proposals: In this case, more thorough revisions are needed with the experimental design, the stimulus material and the interpretation of the experimental data.

Summing-up: The most alarming problem with Wolff and Gentner (2000) is boredom effects due to the huge number of similar tasks, which might have influenced participants' performance. The confrontation of the experimental data and rival theories is in need of refinement, too. Nevertheless, these experiments are clearly improvable, that is, new experimental cycles can be initiated which may produce plausible experimental data.

4.6 Gernsbacher, Keysar, Robertson & Werner (2001)

Experiment 1

Description: The authors intended to test the hypothesis that the basic-level meaning of the vehicle is suppressed during metaphor comprehension. Half of the prime sentences were metaphorical (*That defense lawyer is a shark*), the other half involved their literal counterparts in the sense that the metaphor topic was changed for a member of the basic-level category represented by the vehicle (*That large hammerhead is a shark*). There were two kinds of target sentences: half of them were property statements related to the metaphorical subordinate category (*Sharks are tenacious*), while the other half were related to the literal basic-level category (*Sharks are good swimmers*). Participants had to decide whether the sentences presented made sense or not. If Glucksberg's ACV and the above hypothesis hold, then participants should verify superordinate-level property statements more rapidly after metaphor-prime sen-

[7] For example:
Some arguments are wars. vs. *Some conversations are wars.*
Some lies are boomerangs. vs. *Some statements are boomerangs.*
Some saunas are ovens. vs. *Some rooms are ovens.*
Some suburbs are parasites. vs. *Some towns are parasites.*

[8] Cf. Section 4.8.

tences than after literal-prime sentences, and they should verify basic-level property statements more slowly after metaphorical than literal primes. The whole stimulus material can be found on the first author's homepage.

Evaluation: First, besides the 48 experimental sentence pairs, there were also 144 filler pairs which had a similar structure but at least one member of the statements did not make sense. Despite this precaution, the huge number of tasks of the same structure might have led to monotony and mechanical decision-making following certain conscious considerations, or to the development of conscious strategies. Although only the results of participants with a performance under 66% were excluded during the authentication of the perceptual data, data from 16% of participants had to be eliminated. Secondly, a related problem was that the instructions not only explicitly mentioned metaphors, but a short explanation was also provided, where metaphors were described as a kind of analogy or simile.[9] The explanation may have tempted subjects to interpret their task in such a way that they have to deal with correct or defective analogies on the one hand, and class-member statements, on the other. Against this background, it is doubtful whether this experiment was capable of investigating people's natural linguistic behaviour. A third possible problematic point was identified by the authors: namely, the longer verification time of basic-level properties after metaphorical primes should be rather interpreted as a faster verification time after literal primes, because they contain basic-level terms (such as *hammerhead*). Fourth, the experimental data are – in contrast to the authors' view – not capable of discriminating between predictions based on, for instance, Gentner's SMT and Glucksberg's ACV. For example, if metaphor processing involves structural alignment between the topic and the vehicle as supposed by Gentner, then mentioning a property of the vehicle which cannot be placed in relation to the topic leads to incoherence, while this is not the case with the corresponding literal sentence.

Experiment 2

Description: In order to eliminate the third problem above, in Experiment 2 the same experimental metaphor primes were used but their literal counterparts were changed for nonsense-primes such as *His English notebook is a shark*. The authors put forward the prediction that the verification of basic-level property statements should be slower after metaphorical primes than after nonsense primes.

Evaluation: The same problems, excluding Problem 3, emerge in this case again. For instance, the answers of 27% of participants had to be eliminated. The authors expressed the concern that the advantage of nonsensical primes might be due the circumstance that they contain the vehicle term in its literal, basic meaning and enhance its basic level properties.

Experiment 3

Description: Instead of nonsense-primes, unrelated metaphors (*That new student is a clown*), which did not include the prime vehicle, were used in order to overcome the last problem relating to Experiment 2. This modification, however, leads to another problem: namely, that

[9] "In this experiment, many of the sentences are metaphorical. A metaphor is a figure of speech in which a word suggests a likeness or analogy between two things." (http://www.gernsbacherlab.org/research/language-comprehension-research/experimental-stimuli/experiment-1-materials-literal/)

while the metaphorical prime and the basic-level target both contain the vehicle term (*shark*), the same does not hold for the unrelated prime sentence. Therefore, lexical priming may influence the reaction times. In fact, in this case, basic-level relevant targets were, contrary to the previous experiments, significantly shorter after metaphors than after unrelated sentences. The authors tried to eliminate this distorting effect by subtracting a "penalty" from the average reaction times of unrelated sentences, or with the help of statistical means, namely, with computed z-scores for each prime type.

Evaluation: Problems 1, 2 and 4 mentioned in relation to Experiment 1 can be raised in this case again; thus, for instance, 30% of the perceptual data had to be rejected due to too high error rates. The statistical analysis is questionable, too. First, it is impossible to determine the exact value of the "penalty" for unrelated sentences, and different values lead to totally different constellations. Second, the method described of transforming the results into z-scores seems to be problematical, too. Above all, it is not clear what the comparison between the calculated z-scores of the basic-level relevant targets of metaphorical and unrelated primes in Experiment 3 might mean. The former indicates the value of the basic-relevant target verification times expressed in standard deviation units – relative to the standard deviation of verification times pertaining to the *metaphorical primes* (= 225.77 ms). The latter, however, shows the value of the basic-relevant target verification times expressed in standard deviation units – against the standard deviation of the *unrelated primes* (= 264.49 ms). We would obtain a different scenario if we used the mean and standard deviation of *all observations* gained in Experiment 3 for calculating the z-scores, because this transformation would not change the relationship between the results. To sum up, making use of penalties or standardisation instead of re-designing the experiment does not seem to be a viable option.

Proposals: It is not clear how the revealed errors could be corrected; thus, no improved versions seem to be available.

Summing-up: In addition to problems similar to those found in Wolff & Gentner (2000), the extremely high error rates and shortcomings in the statistical analysis of the perceptual data make the experiments as sources unreliable; that is, the experimental data gained cannot be regarded as plausible.

4.7 Gibbs, Lima & Francozo (2004)

Description: American and Brazilian participants, respectively, received a list of expressions closely related, possibly related, or unrelated to symptoms of hunger. The expressions belonged to three types: local symptoms referred to body parts (*one has a stomach ache*), general symptoms referred to the whole body (*become dizzy*), while behavioural symptoms referred to behaviours that may be consequences of being hungry (*become depressed*). Subjects had to rate each item on a 7-point scale "as to whether they had experienced the effect mentioned when feeling hungry". In the second part of the experiment, another group of participants was first asked to rate the relevance of a list of expressions possibly related to the feelings of a person who is in love, who lusts after somebody or something, or who has a desire ("body questions") on a 7-point scale. The same participants also filled in a questionnaire about a list of linguistic expressions and evaluated their acceptability when talking about love, lust and other types of desire, respectively ("linguistic questions").

Evaluation: This experiment collects and analyses people's conscious reflections on symptoms that cannot be equated – contrary to the authors' supposition – with (more) direct investigation of their bodily sensations, mental representations or conceptual backgrounds. Thus, the experiment does not touch upon metaphor processing but investigates, as the authors correctly put it, "people's folk knowledge about hunger" and desire, and their conscious judgement of linguistic expressions. Second, it is highly problematic that the same group of subjects provided ratings to the "body questions" and to the "linguistic questions". This step allows interferences in the answers to the two kinds of questions. Thus, one cannot rule out that participants' ratings on the "body questions" were influenced by their linguistic knowledge, or by their implicit theories about the meaning of the relevant metaphorical expressions based partially on stereotypes offered by idioms. Third, there is an important difference between the wordings of the tasks, which might have influenced participants' answers. Namely, while questions related to the symptoms of hunger pertained to the experiences participants had, the "body questions" required participants to imagine the feelings of somebody being in love, and the "linguistic questions" asked them to decide "whether it was an acceptable way of talking in their respective language". Fourth, the alleged correlation between data sets is not supported by calculations. The experimental data rather suggest that both the strongly and the weakly relevant hunger symptoms are only moderately or weakly relevant in relation to body symptoms of desire as well as in respect to linguistic expressions about desire. Fifth, data relating to "moderately related" symptoms had been omitted from the analyses.[10] Sixth, the statistical analysis of the perceptual data is defective. No proper analyses are provided, and the partial analyses infringe the rules of the use of statistical tools.[11] Since there were three kinds of desire analysed in this experiment, the last statement means that two-thirds of the English data (and one-third of the whole data set) was statistically not significant.

Proposals: Since the basic idea of the experiment is inherently flawed, this experimental design cannot be improved.

Summing-up: This experiment overcomes several problems typically related to earlier experiments in favour of CMT. It goes beyond the analysis of purely linguistic manifestations and intends to tap into "embodied experiences", by making use of a widened database. Despite this, it is again people's conscious reflections which are studied, and the statistical analyses are clearly deficient. Therefore, this experiment cannot be turned into a reliable source which could provide plausible experimental data about metaphor processing.

5 Summary and prospects

In the previous section, we have seen how the application of the criteria proposed in Part I of this paper can be used in the *re-evaluation of the plausibility* of statements related to different components of the experimental procedure, and via this, in the *revision of the components* themselves. It has also become clear that the weight and impact of errors can be judged only

[10] Cf. Gibbs et al. (2004: 1204).

[11] Cf. "The findings for both the Body and Linguistic questions are *generally consistent* across English and Portuguese for the three types of symptoms for the three types of desire (love, lust, other). Each difference between the strong and weak items for each type of desire was statistically significant, with the exception of love and other desire for English speakers which were only *marginally different*." (Gibbs et al. 2004: 1206)

in the context of the given experimental report, that is, in relation to the argumentation process at issue, by taking into consideration all details of the experiments at our disposal.

To sum up, the précis of our analyses is that *psycholinguistic experiments on metaphors should be turned into a much more thorough and effective cyclic re-evaluation process*. The main points to be considered when moving in this direction should be the following:

- Experiments should be, in harmony with requirements relating to scientific experiments in general, *repeatable* and *actually repeated*. With this end in view, the whole stimulus material, the whole set of the perceptual data and all important details of the statistical methods applied should be made public, for example, on the author's homepage, or on a homepage which could be devoted to psycholinguistic experiments with a kind of data bank of all experiments conducted so far. Experiments should not be regarded as reliable data sources till they are repeated and the replication reinforces their results.
- As our analyses have shown, the Achilles heel of many psycholinguistic experiments is their stimulus material. This provides a further argument for the requirement that the whole stimulus material should be available so that the evaluation of the experiment may also involve a *thought experiment*. Namely, readers should be in a position to become *virtual participants* in the experiment. In this way, it can be more effectively checked whether real participants might have, for example, made use of strategic considerations.
- A related point is that the choice of participants should be controlled for. Thus, linguists, and students of linguistics or psychology should be excluded from psycholinguistic experiments because they might reveal the aim of the experiment more easily.
- The whole set of perceptual data should be made public in order to make it possible to check whether the conditions of application of the chosen statistical method are fulfilled and the calculations are correct, or if possible, alternative analyses can be carried out.
- It should be made clear whether the experimental data are suitable for providing evidence about metaphor processing or pertain only to conscious judgements about the usage of metaphorical expressions.
- Semantic priming should be controlled for more effectively.
- If the repetitions lead to conflicting results, then thorough comparative analyses should be carried out.
- The relationship between the experimental data and rival theories should be made manifest. That is, it should be carefully determined which predictions from the rival theories can be drawn, and the experiment should provide data which are in harmony with the predictions of only one of these rivals.
- Although in this paper I have not touched upon the introductory sections of papers dealing with psycholinguistic experiments, it is often the case that researchers present rival theories in such a way that they strongly simplify and distort them. A similar problem is that the author's own theory and the data supporting it are presented as unquestionable facts.

These proposals cannot, of course, guarantee that psycholinguistic experiments will provide incontestable data for theories about metaphor processing. From the perspective of the p-model's view of experiments it follows that they are data sources that may provide plausible but not certainly true data. Nevertheless, the acknowledgement of the fallibility of experiments does not mean that the reliability and importance of experiments would be questioned. On the contrary: if one is aware of the strengths and possible weak points of these data sources, and as many details of the experiments are made public as possible, then one can

search consciously for errors, reveal potential error sources, revise the experimental design, and develop more refined and elaborated versions of earlier experiments or construct new kinds of experiments. Therefore, strictness and thoroughness in the analysis of experiments, the elaboration of control experiments, and trying out new designs are not destructive activities but might, on the contrary, be the key to the flourishing of this field of research, and lead to a more open and straightforward atmosphere and to more reliable data due to the *collective efforts of the whole scientific community*. This means that a radical change of view is needed. Experiments should not be viewed as single, unique acts conducted by a (small group of similarly minded) researcher(s), but *experimental complexes* should be elaborated step by step by the participation of researchers belonging to different theoretical backgrounds or even to rival approaches. Such an experimental complex would involve chains of closely related experiments: replications of the original experiment, as well as different kinds of control experiments, counter experiments and more and more refined and varied versions of the original experiment. Rákosi (manuscript) and Rákosi (in preparation) explicate the concept of 'experimental complex' and investigates the relationship between non-exact replications of experiments and the emergence and resolution of inconsistencies.

References

Bowdle, B.F. & Gentner, D. (1999): Metaphor comprehension: From comparison to categorization. In: *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society*, 90-95.

Gentner, D. & Wolff, P. (1997): Alignment in the processing of metaphor. *Journal of Memory and Language* 37, 331-355.

Gernsbacher, M.A., Keysar, B., Robertson, R.R.W. & Werner, N.K. (2001): The role of suppression and enhancement in understanding metaphors. *Journal of Memory and Language* 45, 433-450.

Gibbs, R.W., Jr. (1992): What do idioms really mean? *Journal of Memory and Language* 31, 485-506.

Gibbs, R.W. (2013): The real complexities of psycholinguistic research on metaphor. *Language Sciences* 40, 45-52.

Gibbs, R.W., Lima, P.L.C., Francozo, E. (2004): Metaphor is grounded in embodied experience. *Journal of Pragmatics* 36, 1189-1210.

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.

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.

Keysar, B. (1989): On the functional equivalence of literal and metaphorical interpretations in discourse. *Journal of Memory and Language* 28, 375-385.

McGlone, M.S. (1996): Conceptual metaphors and figurative language interpretation: Food for thought? *Journal of Memory and Language* 35, 544-565.

Nayak, N.P. & Gibbs, R.W., Jr. (1990): Conceptual knowledge in the interpretation of idioms. *Journal of Experimental Psychology: General* 119(3), 315-330.

Wolff, P. & Gentner, D. (2000): Evidence for role-neutral initial processing of metaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26(2), 529-541.

Csilla Rákosi PhD
MTA-DE Research Group for Theoretical Linguistics
H-4002 Debrecen, Pf. 400
rakosics@gmail.com