Sprachtheorie und germanistische Linguistik (sugl.eu), 27.1 (2017), 3-32

Csilla Rákosi

'Experimental complexes' in psycholinguistic research on metaphor processing*

Abstract

Although reliability is one of the most important requirements experiments should meet, it is almost never checked in psycholinguistic research on metaphor processing. Moreover, most replication attempts are not exact repetitions but involve some kind of modification. This finding appears to be in conflict with the basic idea of replications, since there is no striving for the closest possible repetition of all the details of the original experiment. Rather, replications in psycholinguistic research seem to be intended to fulfil a control function. This motivates the elaboration of the concept of 'experimental complex', which makes it possible to reconstruct and evaluate chains of closely related experiments.
Keywords: psycholinguistic experiments, cognitive theories of metaphor, replication of experiments, metaphor processing, reliability of experiments, cognitive linguistics methodology

* Work on this paper was supported by the MTA-DE-SZTE Research Group for Theoretical Linguistics and the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

> "Two central values of science are openness and reproductibility." (Nosek & Lakens 2014: 139)

1 Introduction

Although experiments are regarded as one of the most important and valuable data sources in cognitive linguistics, their evaluation is often highly controversial. In this research field, it is usually heavily debated whether or not the results of an experiment are reliable and valid. This might sound paradoxical as regards the former criterion, insofar as there is a generally accepted and simple way of checking whether an experiment is reliable, namely *replication*:

> Today is generally assumed that isolated experimental outcomes – »one-offs« – are insignificant. Twentieth-century philosophers of science, most notably Popper, made the reproductibility of experimental results the basic methodological requirement for successful experimentation: if an experiment cannot be re-done, it is invalid (Schickore 2011: 327).

In spite of this, the vast majority of psycholinguistic experiments have not been replicated. There are several, mainly social and psychological factors which have contributed to this situation:

- Papers dealing with novel, original results are considered superior by (psyho)linguistics, and are strongly preferred by journals and researchers alike. Experiments with negative outcomes are rarely publicised, while replications are practically banned from the acknowledged forums of scientific discourse.
- Although the standards applied by (psycho)linguistic journals have gradually become stricter, many experiments are not even replicable due to the lack of a sufficiently detailed description of the experimental procedure in the experimental report.
- Even though the experimental design and the experimental process were documented carefully in the experimental report, there are always details which would be needed in order to produce an exact reproduction of the original experiment. Thus, in practical terms, there is no such thing as a perfect replication – repetitions can only be closer or not so close.
- Replication attempts often lead to contradictory results and to barren controversies between the researchers who have conducted the original experiment and those undertaking the repetition.

There are several alarming signs indicating that this practice cannot be regarded as beneficial. For example, the *Open Science Collaboration* project replicated 100 experiments and correlational studies in psychology, and found, on the basis of five indicators, that "[a] large portion of replications produced weaker evidence for the original findings despite using materials provided by the original authors, review in advance for methodological fidelity, and high statistical power to detect the original effect sizes" (Nosek et al. 2015). Nosek et al. (2014) or Meyer & Chabris (2014) provide a deep analysis of the destructive consequences of the neglect of replications. On top of this, although the opposite is often declared, the evaluation of experiments lacks clear and generally accepted guidelines. As, for instance, the special issue of *Behavioral and Brain Sciences*, 14 (1991) pp. 119-186 testifies, problems related to peer reviewing in the publication of experimental reports are chronic.

These findings clearly show that the neglect of replications has to be deemed a serious methodological failure, making the reliability of psycholinguistic experiments as data sources dubious. Therefore, the common practice should be rethought and new methodological guidelines should be elaborated and issued. Novel approaches to replications are also paramount from a (general) philosophy of science point of view, since

> […] the very concept of replication has not received much analytic attention. Only recently, a few philosophers have begun examining more systematically the concepts of replication, reproductibility, and robustness or multiple determinations [...]. As yet, no consensus about these concepts, their meaning and significance has emerged (Schickore 2011: 345).

As a first step in this direction, the following question should be answered:

(Q) What role do replications play in the evaluation of psycholinguistic experiments?

We don't intend to answer this question in general, but we will rely on a case study by analysing various replication attempts conducted within cognitive metaphor research. Section 2 will offer a first concise description of the original experiment, its replications, and related counter-experiments. It is important to emphasise that this is a quasi-historical case study. This means that it intends to show the applicability of the presented metatheoretical model on an illustrative

example. Although the role of conventionality, familiarity and aptness is still heavily debated in the recent literature (see, for example, Thibodeau et al. 2016), the experiments analysed in this paper are still referred to in current literature on metaphor processing. In Section 3, a metascientific model of psycholinguistic experiments will be presented with the help of which psycholinguistic experiments can be analysed. This model will then be extended in such a way that the relationship between original experiments and their repetitions can be described. In Section 4, the extended 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 an answer to (Q).

2 Wolff & Gentner (1992) and its replications

2.1 The original experiment: Wolff & Gentner (1992)

Experiment 1: Participants were shown either the target or the base of a metaphor, or a blank line on a computer screen for 1500ms. After a 2500 ms pause, they saw the whole metaphor until a key press and had to type an interpretation of it. It was emphasised that they should start writing only when they had completely formulated their interpretation. They also received the instruction that they have to make use of the words presented and try to make a head start with their interpretation. According to the authors, if Glucksberg's Attributive Categorisation Theory is correct, and metaphor processing starts asymmetrically, by the derivation of a category from the base term which is then applied to the target term, then base primes should be more effective than target primes. This prediction, however, was not supported by the data obtained: there was no significant difference between base and target.

Experiment 2: Wolff and Gentner re-designed the experiment and modified the experimental procedure at two points. First, the role of the primes was made explicit. For example, the base prime *butcher* was presented as the sentence *A something is a butcher*, while if the target word was *surgeon*, the sentence *A surgeon is a something* appeared on the screen. The second change was that there was a fourth priming context, when the whole metaphor served as a prime. The authors put forward the prediction that the lack of a significant dif-

ference between 'both' and 'base' would provide evidence against Gentner's Structure Mapping Theory, stating that the early stage of metaphor processing consists of a matching process of the representations of target and base. There were significant differences between the conditions 'blank' and 'base', 'both' and 'base', and 'both' and 'target', while no significant difference was detected between 'target' and 'blank', and 'base' and 'target', respectively. These results were again found to be incompatible with Glucksberg's Attributive Categorisation Theory but in harmony with Gentner's Structure Mapping Theory.

Experiment 3: This experiment was motivated by a deeper analysis of the perceptual data gained in the previous experiment. The authors raised the hypothesis that the conventionality of metaphors, or more exactly, bases, is a factor that facilitates a processing mode that starts with the base term. While the first two experiments used novel metaphors, this experiment employed only bases with pre-stored, stock meanings. According to the authors, the experimental data obtained in this experiment reinforce this hypothesis and provide supporting evidence for Glucksberg's theory in relation to conventional bases.

2.2 Replication No. 1: Glucksberg, McGlone & Manfredi (1997)

Experiment 2: This experiment is a revised version of the experiments in Wolff & Gentner (1992) insofar as it used the same methodology with some modifications. First, the prime word appeared for 2 seconds instead of 1.5. Second, the applied category system was considerably refined. Primes were selected in such a way that topics were either high-constraint (*lawyer, mind*) or low-constraint (*my brother, life*) and vehicles either ambiguous (*jail*, *shark*) or unambiguous (*garden*, *puppy*); the classifications were checked with the help of control experiments. Third, no interpretations were required, but the space bar had to be struck when subjects understood the metaphor, so that the measurements – in contrast to Wolff & Gentner's experiment – captured the processing time only. Fourth, in order to secure a comprehensive reading, as a final task, participants had to fill in a questionnaire about the metaphors in the experiment. Fifth, the predictions were also different. As we have seen in Section 2.1, from the Attributive Categorisation View Wolff and Gentner inferred the prediction that vehicle primes should be more effective than topic primes. In

contrast, according to the authors' predictions, high-constraint topics and unambiguous vehicles should be effective primes for metaphor comprehension, while low-constraint topics and ambiguous vehicles should be ineffective or less effective. The experimental data clearly support the latter hypothesis.

2.3 Replication No. 2: Gentner & Wolff (1997)

Experiments 1-2: Experiments 1-2 were repetitions of Experiment 2 of Wolff & Gentner (1992) with a few modifications; they were also a reaction to Experiment 2 in Glucksberg et al. (1997). The stimulus material was somewhat wider (32 metaphors instead of 24), there were more participants, and the set of control measures was extended by disabling the backspace key in order to prevent subjects from editing their interpretations. In Experiment 2, participants were asked to press the spacebar as soon as they had an interpretation of the given metaphor. A further difference lay in the timing of the presentation of the stimuli. The ISIs of Experiment 2 in Wolff & Gentner (1992) and Experiment 2 in Gentner & Wolff (1997) were identical, while Experiment 1 in Gentner & Wolff (1997) applied a very short ISI between the prime and the entire metaphor. In both cases, the experimental data were found to be in harmony with the alignment-driven model, but inconsistent with the Attributive Categorisation View, since bases were not quicker than targets, and metaphors preceded by both primes were faster than bases or targets alone.

Experiment 3: The experimental design was a further development of Experiment 3 in Wolff & Gentner (1992). The stimulus material took two additional factors into account. Namely, it consisted of metaphors with high base conventionality and low relational similarity between base and target in order to secure Attributive Categorisation Theory maximally advantageous conditions against the alignment-based Structure Mapping Theory. The stimulus material can be found in the experimental report and was checked with the help of two control experiments. In this case, the experimental data were found to be in harmony with the predictions of the Attributive Categorisation Theory, indicating that under the special conditions described above, abstraction-first processing is preferred.

Experiment 4: Experiment 4 was a more elaborated version of Experiment 3 insofar as it had a 2x2x4 design with factors of base conventionality, relational similarity and prime type (both, base, target, blank). The ISI between prime and metaphor was 0 ms in this experiment. According to the authors' predictions, if the Attributive Categorisation Theory is correct, then there should be a base advantage under all conditions. In contrast, from Gentner's career of metaphor hypothesis it follows that there should be no base advantage for low-conventionality metaphors, there should be a base advantage for all high-conventionality or, at least, for high-conventionality and low-similarity metaphors, and high-conventionality metaphors should be faster than low-conventionality ones. This also means that the authors re-evaluated their theory as well. To wit, they narrowed down the scope of Structure Mapping Theory, and integrated it, together with a similarly reduced Attributive Categorisation Theory, into a more complex version of SMT, extended with the career of metaphor hypothesis.

2.4 Counter-experiments: Jones & Estes (2005, 2006)

Jones & Estes (2005), Experiment 1: The stimulus material, presented in the appendix of the paper, consisted of 32 high-similarity metaphors (16 conventional and 16 novel) from Gentner & Wolff (1997), Experiment 4, as well as 32 matched literal control sentences. After seeing a prime sentence (metaphor or literal control) for 4 seconds on the computer screen, participants had to answer the question of the extent to which the topic is a member of the category defined by the vehicle. They had to press button 1 for "non-member", 2 for "partial member" and 3 for "full member". According to Glucksberg's Attributive Categorization View, both novel and conventional metaphor-primed items should have higher ratings than literal controls, while Gentner's career of metaphor hypothesis leads to the prediction that only conventional metaphors should receive significantly higher ratings. The authors found that experimental data clearly support the former hypothesis.

Experiment 2: In order to rule out the possibility that the grammatical structure of the primes distorts the results, literal controls were omitted and two new control prime types were added: 16 borderline literal items (*A tire is a boat*, *A cucumber is a fruit*) and 16 scrambled

metaphor items (*Hard work is a teddy bear*, *Respect is a vampire*). Both the control and the metaphor stimuli were slightly reformulated in order to make them more natural-sounding. A further modification was that a 7-point scale was applied for the ratings. There was also an unprimed condition; that is, half of the participants made ratings without seeing a prime sentence, the other half obtained primes before providing judgements. The results showed the same pattern as in Experiment 1, and priming increased the categorisation ratings.

Experiment 3: The authors raised the conjecture that conventional items might have been more apt than novel items. Therefore, they changed the factor 'conventionality' to 'aptness'. The stimulus material included 32 high apt and 32 low apt metaphors, collected from 4 papers by different authors. A separate group of participants provided the aptness ratings; the high apt metaphors were significantly more apt than the less apt ones. Aptness was found to increase class inclusion effectively.

Jones & Estes (2006), Experiment 3: This experiment relied on the same stimulus material as Experiments 1 and 2 in Jones & Estes (2006) and was a revised version of Experiments 2 and 3 in Jones & Estes (2005). Namely, it used the same methodology but besides aptness, it also controlled conventionality – thus, it tested the factors investigated by Experiments 2 and 3 together. The experimental data obtained support Glucksberg's ACV and contradict the career of metaphor hypothesis, because category membership ratings were higher for the high apt metaphors than for low apt ones, while no difference was found between novel and conventional metaphors. Further, there was no interaction between conventionality and aptness.

2.5 Remarks

The most striking feature of the replications is that they are not exact repetitions but rather modified or refined versions of the original or the previous experiment. The modifications pertain to different aspects of the experimental design or the relationship between theory and predictions. A further important point is that in this respect there is no difference between those repetitions conducted by the original authors and those conducted by adherents of rival approaches.

That is, the original experiment belongs to a series of closely related experiments which try to rule out possible systematic errors or make use of a more differentiated stimulus material and research hypothesis. Similarly, the counter-experiments by Jones and Estes are nothing other than variations of the starting experiment, which make use of the same stimulus material as one of Wolff and Gentner's experiments. Nevertheless, there is an important difference. The outcome of the experiments conducted by the authors of the original experiment is interpreted in such a way that the results either reinforce the original research hypothesis, or motivate its further refinement and the elaboration of a new theory-version. In contrast, the experimental data gained by adherents of rival approaches are regarded as conflicting with the original results and motivating the rejection of the original theory. To put it differently, while repetitions by the researcher who conducted the original experiment seem to increase the plausibility of the data originating from the original experiment, related experiments (non-exact replications or counter-experiments) conducted by adherents of rival approaches decrease it. Therefore, the question emerges of how such "cumulative" contradictions can be resolved.

3 Metatheoretical background

3.1 A model of psycholinguistic experiments

The basic idea of the model of psycholinguistic experiments as presented in Rákosi (2012, 2014, 2017) and Kertész & Rákosi (2012) is that experiments are not linear processes, leading from an unimpeachable experimental design and a set of completely secure observations to entirely reliable experimental results demonstrating or falsifying theories, but *open, cyclic problem solving processes built on and producing uncertain information*, organised and conducted by a *plausible argumentation process*.[1] This latter process governs the relationship among the components of the experimental process: the hypotheses of the experimental design, the theoretical model of phenomena, the theoretical model of the experimental apparatus, the statements describing the events of the experimental procedure, the

[1] For lack of space, we cannot go into the details of the model of psycholinguistic experiments and the p-model of plausible argumentation; we can only delineate their basic ideas here.

statements capturing the results of the interpretation and authentication of perceptual data, as well as the theory being tested and its rivals, etc. See Figure 1.

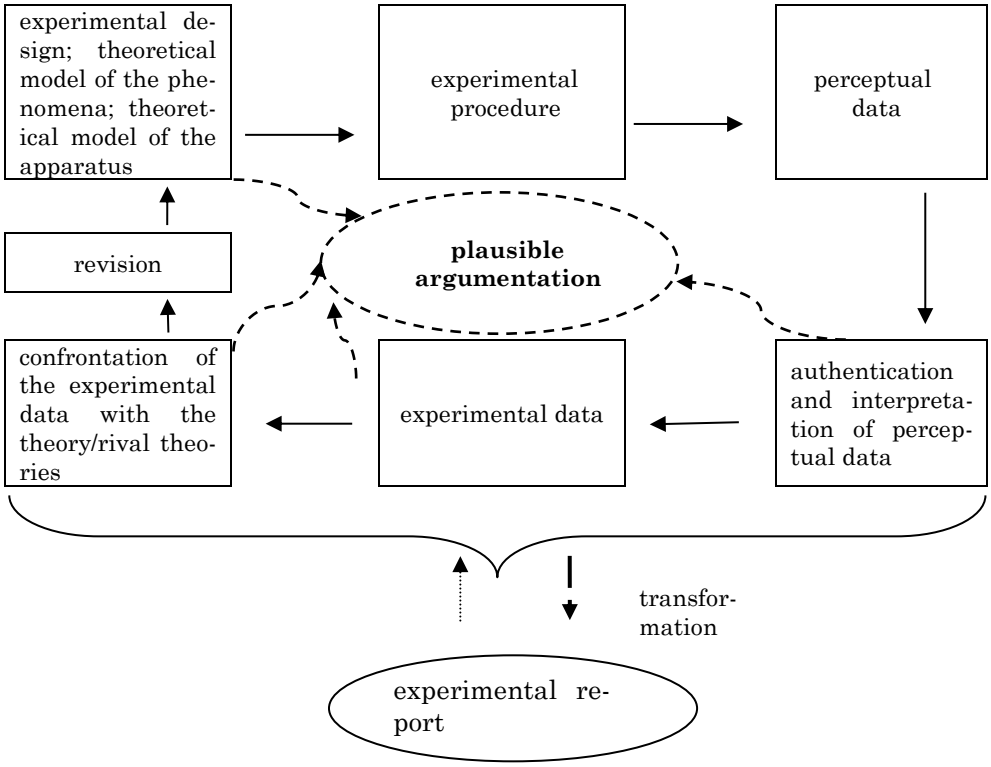

Figure 1[2]

Plausible argumentation is a cyclic process that continuously *re-evaluates* the plausibility (acceptability) of statements in the light of new pieces of information and tries to resolve conflicting assessments of the plausibility of a given statement.[3]

The *effectiveness* and *comprehensiveness* of the plausible argumentation process organising the conduct and control of experiments

[2] Simple arrows indicate successive stages of the experimental process; dotted arrows signify the non-public argumentation process which organises the experimental process.

[3] That is, it may happen that a statement is plausible on the basis of a source but at the same time, some other source makes its negation plausible.

largely determines the acceptability of the results obtained. This argumentation process is, however, not public but remains a *private affair* of the experimenters. It is transformed into an experimental report/paper summarising the results of the experiment and making them available to the scientific community. Clearly, this transformation can be regarded as acceptable if *it does not change the plausibility value of the statements (data, hypotheses) of the original argumentation*. This is of utmost importance because there is a danger that the researcher eliminates relevant information from the published report so that important decisions remain outside public control, and overestimates the plausibility of the results. From this it follows that the acceptability of the experimental report is also influenced by the *transparency of the transformation of the non-public argumentation process into its public version*.

3.2 The relationship between original experiments and replications: Experimental complexes

If we summarise the moral of the remarks relating to the experiments presented in Section 2, we can reach the conclusion that most replication attempts are *not exact repetitions* but involve some kind of *modification*. Thus, they cannot be described neither as 'multiple repetitions of the same experiment' or 'procedural replications', nor as 'multiple determinations of experimental results', that is, attempts at "obtaining similar results in different experimental settings" (Schickore 2011: 328). This finding appears to be in conflict with the basic idea of replications, since there is no striving for the closest possible repetition of all details of the original experiment. Rather, replications seem to be intended to fulfil a control function. To put it differently, a *cyclic process of re-evaluation* is at work

- among closely related experiments conducted by the same authors, and usually published within a research article in order to rule out some possible sources of systematic error, refine the research hypothesis, and/or increase the reliability of the results, and
- among original experiments and non-exact replications by other authors which apply more differentiated stimulus material and/or intend to test a more elaborated research hypothesis, as well as

– among original experiments and counter-experiments which make use of the same stimulus material but apply a different method in order to provide evidence against the original experiment's results.

From this it follows that *the evaluation of psycholinguistic experiments has to transgress the boundary of single experiments*. This motivates the elaboration of the concept of the 'experimental complex':

(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, 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 implausible. Such experimental complexes are considerably more complex than single experiments, because they 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 (or its non-exact replications) as well as among experiments belonging to the experimental complex,
– different kinds of *problems* as well as *solution attempts*,
– *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.

As Figure 2 shows, experimental complexes have basically the same cyclic structure as single experiments:

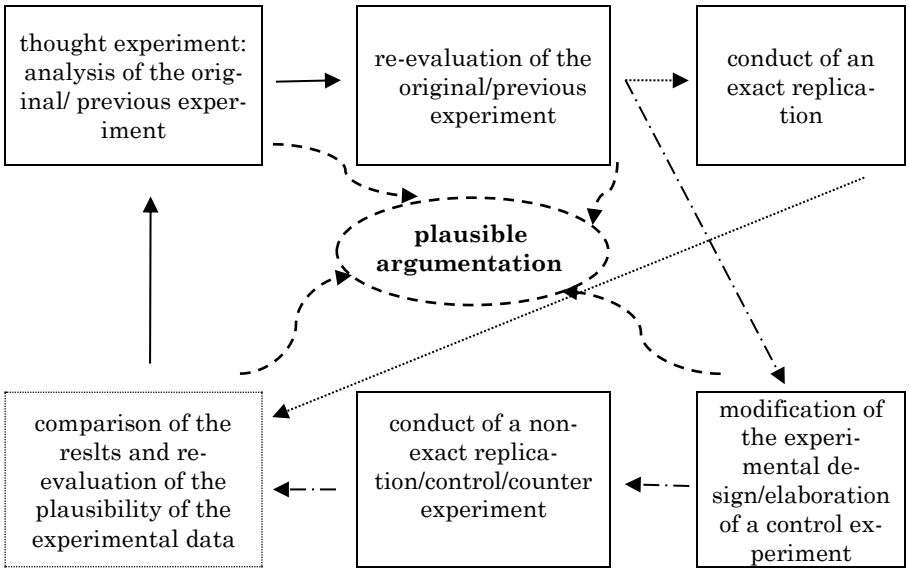

Figure 2

In the long run, non-exact replications may provide increasingly similar results, but it is also possible that existing conflicts deepen and multiply. In order to provide tools for the description of such situations, we introduce the following concepts:

(D2) An experiment is the *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).

This definition stipulates very strict criteria. These are only fulfilled if a series of non-exact and exact replications leads to an experiment that is, at least temporarily, stable and generally accepted by the

members of the given research field. In such cases, the experimental complex is convergent:

(D3) An experimental complex is *convergent* if it has a limit; otherwise, it is *divergent*.

However, 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. A further important remark is that the limit of a convergent experimental complex may be inconsistent with the outcome of some earlier member of the chain of non-exact replications to which it belongs, or with experimental data originating from other experiments belonging to some other experimental complex. Moreover, an experimental complex may have many limits during its development. These are in most cases at variance with each other, and the later ones always count as revisions of the earlier ones. Nevertheless, any modification may not only rule out possible systematic errors but also lead to the emergence of new ones. Against this background, one can distinguish between progressive and stagnating non-exact replications:

(D4) A non-exact replication is *progressive* if it eliminates at least one 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*.

Progressive replications provide well-motivated and effective re-evaluations of the original experiment, mostly produce more plausible experimental data, and may bring us closer to the limit of the experimental complex. It is not required, however, that they eliminate all problems of the original experiment or their predecessors, or that they are free of (known) error types.

There are three basic types of scenarios:

1. The experimental complex is convergent. See Figure 3:

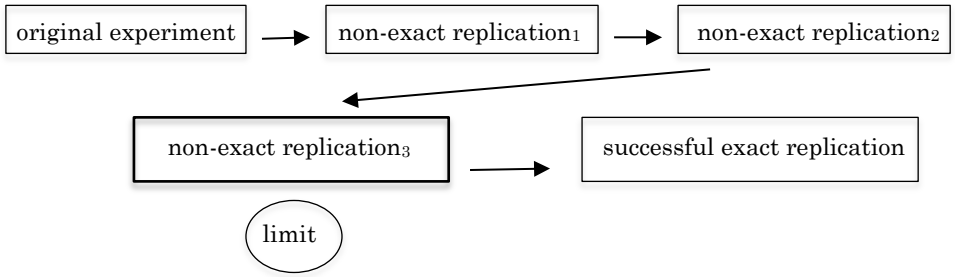

Figure 3

2. The experimental complex is divergent because the final non-exact replication is not reliable. See Figure 4:

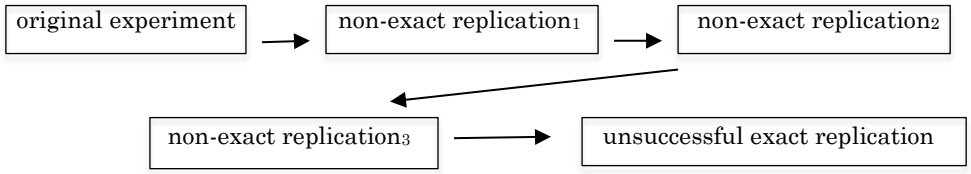

Figure 4

3. The experimental complex is divergent because the final non-exact replication was shown to be problematic (for example, it is not valid) and it is not clear whether and if so, how a revised version could be designed. See Figure 5:

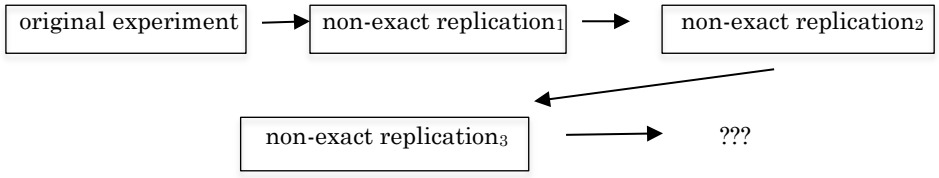

Figure 5

Of course, there are many further possible scenarios, which may be considerably more complex. For example, a convergent experimental complex may have "dead ends", i.e., non-exact replications which cannot be continued. In such cases, the process turns back to an earlier

stage and a new series of replications is conducted. It may also happen that an experimental complex has more limits. In such cases, different revisions of the original experiments have led to conflicting results, and in the given informational state, it is unclear how this inconsistency can be resolved.

It is also important to emphasise that experimental complexes are not isolated entities but may have different kinds of relationships to other experimental complexes. Experimental complexes may also overlap in the sense that an experiment may also belong to two complexes – indeed, of course, in different roles (for example, as a non-exact replication and as a counter-experiment). The description of such constellations, however, should be the subject of another paper.

In the next section, we will reconstruct the experiments briefly presented in Section 2 with the help of this model. Thus, our aim will be to find out whether there is a convergent experimental complex among them. The re-evaluation of an experimental complex cannot be reduced to the analysis of its final state; the whole process has to be reconstructed. This boils down to the following steps:

– the separate reconstruction and re-evaluation of the experiments belonging to the experimental complex (plausibility of the experimental data);[4]
– the reconstruction and re-evaluation of the relationship between the experiments (checking the progressivity of the replications);
– the evaluation of the convergence/divergence of the experimental complex.

A thorough analysis along these lines would be, however, lengthy. Therefore, Section 4 will focus on the progressivity of the non-exact replications, and the evaluation of the convergence of the experimental complex. The analyses presented are not intended to be complete; their task is solely to illustrate the workability of the model presented in this section.

[4] For the metatheoretical background to this step, see Rákosi (2016a,b).

4 Reconstruction of the experimental complex evolving from Wolff & Gentner (1992)

As Figure 6 shows, the experimental complex evolving from Experiment 1 in Wolff & Gentner (1992) involves the original experiment (OE), 7 non-exact replications (NR1-4) and 4 counter-experiments (COU1-4):

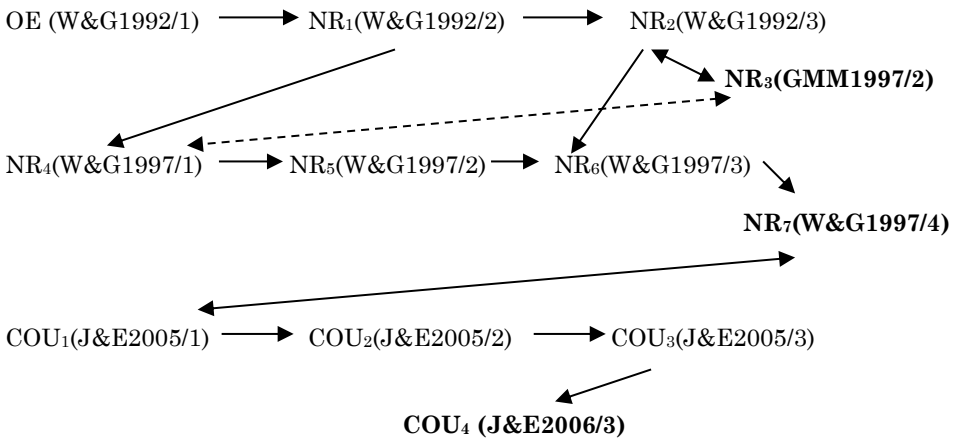

Figure 6

Among them, three chains of experiments can be identified:

– NR_1, NR_2, NR_4, NR_5, NR_6 and NR_7 are more and more elaborated versions of the original experiment, whose results seem to be in harmony;
– NR_3 is a non-exact replication of NR_2, leading to a conflicting result;
– the counter-experiments COU_1, COU_2, COU_3 and COU_4 (where COU_2, COU_3 and COU_4 are non-exact replications of COU_1) make up the third chain, with a varying result again.

Thus, we have two limit-candidates: NR_3 (Section 2.2), NR_6 (Section 2.3), as well as a limit-candidate of a series of counter-experiments: COU_4 (Section 2.4). The first step of the re-evaluation of this experimental complex should be the reconstruction and analysis of the three chains of experiments.

4.1 The limit-candidate by Gentner and Wolff

OE (cf. Section 2.1): The first step of the re-evaluation consists of *the identification of the problematic points of the original experiment*:

Problem 1: The times measured were not processing times but the times required to process a metaphor and formulate an interpretation. Therefore, it might be the case that there was a major difference in processing times but this was masked by the elaboration of the given interpretation. This seems to be a strong possibility because thinking out an interpretation takes longer than processing a metaphorical sentence.

Problem 2: The primes did not reveal the role of the presented words, that is, participants could not know whether the word on the screen will be a base or a target. They might have made false starts, and, in order to avoid those, have applied conscious strategies instead of making spontaneous head starts.

Problem 3: It is not clear whether the choice of the presentation time of the primes and the ISI were correct and the experiment touches upon the early stage of metaphor processing.

Problem 4: The stimulus material contained solely novel metaphors in the sense that none of the applied base terms had conventional metaphorical meaning. This reduces the generality of the investigations.

Problem 5: The stimulus material is missing in the experimental report. Therefore, its correctness cannot be checked.

As the second step, we have to *check the progressivity of the non-exact replications*:

NR_1 (cf. Section 2.1): The addition of the condition 'both' weakens the strength of Problem 1, since the results show that the experiment was sensitive enough to detect relevant differences. Problem 2 has been successfully prevented with the modification of the stimulus material. Problems No. 3, 4, and 5, however, emerge in this case again. Moreover, two new problems arise:

Problem 6: There is a conflict between the experimental data and the research hypothesis. Namely, if in the first stage of metaphor processing, the role of the base and target is symmetrical, then there should be a significant difference not only between bases and blanks but also between targets and blanks.

Problem 7: There is a conflict between the results of the original experiment and its non-exact replication. That is, the original experiment yielded a significant difference not only between bases and blanks but also between targets and blanks, while this was not the case with its first non-exact replication.

This means that NE_1 is a progressive replication but it cannot be regarded as a limit of this experimental complex.

NR_2 (cf. Section 2.1): Experiment 3 is a progressive replication, too, because it addresses Problem 4, and extends the investigations to conventional metaphors. Nevertheless, it leaves Problems 3, 5, 6, and 7 open and raises Problems 8 and 9:

Problem 8: The relationship of the experimental data and rival hypotheses is indeterminate. To wit, frequently used, conventional base terms might guarantee shorter interpretation times by facilitating head starts. Thus, it is not clear how to distinguish matching-first models, speeded up with head starts, from mapping-first models. Unfortunately, base conventionality is a factor that cannot be balanced, because there are no conventional target terms that could influence the target primes' interpretation times in a similar manner.

Problem 9: The stimulus material is comprised solely of conventional metaphors. Thus, the experiment does not allow a direct comparison of novel and conventional metaphors.

NR_4 (cf. Section 2.3): Experiment 1 in Gentner & Wolff (1997) is a progressive non-exact replication of NE_1. Namely, both the stimulus material and the number of participants have been increased, and Problem 5 was solved. Nevertheless, Problems 4, 6 and 7 remained untouched, and the application of different ISIs did not lead to similar experimental data; thus, Problem 3 is open, too.

NR_5 (cf. Section 2.3): In Experiment 2, the impact of Problem 1 was reduced; Problems 3 and 7, however, have become more serious, leading to Problem 10:

Problem 10: The interpretation of the perceptual data is deficient, because in NR_4, the authors found a significant difference between blanks and targets or bases alone, and interpreted this finding as "indicating that the primes were effective". In NR_5, however, no significant difference was found among these conditions, and the authors did not comment on this result. Thus, an unreflected and unsolved conflict between the results of similar experiments emerged. There are further differences between the results of NR_4 and NR_5, which require explanation (see Kertész & Rákosi 2012, p. 232).

NR_6 (cf. Section 2.3): Experiment 3 of Gentner & Wolff (1997) is a non-exact replication of NR_2 as well as NR_5. Its progressivity results from a refinement of the research hypothesis and the circumstance that it addresses Problem 4. The attempted solution, however, once again raises new problems:

Problem 11: The authors found in NR_2 that high base conventionality is alone effective and led to the same results. Thus, the role of the factors of conventionality and relational similarity has been left open, and a new conflict between non-exact replications has emerged.

Problem 12: The wording of the metaphors presented was changed from the earlier versions of this experiment type. Namely, in the experiments in OE, NR_1, NR_2 as well as NR_4, metaphors of the form "An *X* is a *Y*" were used, while in NR_5 and NR_6, the formulation "That *X* is a *Y*" was chosen. The former statement is a generalisation stating that very *X* is a *Y*, while the latter is a statement about a singular exemplar of a category. This difference might have influenced participants' expectations and behaviour, since statements about individuals are more frequently acceptable, while generalisations can turn out to be often false or awkward.

Problem 13: There is an inconsistency in the judgement of the degree of conventionality with the stimulus material of the ex-

> periments within this experimental complex. Namely, while the authors emphasise that OE, NR_1, NR_3 and NR_4 made use of novel metaphors, they also state that "[R]esults from these ratings indicated that the bases for the metaphors used in Experiments 1 and 2 were fairly high in conventionality (M = 4.86) [on a scale from 1 to 7 – Cs. R.]. The bases for the new metaphors constructed for Experiment 3 were rated somewhat higher in conventionality (M = 5.72)" (Gentner & Wolff 1997: 341).

NR_7 (cf. Section 2.3): This non-exact replication of NR_6 is progressive because it tackles Problems 4, 9 and 11, and raises a new, more refined research hypothesis. Indeed, Problem 3 arises again, since it is also not clear what the shrinking of the ISI between primes and targets to 0 ms motivated. Variances with the outcome of the experiments using a longer ISI might also be due to this factor.

Table 1 gives an overview of the re-evaluation process in this chain of non-exact replications:[5]

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
OE	E	E	E	E	E								
NR_1	P	S	O	O	O	E	E						
NR_2	O	S	O	P	O	O	O	E	E				
NR_4	O	S	O	O	S	O	O						
NR_5	P	S	O	O	S	O	O			E			
NR_6	P	S	O	P	S	O	O	O	O		E	E	E
NR_7	P	S	O	S	S	O	O	O	S	O	S	O	O

Table 1

To sum up, NR_7 provides the most plausible experimental data from the members of this series of experiments. Despite this, it cannot be regarded as a limit of this experimental complex. This verdict is based on the finding that not all problems have been resolved, and the elaboration and conduct of new, improved versions seems to be possible.

---

[5] In Tables 1-3, '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.

4.2 The limit-candidate by Glucksberg, McGlone & Manfredi

NR_3 (cf. Section 2.2): The progressivity of the non-exact replication of the original experiment by Glucksberg et al. is due to the extension of the research hypothesis with further possibly relevant factors, and the provision of solutions to Problems 1, 2, 4, 5, 8 and 9. Nevertheless, new problems emerge here, too:

Problem 14: Since the instructions contained explicit reference to metaphors,[6] the aim of the experiment was not masked.

Problem 15: Both high-constraint topics and unambiguous vehicles are less susceptible to causing false starts than low-constraint topics and ambiguous vehicles. Therefore, the former primes' advantage over the latter might be partially due to this circumstance, independently of the processing mode of metaphors.

Problem 16: The relationship of the experimental data and rival hypotheses is indeterminate. First, from Wolff and Gentner's interpretation of the Attributive Categorisation View it would follow that unambiguous vehicles should be faster than high-constraint topics. Second, the experimental data obtained seem to be consistent with Gentner's Structure Mapping Theory, since both high constraint topics and unambiguous vehicles offer fewer properties for matching and both may facilitate the projection of candidate inferences from vehicle to topic.

It is easy to see that Problem 16 is analogous to Problem 8. As Table 2 shows, NE_3 cannot be regarded as a limit of this experimental complex, either, although it is clearly progressive and produces more plausible experimental data than its predecessors:

---

[6] Cf.: "In this study, you will be asked to read a series of metaphors. Metaphors are figurative statements such as Shakespeare's *All the world's a strange* or the common expression *Some lawyers are sharks*." (Glucksberg et al. 1997: 61)

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P14	P15	P16
OE	E	E	E	E	E							
NR_1	O	S	O	O	O	E	E					
NR_2	O	S	O	P	O	O		E	E			
NR_3	S	S	O	S	S			S	S	E	E	E

Table 2

4.3 Counter-experiments by Jones and Estes

COU_1 (cf. Section 2.4): The first experiment in Jones & Estes (2005) is a counter-experiment to NR_7. This means two things. First, it makes use of the stimulus material of Gentner & Wolff (1997)'s Experiment 4, extending it with literal control sentences, but applies a different methodology: instead of measuring interpretation times, it collects categorisation ratings. Second, it aims to provide evidence for a hypothesis that was rejected by the authors of NR_7. The plausibility of the experimental data is questioned by the following problems:

Problem 17: This experiment does not test mental processing directly but investigates subjects' conscious considerations and might invite them to create naïve theories about language.[7] To wit, it may be the case that the experiment shows solely that people are capable of, and willing to, consciously interpret metaphors as (more or less apt) categorical statements. From this, however, one cannot conclude that they process metaphorical expressions as categorical statements.

Problem 18: The metaphorical stimuli and the literal stimuli applied have a different grammatical structure.[8] While the metaphors had a nominal structure and stated that the topic is a kind of the vehicle, inviting a 2- or 3-rating, the lit-

[7] See, for example, the formulation of the tasks: "To what extent are ARGUMENTS a member of the category WAR?"

[8] For example: *That sauna is an oven* (conventional metaphor); *That sauna is located behind an oven* (conventional literal); *That canary is a violin* (novel metaphor); *That canary flew over a violin* (novel literal).

Cf.: "Another alternative explanation of Experiment 1 is that the nominal structure (i.e., *That X is a Y*) of the metaphorical primes may have induced a task demand, such that participants were more likely to judge that an X is a Y after reading the prime *That X is a Y*." (Jones & Estes 2005: 116)

eral counterparts interpret topic and vehicle as two different participants of a scene and tend to suggest a 1-rating.

Problem 19: The literal sentences sound odd in several cases in comparison to their metaphorical counterparts, which might also have influenced participants' decisions.

Problem 20: A further problem with the task given to the participants is that Glucksberg's ACV states that metaphors are interpreted in such a way that the topic belongs to an *ad hoc* category, and the vehicle is a typical member of this category. It is not required, however, that this *ad hoc* category is that of the vehicle *per se*; instead, the vehicle term usually exemplifies an abstract category which does not have a name (cf. Glucksberg et al. 1997: 52).

Problem 21: It is not clear why there is a main effect of conventionality as well.

Problem 22: The usage of a 3-point scale is clearly a less sensitive tool than a 1-7 scale would be.

COU$_2$ (cf. Section 2.4): This experiment is a progressive non-exact replication of COU$_1$, since it solves Problems 18, 19 and 22. Problems 17, 20, and 21, however, remained unsolved, and two new problems emerged:

Problem 23: The low values with both novel and conventional metaphors (2.47 vs. 3.11) suggest that metaphors are not viewed as a kind of categorisation, since these scores are below the scalar midpoint. This does not, however, mean that people would not process metaphors as categorical statements unconsciously (cf. Problem 17).

Problem 24: As an infelicitous side effect of the extension of the stimulus material, the number of tasks was too high. This might have led to boredom effects and/or the use of conscious strategies.

COU$_3$ (cf. Section 2.4): Experiment 3 is a progressive non-exact replication, too: it addresses Problem 21 and provides a solution for Problem 23. As for the latter, high apt metaphors received rather high categorisation ratings (4.23), and low apt metaphors obtained clearly low values (2.29). Nevertheless, a new problem unfolded, which seems to be, however, less severe than Problem 23 was:

Problem 25: Even high apt metaphors were evaluated lower (4.23) than borderline literals (4.98).

COU$_4$ (cf. Section 2.4): Experiment 3 in Jones & Estes (2006) is a progressive non-exact replication of COU$_2$ and COU$_3$, because it investigates both possibly relevant factors jointly, and provides a more satisfactory solution to Problem 21. Despite this, this experiment inherited the problems relating to the stimulus material of Experiments 1-2 in Jones & Estes (2006) as well as several weak points of the methodology used in Experiment 2 in Jones & Estes (2005). Thus, Problems 24 and 25 have become even more serious,[9] and the following problem should be added to those already presented:

Problem 26: 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).

Table 3 clearly indicates that COU$_4$ cannot be regarded as a limit of this experimental complex, either, because it is still multiply problematic:

[9] High apt metaphors obtained the average rating 3.63, low apt metaphors 2.28, and borderline literals 5.17.

	P17	P18	P19	P20	P21	P22	P23	P24	P25	P26
COU_1	E	E	E	E	E	E				
COU_2	O	S	S	O	O	S	E	E		
COU_3	O	S	S	O	S	S	S	O	E	
COU_4	O	S	S	O	S	S	S	O	O	E

Table 3

Moreover, Problem 17 is a problem which calls the plausibility of the data obtained from these experiments into question. Thus, the data originating from $COU_{1\text{-}4}$ cannot be regarded as plausible in connection with metaphor processing.

4.4 Evaluation of the experimental complex

The reconstruction of the three chains of experiments shows that the experimental data originating from experiments by the same researchers have, in most cases,[10] become more plausible. Despite this, our analyses lead to the conclusion that this experimental complex is not convergent. All of the three limit-candidates contain unsolved problems, which motivate the elaboration of further non-exact replications. Nevertheless, the reliability of the experiments, i.e., the stability of the results, did not increase, because there was no perfect harmony among the corresponding results of the replications, and there were substantial differences between the experimental designs, as well. We have also seen that the two chains of non-exact replications and the counter-experiments lead to conflicting results. Such contradictions cannot be resolved simply by a mechanical comparison of the plausibility value of the last member of the chains of experiments. For instance, in this case, it would be a failure to choose the more plausible limit-candidate and reject the other one. Instead, non-exact replications to both NR_7 and NR_3 should be elaborated and conducted, and an online version of $COU_{1\text{-}4}$ should be developed. That is, the conflict between experiments can be resolved by the continuation of the process of re-evaluation with the help of new non-exact replications. In most cases, it is not the current state of the cyclic process of re-evaluation that is decisive but *the assessment of future prospects*. Another possibility of conflict resolution among experiments is the

[10] The relationship between NR_1 and NR_2, as well as NR_4 and NR_5 can be regarded as complementary rather than consecutive.

application of the tools of statistical meta-analysis (see Rákosi, in preparation; Rákosi, manuscript).

5 Summary

In Section 1, we raised the following question:

(Q) What role do replications play in the evaluation of psycholinguistic experiments?

On the basis of our considerations in Section 3 as well as the moral of the case study in Section 4, the following answer presents itself:

(A) (a) Non-exact replications may lead to more plausible (acceptable) experimental data. The increasing plausibility (acceptability) is due to the progressivity of the non-exact replications, i.e., it results from the successes in the problem solving process and/or the refinement of the research hypothesis.

(b) This is, however, not a steady growth, because the elaboration and conduct of more refined versions of the original experiment may give rise to the emergence of new problems, too.[11]

(c) Non-exact replications do not seem to be weaker tools than exact repetitions. For instance, a replication making use of an improved set of stimuli may provide even a stronger piece of evidence than one using the same stimulus material.

(d) Thus, checks for reliability and validity cannot be separated from each other. Successful non-exact replications motivated by problems (such as concerns about the validity) of the original experiment may also increase the latter's reliability, if there is harmony between their corresponding results.

(e) Convergent experimental complexes may be the result of a co-operation between exact and non-exact replications. While non-exact replications have to eliminate all known problem sources, exact replications can secure the reliability of the re-

[11] For the clarification of the criteria with the help of which aspects of the effectiveness of the problem solving process can be differentiated and judged, as well as for the strategies of inconsistency resolution between replications of psycholinguistic experiments, see Rákosi (2016a,b, 2017).

sults. Of course, 'multiple determinations of experimental results' (cf. Section 3.2), that is, experiments belonging to other experimental complexes may increase the plausibility of the results, too.

(f) Nevertheless, the concept of 'exact replication' could be extended so that it also covers cases when the only difference between the original experiment and its replication lies in the stimulus material. That is, generalizability requires that the application of the same experimental method produces the same results when applied to linguistic material constructed along the same theoretical principles.

References

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.

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

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.

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.

Meyer, M.N. & Chabris, C. (2014): Why psychologists' food fight matters.
http://www.slate.com/articles/health_and_science/science/2014/07/repliction_controversy_in_psychology_bullying_file_drawer_effect_blog_posts.html.

Nosek, B.A. et al. (2015): Estimating the reproducibility of psychological science. *Science* 28, Vol. 349, no. 6251.
DOI: 10.1126/science.aac4716.

Nosek, B.A. & Lakens, D. (2014): Registered Reports. A Method to Increase the Credibility of Published Results. *Social Psychology* 45(3), 137-141. DOI: 10.1027/1864-9335/a000192.

Open Science Collaboration (2015): Estimating the reproducibility of psychological science. *Science* 349(6251), aac4716.
DOI: 10.1126/science.aac4716.

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.

Rákosi, Cs. (2016a): On the evaluation of psycholinguistic experiments on metaphor. Part I: The metatheoretical background. *Argumentum* 12, 278-287.
http://argumentum.unideb.hu/2016-anyagok/rakosics1.pdf

Rákosi, Cs. (2016b): On the evaluation of psycholinguistic experiments on metaphor. Part II: Case studies. *Argumentum* 12, 288-302.
http://argumentum.unideb.hu/2016-anyagok/rakosics2.pdf

Rákosi, Cs. (2017): Replication of psycholinguistic experiments and the resolution of inconsistencies. *Journal of Psycholinguistic Research*, DOI:10.1007/s10936-017-9492-0.

Rákosi, Cs. (manuscript): *Do metaphors influence thinking about crime? A meta-analytic study.*

Rákosi, Cs. (in preparation): *The role of conventionality, familiarity and aptness in metaphor processing. A meta-analytic study*.

Schickore, J. (2011): The significance of re-doing experiments: A contribution to historically informed methodology. *Erkenntnis* 75, 325-347.

Thibodeau, P.H., Sikos, L., & Durgin, F.H. (2016): What Do We Learn From Rating Metaphors? *Proceedings Of The 38th Annual Conference Of The Cognitive Science Society,* 1769-1774. http://works.swarthmore.edu/fac-psychology/919.

Wolff, P. & Gentner, D. (1992): The time course of metaphor comprehension. *Proceedings of the fourteenth annual conference of the Cognitive Science Society*. Hillsdale: Erlbaum.

Dr. Csilla Rákosi
MTA-DE-SZTE Research Group for Theoretical Linguistics
University of Debrecen
Pf. 400
H-4002 Debrecen
rakosics@gmail.com