MTA doktori értekezés

Foundational quandaries in Cognitive Linguistics: Uncertainty, inconsistency, and the evaluation of theories

Rákosi Csilla

MTA-DE-SZTE Elméleti Nyelvészeti Kutatócsoport

Debrecen, 2019

Table of contents

> *"Every practicing scientist, past and present, adheres to certain views about how science should be performed, about what counts as an adequate explanation, about the use of experimental controls, and the like. These norms, which a scientist brings to bear in his assessment of theories, have been perhaps the single major source for most of the controversies in the history of science, and for the generation of many of the most acute conceptual problems with which scientists have had to cope." (Larry Laudan 1977: 58)*

1. Preface

While many researchers are uninterested in foundational issues and seem to be of the opinion that linguistics can be practised without making explicit the background assumptions of theories, there is a deep feeling of unease about such issues that is openly expressed over and over again. Critique is offered by researchers belonging to different schools and is levelled at different aspects of linguistic theorising. In a paper on the methodological foundations of linguistics, Raffaele Simone points to an inherent tension within linguistics. She argues that despite this diversity of criticism, there have been two basic strivings since the beginnings of this discipline. The first one is, as Simone calls it, *"Saussure's dream"*, according to which one should

> "provide linguistics with an appropriate *method*, one not borrowed more or less mechanically from other sciences, but designed to be peculiarly and strictly of its own." (Simone 2004: 238; emphasis as in the original)

Simone labels the second endeavour *reductionism*:

> "[…] two different types of reduction have taken place: (a) the reduction of linguistics to some other science, and (b) the reduction of language data to some other entity." (Simone 2004: 247)

Although the tenability of this kind of total reductionism can be questioned,[1] we can add a third type of reduction which is clearly present in linguistics and is of central importance for us: *methodological reduction,* meaning that linguists often try to borrow methodological tools and norms from other disciplines.

The presence of the two strivings can be traced back to the same cause: the scientificity of linguistics is often felt to be unsatisfactory in comparison to the standards of natural sciences or even social sciences.[2] This inferiority complex is mostly articulated as the requirement to

[1] Simone refers, among others, to Chomsky's statement that linguistics is nothing else but a branch of psychology. We should not forget, however, that generative grammar is a long way from applying same methodology as cognitive psychology.

[2] "[L]anguage should be analysed by the methodology of the natural sciences, and there is no room for constraints on linguistic inquiry beyond those typical of all scientific work." (Smith 2000: vii)
"Linguistics is not the only discipline nowadays in which intellectual leaders fail to respect traditional scholarly norms." (Sampson 2007b: 127)

turn linguistics into a "mature empirical science".[3] The following *general requirements* have been imposed and found wide acceptance among linguists:

(GR) (a) Theory formation (that is, generation of hypotheses) and testing of the theory have to be strictly separated.
(b) The hypotheses of empirical linguistic theories have to be connected by valid deductive inferences.
(c) Linguistic theories have to be free of inconsistencies.
(d) Data are immediately given and primary to the theory.
(e) The hypotheses of empirical linguistic theories have to be tested with the help of reliable data that can be regarded as facts constituting a firm and secure basis of research. Such data are called 'evidence'.

We will examine one of the strategies that have been proposed in order to fulfil (GR) and get rid of the inferiority complex in linguistics.[4] It is relatively new in this form and was put forward by, among others, Geeraerts (2006), Lehmann (2004) and Sampson (2007b). It contains, among others, the following principles (*special requirements*):

(SR) (a) Linguistics has to rely on evidence that is *intersubjectively controllable*. The objectivity of data can be secured by systematic and controlled *observation* such as psycholinguistic and neurolinguistic experiments, use of corpora, surveys, and fieldwork. Evidence consists of observation statements capturing different perceptible manifestations of linguistic behaviour.[5]
(b) Data gained by proper application of these methods can be treated as irrevisable *facts* within the given theory.[6]
(c) Linguistics has to apply *procedures* that relate higher-level abstractions or unobservable phenomena to evidence.[7]

---

[3] "[…] one of the major strivings of modern linguistics has been precisely that of meeting the requirements of an empirical science, namely one that is careful with data and sensitive to its nature." (Simone 2004: 246)

[4] There are, of course, several other views as well. The choice of the highlighted strategy is motivated by the circumstance that it is relatively elaborated and seems to be influential.

[5] "What makes a theory empirical is that it is answerable to interpersonally-observable data." (Sampson 2007b: 115)
"Empirical research is *data-driven*. You cannot easily draw conclusions from single cases and isolated observations, and the more data you can collect to study a particular phenomenon, the better your conclusions will get. The observations could come from many sources […]: you could collect them as they exist […], but you could also elicit them by doing experimental research, or by doing survey research […] (Geeraerts 2006: 23; emphasis as in the original)

[6] "[Something] may nevertheless function as a datum in some research that assigns it the role of unquestionable evidence in the argumentation." (Lehmann 2004: 181)
"[…] linguistics at large does not possess a common empirical ground, in the form of a set of observations derived through a generally accepted method, that plays the same role that experimentation does in psycholinguistics." (Geeraerts 2006: 26)

[7] "In general, for a datum to be accepted as such in the discipline, there must be operational procedures of relating secondary to primary data, and primary data to the ultimate substrate. Such procedures are part of the methodology of that discipline, viz. of the methods that *allow* scientists to *control* the relationship

(d) Linguistic hypotheses have to be *operationalised* which means that they should be appropriate for evaluation by quantitative methods.[8]

(e) Since data are hard facts, any conflict between them and hypotheses of the theory has to lead to the instant and automatic *falsification* of the theory.[9]

If we take a closer look at (GR) and (SR), we have to say that they can be *questioned* at several points:

(a) *(SR) stipulates criteria that are so strong that no linguistic theory is capable of fulfilling them*. First, (SR)(a) requires the elimination of subjectivity from linguistic research. Therefore, it sharply rejects the use of introspective data and wants to exclude linguistic intuition from the interpretation of data.[10] Nevertheless, as was shown in the current literature on linguistic data and evidence, neither work with corpora, nor experiments can be carried out and interpreted without the use of the linguist's linguistic intuition and without (to some extent) arbitrary (therefore, subjective) decisions.[11]

Second, as a consequence of the above, in opposition to (SR)(b), neither corpus data, data gained by experiments, nor introspective data can be regarded as perfectly reliable. One of the most important insights of the current literature on linguistic data and evidence is that all data

---

between the theory and the data. […] If there are no such operational procedures, then firstly there is no basis on which *the datum can be taken for granted*, which means that it is not a datum in the sense of our definition; and secondly, there is no way of relating a theory to a perceptible epistemic object, which means it is *not an empirical theory*." (Lehmann 2004: 185f.; emphasis added)

"[…] linguistics should primarily develop an independent *observational language* that the different *theoretical languages* of linguistics can be mapped onto […]." (Geeraerts 2006: 27; emphasis as in the original)

[8] "Empirical research involves *quantitative methods*. In order to get a good grip on the broad observational basis that you will start from, you need techniques to come to terms with the amount of material involved. […] Empirical research requires the *operationalization of hypotheses*. It is not sufficient to think up a plausible and intriguing hypothesis: you also have to formulate it in such a way that it can be put to the test. That is what is meant by "operationalization": turning your hypothesis into concrete data." (Geeraerts 2006: 24; emphasis as in the original)

[9] "To be truly scientific, a theory should make sufficiently strong claims that are open to rebuttal by experimentation or direct observation. This principle, most famously reduced to the single term falsifiability (e.g. see Popper 1959), is tightly woven into the practice of modern day linguistics […]." (Veale 2006: 466)

"[…] there is a common, commonly accepted way in psycholinguistics of settling theoretical disputes: experimentation. Given a number of conditions, experimental results decide between competing analyses, and psycholinguists predominantly accept the experimental paradigm as the cornerstone of their discipline. The conditions that need to be fulfilled to make the paradigm work are in principle simple: the experiment has to be adequately carried out, and it has to be properly designed in order to be distinctive with regard to the competing theories." (Geeraerts 2006: 26)

[10] "It is startling to find 20th- and 21st-century scientists maintaining that theories in any branch of science ought explicitly to be based on what people subjectively 'know' or 'intuit' to be the case, rather than on objective, interpersonally-observable data." (Sampson 2007a: 14)

"If linguistics is indeed based on intuition, then it is not a science […] Science relies exclusively on the empirical." (Sampson 1975: 60)

[11] Cf. Kertész & Rákosi (2008a, b, c, 2012); Schütze (1996); Lehmann (2004); Penke & Rosenbach (2004); Kepser & Reis (2005); Borsley (2005); Stefanowitsch & Gries (eds.)(2007); Sternefeld (ed.)(2007); Consten & Loll (2014).

types have to be assumed to be problematic, and they are inevitably highly theory- and problem-dependent. Although linguistic data cannot be treated as irrevisable facts, the everyday practice of linguistic research and the metascientific reflection of a genuinely wide group of linguists testify that all data types should be considered as legitimate (at least in principle), and can be used together, in combination, to make the results more reliable.[12]

Third, the means for fulfilling (SR)(c) are lacking: the connection between perceptible properties of linguistic behaviour ("observational terms") and the conceptual apparatus of the theory ("theoretical terms") is missing – and left to the (subjective) interpretation of linguists. Therefore, corpus linguists, linguists carrying out experiments, cognitive linguists etc. in most cases do not work with observable data but with (more or less abstract) theoretical constructs.

Fourth, (SR)(d) is only partly realised. It is highly doubtful whether quantitative methods can be applied in every field of linguistic research, or can be applied without also doing research using qualitative tools. There seem to be principled reasons for the failure of this requirement.

Fifth, the fallibility of linguistic data undermines the requirement of falsifiability as formulated in (SR)(e). In a conflict between data and the hypotheses of a theory, it is not clear which one should be given up.

These problems cast doubt on (GR) as well. The uncertainty, problem- and theory-dependence of linguistic data is irreconcilable with (GR)(a), (d) and (e). In opposition to (GR)(b), most linguistic theories do not have a deductive structure but they make use of several kinds of non-deductive inferences such as analogy, part-whole inference, induction etc. The application of several different data types and the uncertainty of the data leads to a higher possibility of the emergence of inconsistencies both between data and hypotheses and among the hypotheses of the theory as well, casting doubt upon (GR)(c).

(b) *There are other specifications of (GR) which are incompatible with (SR).* There is, among others, a second strategy that is significantly older, and is applied by many generative linguists. It is based on the use of introspective data and is an elaboration of (GR), too.[13] Although these

[12] For an overview, see Kertész & Rákosi (2008a, b, c, 2012).

[13] The parallelism between the norms of this strategy summarised as (SR') on the one hand and (SR) on the other hand is striking:

(SR') (a) Linguistics has to rely on evidence that is *intersubjectively controllable*. The objectivity of data can be secured by a special type of experiment, namely, with the help of collecting and observing grammaticality/acceptability judgements of native speakers (cf. e.g., Chomsky, 1965, p. 18; Chomsky 1969: 56).

(b) Since linguistic competence is supposed to be homogeneous within a language community (and eventual differences can be considered as performance errors), data gained by the proper application of this method can be treated as irrevisable *facts* within the given theory (cf. e.g., Chomsky 1969 [1957]: 13-16; Andor 2004: 98).

(c) Linguistics has to develop higher-level abstractions that, on the one hand, make it possible to make *testable predictions* and, on the other hand, enable us to formulate *general laws* of linguistic competence (Chomsky 1969 [1957]: 49-50).

(d) Linguistics has to elaborate an *evaluation procedure* that compares possible grammars, and determines which of them meets the criteria of external adequacy and generality (explanatory adequacy) to a greater extent (Chomsky 1969 [1957]: 49-60).

two strategies have the same goal, and share the same metascientific commitments, they sharply criticise and reject each other's views. The major difference lies in their concept of empiricalness: while adherents of (SR) accept only observation statements based on perception as evidence, followers of the generativist tradition use the term 'observation' and 'experiment' in a wider sense, or even abandon using the first term. Thus, they find introspective data perfectly acceptable – and do this with reference to (GR).[14]

(c) *(GR) and (SR) do not describe the practice of scientific theorising in natural sciences properly*. Neither (GR) nor (SR) stem from the study and thorough analysis of scientific research in physics, biology, medicine etc. but they adopt highly abstract tenets of the *standard view of the analytical philosophy of science*, initiated by the logical positivists in the 1920s. The elements of the standard view, however, have never been accepted methodological principles of natural sciences but remain alien to everyday research practice. As Machamer puts it,

> "[t]he logical positivists, though some of them had studied physics, had little influence on the practice of physics, though their criteria for an ideal science and their models for explanations did have substantial influence on the social sciences as they tried to model themselves on physics, i.e. on 'hard' science." (Machamer 2002: 12)

This discrepancy between "ideal" and "real" science has been recognised by philosophers of science since the 1960s. With the historical and sociological turn in the philosophy of science, the standard view of the analytical philosophy of science has become outdated.[15] Although its importance from the point of view of the history of the philosophy of science is, of course, indisputable, it no longer belongs to the mainstream trends of the philosophy of science. Therefore, the position of linguistics is highly anachronistic since it still greatly relies on a number

---

[14] Chomsky argued that introspective data – although they do not possess spatiotemporal coordinates – fulfil the function which (GR) requires from empirical evidence:

> "An experiment is called work with an informant, in which you design questions that you ask the informant to elicit data that will bear on the questions that you're investigating, and will seek *to provide evidence* that will help you answer these questions that are arising within a theoretical framework. Well, that's *the same kind of thing* they do in the physics department or the chemistry department or the biology department. To say that it's not empirical is to use the word 'empirical' in an extremely odd way." (Andor 2004: 98; emphasis added)

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

of obsolete elements that have already been eliminated from among the tools of the philosophers of science (cf. Kertész 2004; Kertész & Rákosi 2008a,b, 2012, 2014).

(d) *Contemporary philosophy of science rejects the idea of providing general, uniform norms for scientific theorising*. Therefore, only tentative hypotheses with more or less restricted scope can be formulated on the basis of detailed case studies focusing on different aspects of research practice in special fields of scientific theorising and from diverse historical periods.[16] As opposed to these insights, (SR) still tries to derive norms of linguistics from the alleged principles of scientific theorising in general.

At this point, of course, the question emerges of what linguists should do. Further insistence on (GR) and (SR) seems to be hopeless. Moreover, it is also doubtful whether any kind of reductionism is possible. Another option would be the fulfilment of Saussure's dream, that is, the elaboration of a new, specific methodology for linguistics. This strategy would be in accord with the recent stance of the philosophy of science as mentioned in (d) above. Despite this, it appears highly *risky*. First, the silent majority of linguists do not reflect on foundational issues systematically but, at best, occasionally. Second, there are no generally accepted methods (such as data handling techniques, strategies for the treatment of inconsistencies, tools for the evaluation and comparison of rival approaches) and standards (for example, what types of data and evidence are legitimate, when is a contradiction tolerable etc.). Therefore, it is not clear whether the enormous diversity of methods, theories and norms in linguistics makes any kind of generalisation, comparison and evaluation possible. This motivates raising the following series of problems, which belong to the most fundamental and thorny issues of cognitive linguistic research:

(GP) (a) How can the uncertainty of data be treated in cognitive linguistics?
(b) What are the methods of inconsistency resolution in cognitive linguistics?
(c) Which guidelines should govern the evaluation of theories in cognitive linguistics?

Due to the diversity of approaches and the methodological pluralism within cognitive linguistics research, we will not answer (GP) in general. Instead, we will narrow it down to a more

[16] "A consensus did emerge among philosophers of science. It was not a consensus that dealt with the concepts of science, but rather a consensus about the 'new' way in which philosophy of science must be done. Philosophers of science could no longer get along without knowing science and/or its history in considerable depth. They, hereafter, would have to work within science as actually practiced, and be able to discourse with practicing scientists about what was going on. […] The turn to science itself meant that philosophers not only had to learn science at a fairly high level, but actually had to be capable of thinking about (at least some) science in all its intricate detail. In some cases philosophers actually practiced science, usually theoretical or mathematical. This emphasis on the details of science led various practitioners into doing the philosophy of the special sciences. […] One interesting implication of this work in the specialized sciences is that many philosophers have clearly rejected any form of a science/philosophy dichotomy, and find it quite congenial to conceive of themselves as, at least in part of their work, 'theoretical' scientists. Their goal is to actually make clarifying and, sometimes, substantive changes in the theories and practices of the sciences they study." (Machamer 2002: 9ff.)

specific problem. One of the most important developments in cognitive linguistics is the acceptance and application of a wide range of data types. Experiments count as frequently used and extremely valuable data sources; experimental data are regarded as the prototypical empirical data type, the counterpart of experimental data used in the natural sciences. Therefore, it seems to be highly instructive to investigate its strengths, usefulness and limits, and compare them to those of introspective data, which are at the other end of a progressive-traditional spectrum. Nonetheless, the standards which could govern the experimental procedure and the evaluation of the results are missing. Against this background, (SP) presents itself as an interesting special case of (GP):

(SP) (a) How can the uncertainty of experimental data be treated in cognitive linguistics?
(b) What are the methods of the treatment of inconsistencies emerging from conflicting results of experiments in cognitive linguistics?
(c) Which guidelines should govern the evaluation of theories with respect to experimental results in cognitive linguistics?

The search for a solution of (SP) is a *metascientific* undertaking. Metascience is the inquiry of scientific research (scientific activities and/or products) by means of scientific methods. In our case, this means that we intend to systematically investigate the *nature and limits of cognitive linguistic experiments by applying the tools of argumentation theory, the philosophy of science and statistical meta-analysis*.

The three parts of the book will correspond to the three sub-problems of (SP). All parts are organized around a *paradox* and make use of instructive *case studies* to demonstrate the workability of the presented metascientific model.

Part I will be devoted to the *uncertainty of experimental data* in cognitive linguistics. While introspection is often criticized for being burdened by subjectivity (and rejected as an unreliable data source), experimental data are deemed, according to the established view in cognitive linguistics, to be a firm and objective base for the empirical study of linguistic behaviour. Yet there is a relatively new insight among linguists working with experimental data that measurements do not mirror linguistic stimuli directly (cf. Schlesewsky 2009: 170). Experiments involve several external factors such as the task environment, the capacity of memory, etc. whose impact on the results is not clear and thus compel us to deem experiments not perfectly reliable data sources. A related problem is that in order to the control the parameters that might influence the outcome of the experiment as strongly as possible, researchers have to construct artificial situations which therefore yield less natural data. These insights echo similar problems well-known in relation to experiments in the natural sciences. Therefore, it seems to be straightforward to search for analogies between experiments in cognitive linguistics and in science, and make use of (or adapt) metascientific-methodological tools elaborated by the philosophy of science in order to model experiments. If we try to fulfil this task, however, we encounter severe difficulties. Namely, according to the common view, the experimental report should be transparent in the sense that it should provide *direct access* to all relevant facets and details of the experimental procedure, avoid argumentative tools and convey only the "hard facts". This is, however, practically never the case with experiments in cognitive linguistics, because experimental papers usually present *a thematically more comprehensive and highly*

persuasive account of the experimental procedure, which embeds the experimental results into a coherent and general picture of the experimental process and the related cognitive theories. This picture is, however, *informatively deficient* in comparison to the requirements of total reconstructibility. This leads to the *rhetorical paradox of experiments in cognitive linguistics*:

(RPE) The reliability of experiments as data sources in cognitive linguistics is both directly *and* inversely proportional to the rhetoricity of the experimental report.

At this point, however, we have to face a problem: a solution to (RPE) – that is, the clarification of the role of argumentation in the experimental reports – seems to be the prerequisite for the reconstruction and evaluation of experiments. On the other hand, (RPE) cannot be resolved without our being able to judge the reliability of experiments as data sources independently of the experimental report. We will argue in Part I that the *circularity* between the resolution of the paradox and the modelling and evaluation of experiments is apparent, because it is possible to develop *an argumentation theoretical model of experiments* which allows us to describe and re-evaluate the role and impact of argumentation both in the experimental procedure and in the experimental report, as well as the relationship of these two pieces of argumentation to each other and to other components of experiments. The reliability of experiments as data sources, however, depends not only on their inner structure but also on their relationship to related experiments. Therefore, we have to extend our model to larger units, i.e. sequences of similar experiments ('experimental complexes').

Part II deals with the emergence, function and treatment of *inconsistencies related to experimental data* in cognitive linguistics. One of the major sources of inconsistencies in experimental research are exact and non-exact replications (unaltered or modified replications of the same experiment), as well as methodological variants (experiments which investigate the relationship between the same variables with different methods). Such experiments are conducted in order to increase the experiments' reliability and validity. Harmony in the results is then interpreted as indicating stability and the absence of problems which were suspected to burden the original experiment. The new version's outcome, however, often contradicts the results of the original experiment. Moreover, it often happens that while one problem is solved, new problems also arise which burden the modified version of the original experiment. This yields the Paradox of Problem-Solving Efficacy:

(PPSE) Non-exact explications and methodological variants are
- (a) *effective tools of problem-solving* in cognitive linguistics because by resolving problems they lead to more plausible experimental results; and they are also
- (b) *ineffective tools of problem-solving* because they trigger cumulative contradictions among different replications and methodological variants of an experiment and lead to the emergence of new problems.

A key point in relation to the resolution of (PPSE) is the elaboration of a metascientific tool which enables us to reconstruct the inconsistencies and other problems, identify their causes and function, assess their possible treatment, and evaluate the progress of the problem-solving

process. As we will see, not all inconsistencies are equal; accordingly, the function they fulfil as well as their treatment is different, too.

Part III focuses on the often problematic *relationship between theories and experimental data* in cognitive linguistics. There are no clear and easily applicable criteria for determining the strength of support an experiment or a series of experiments provides for or against a theory. We have to take into consideration the peculiarities of how predictions can be drawn from a theory, as well as how the results of a series of experiments can be summarised and compared to the predictions of rival theories. At this point, however, we have to face the *Paradox of Error Tolerance*:

(PET) When determining the strength of support provided by an experimental complex to a hypothesis/theory,
  (a) *the elimination of errors is a top priority*, because it is the detection and elimination of problems which makes experiments more reliable data sources;
  (b) *the elimination of errors is not a top priority*, because comprehensibility, that is, the involvement of all relevant experiments and the accumulation of all available pieces of information should be ranked higher.

We will argue that both methods can be useful, and they can be applied in parallel, because their results can complement and control each other.

All three parts of the book make extensive use of the *method of case studies*. Case studies can be seen as tools of the *empirical testing* of the hypotheses raised by the philosophy of science, and more generally, the *naturalization of philosophy of science*.[17] They make it possible to carry out detailed, fine-grained analyses, and confront metascientific-philosophical ideas with the realities of research practice. Nonetheless, their application raises some doubt. First, the question is whether it is allowed to *generalise the results* of single cases. A second concern is *selection bias*: how to justify the choice of the cases. Pitt (2001) conjoins these two problems to the *Dilemma of Case Studies*:

> "On the one hand, if the case is selected because it exemplifies the philosophical point being articulated, then it is not clear that the philosophical claims have been supported, because it could be argued that the historical data was manipulated to fit the point. On the other hand, if one starts with a case study, it is not clear where to go from there – for it is unreasonable to generalize one case or even two or three." (Pitt 2001: 373)

Several solutions have been proposed and discussed for this dilemma. One important idea raised by Chang (2011) pertains to the relationship between case studies and the philosophy of science: it should not be viewed as a hierarchical relationship between the particular (single cases) and the general (comprehensive theory) but as a *cyclic relationship between the concrete and the abstract*. Clearly, multiple returns from case studies to the metascientific model and back again make it possible to *gradually modify one's hypotheses as well as interpretations of*

[17] Cf. Giere (2011: 60f.), Scholl & Räz (2016: 72ff.).

concrete cases. A second significant point is that the selection of the case studies should be carefully based on a *set of criteria laid down in advance*. A list presented by Scholl & Räz (2016: 77ff.) contains four methods for preventing selection bias:

1. Hard cases:	Instead of selecting cases which illustrate the theory nicely, one may choose challenging ones which at first sight seem to refute the philosophical theory at issue; thus they can be really good tests of the theory.
2. Paradigm cases:	One may select cases which are regarded as typical instances in the given research field. Thus, the choice is not governed by the researcher's points of view, and generalization from a few instances is well-founded, too.
3. Big cases:	Famous, well-known cases which were decisive for the development of the research field at issue are interesting objects of case studies, too, although generalizability may be a problem.
4. Randomized cases:	In possession of a database of cases, randomization offers a widely accepted way of avoiding selection bias. Moreover, due to the variegation, representativeness, and via this, generalizability is secured, too.

The selection of the experiments for the case studies followed none of the above methods thoroughly but was governed by other motives. Above all, narrowing down the topic to experiments on metaphor processing resulted from *practical considerations*. Namely, within cognitive linguistics, this research field has the longest tradition with experimentation. This is the only research area in which it was possible to find a sufficient number of high-quality experiments. Within the realm of experiments on metaphor processing, the main idea was *representativeness*: I strived to find experiments which were conducted by the leading figures of the three most influential theories on metaphor processing, or which tested these approaches: Lakoff and Johnson's Conceptual Metaphor Theory (CMT), Gentner's Structure Mapping Theory (SMT) and its successor, the Career of Metaphor Hypothesis (CMH) and Glucksberg's Attributive Categorization View (ACV) and its refined version, the Interactive Property Attribution Model (IPAM).

The starting point of the metascientific model of cognitive linguistic experiments put forward in Parts I-III is the p-model elaborated by András Kertész and Csilla Rákosi (Kertész & Rákosi 2012, 2014). The application of the p-model to experiments and its extension to series of experiments, to the problem of inconsistency related to experiments and to the relationship between cognitive linguistic theories and experimental evidence was published in Rákosi (2011a, b), Rákosi (2012), Rákosi (2014), Rákosi (2016a, b), Rákosi (2017a, b), Rákosi (2018a, b).

PART I. THE TREATMENT OF THE UNCERTAINTY OF EXPERIMENTAL DATA IN COGNITIVE LINGUISTICS

2. Introduction: The rhetorical paradox of experiments (RPE) in cognitive linguistics

According to Geeraerts' diagnosis, one important step that cognitive linguistics should take in order for the field to reach the status of a scientific enterprise, is the application of empirical methods used successfully within other branches of cognitive science:

> "Cognitive Linguistics, if we may believe the name, is a *cognitive science*, i.e. it is one of those scientific disciplines that study the mind […]. It would seem obvious then, that the methods that have proved their value in the cognitive sciences at large have a strong position in Cognitive Linguistics: the experimental techniques of psychology, computer modelling, and neurophysiologic research." (Geeraerts 2006: 28; emphasis as in the original)

Thus, the recent development, namely that reference to experiments is regarded as one of the most powerful tools in argumentations in favour of, or against, cognitive linguistic theories, might be interpreted in such a way that in cognitive linguistics, similarly to psychology and other cognitive sciences, the idea of treating experimental results as strong evidence for, or against, theories is prevalent:

> "[...] there is a common, commonly accepted way in psycholinguistics of settling theoretical disputes: experimentation. Given a number of conditions, experimental results decide between competing analyses, and psycholinguists predominantly accept the experimental paradigm as the cornerstone of their discipline." (Geeraerts 2006: 26)

This authority is usually based on the view that experiments allow for confronting hypotheses of theories *directly* with empirical evidence. In this vein, experiments have to be objective and intersubjectively controllable, and apply feasible, well-established procedures providing completely reliable experimental data:

> "The conditions that need to be fulfilled to make the paradigm work are in principle simple: the experiment has to be adequately carried out, and it has to be properly designed in order to be distinctive with regard to the competing theories. That is to say, you need good experimental training (knowledge of techniques and analytical tools), and you need the ability to define a relevant experimental design. The bulk of the effort in psycholinguistic research, in other words, involves attending to these two conditions: setting up adequate designs, and carrying out the design while paying due caution to experimental validity." (Geeraerts 2006: 26)

The experimental report has to transmit these characteristics and must not make use of rhetorical tools aimed merely at persuading the reader. From this it follows that the reliability of experiments is supposed to be *inversely proportional* to the rhetoricity of the experimental report.

If, however, we take a closer look at papers dealing with experiments in cognitive metaphor research, we never actually see the "raw" (numerical) data capturing some observation of linguistic behaviour and a chain of deductively valid inferences leading to the result of the experiment and the latter's confrontation with some hypotheses or theories (i.e. confirmation or falsification). Instead, a typical experimental report seems to be a highly complex argumen-

tation process which is not deductive. It usually contains, among other things, the following components (not necessarily in this order):

– main tenets, explanatory power, and other strengths of the preferred theory;
– central hypotheses and weak points of the rival theories;
– description of a phenomenon in connection with which the theory and its rivals propose different predictions;
– motivation and description of the experiment to be conducted and conjectures about its outcome;
– details and shortcomings of earlier similar experiments;
– description and results of control experiments aimed at ruling out some known possible systematic errors;
– no "raw data" (individual measurements) at all;
– some excerpts from the stimulus material used;
– type and upshot of statistical analyses;
– presentation of considerations concerning the interpretation and reliability of the results;
– if there seem to be shortcomings in the experiment, then a second experiment is proposed, carried out and its results are analysed, too;
– the impact of the conducted experiment on the theory at hand and its rivals;
– proposals for further inquiry in the given topic etc.

It is plain to see that the relationship between the "raw data" (that is, the complete set of individual measurements) and hypotheses of the linguistic theory or theories at issue cannot be reconstructed on the basis of the information provided in the experimental report. Consequently, far from being direct and transparent this relationship is quite fragmentary. Despite this, it is the experimental report on the basis of which one decides whether the given experiment is a reliable data source. Compelling, lucid and reasonable experimental reports are regarded as indications of good, reliable experiments and, conversely, poor, shaky, faulty experimental reports also lead to the rejection of the experiment itself. Therefore, the authority of experiments does not stem from an impersonal and straightforward linkage between "empirical facts" and hypotheses. Rather, it seems to depend crucially on the peculiarities and plausibility of the argumentation put forward in the experimental report, on its persuasiveness and its convincing force. From this we obtain that the reliability of experiments is *directly proportional* to the rhetoricity of the experimental report.

Thus, our considerations have led to a paradox:

(RPE) *The rhetorical paradox of experiments in cognitive linguistics*:
The reliability of experiments as data sources in cognitive linguistics is both directly *and* inversely proportional to the rhetoricity of the experimental report.

If we examine the two contradictory members of (RPE), two promising starting points present themselves:

1) While the view which considers rhetorical tools unnecessary and worthless is a *methodological rule*, the opposite view refers to the *practice* of linguistic research. This should motivate us to raise the question of whether the first view is an adequate norm, and, if not, then whether and to what extent is the practice of presenting the results of experiments in cognitive linguistics acceptable? That is, the criteria for judging the rhetoricity of experiments should be revealed.
2) The two contradictory views use the concept 'rhetoric' in different senses. The first view reduces it to *irrational tricks and manoeuvres*, erroneously claiming that the experiment provides reliable results. In sharp contrast to this, the second view allows room for interpreting 'rhetoric' as *rational argumentation* that may be fully legitimate and should be an important constituent of scientific experiments.

Nonetheless, the impact of argumentative tools on the reliability of experiments as data sources in cognitive linguistics can only be judged *in the context of all factors which might influence the reliability of experiments as data sources*. Therefore, Section 3 will be devoted to the elaboration of a metascientific tool which allows us to reconstruct the structure of experiments in cognitive linguistics. In Section 4, we will try to reveal the nature and function of argumentation in experiments. In possession of a meta-scientific model of experiments which clarifies how the structure of experiments in cognitive science can be reconstructed and how their components influence the reliability of experiments, in Section 5 we will show how the reliability of single experiments as data sources can be determined and re-evaluated in cognitive linguistics. Experiments, however, have not only a private but also a social life. Consequently, the uncertainty of experiments originates not solely from their design, conduct, the interpretation of their results, etc., but also from the harmony of their results with other experiments. This means two things. Firstly, an analysis of the inner life of experiments may check their validity and reveal the problems which burden them and might prevent an experiment from producing plausible experimental data. It does not reveal, however, whether an experiment is reliable in a more restricted (traditional) sense of the term; that is, whether exact replications would produce similar results. This can be decided only by conducting a series of replications. Secondly, successful non-exact replications motivated by problems related to the original experiment (such as concerns about its validity) may also increase the experiment's reliability, if there is harmony between their corresponding results. Thus, Section 6 will be devoted to the social life of experiments. It presents a metascientific model of the multifaceted relations between closely related experiments (exact and non-exact replications, control experiments, counter-experiments), and discusses how the reliability of the related experiments as data sources can be determined. The workability of the metascientific model presented in Part I will be illustrated with the help of several small-scale case studies related to different theories of metaphor processing and to different timepoints; therefore, they can be regarded as representative of this research field. In Section 7, we will summarise our results and put forward a possible resolution to (RPE).

3. Metascientific modelling of experiments as data sources in cognitive linguistics

In order to elaborate a feasible metascientific model of the inner life of experiments in cognitive linguistics, we will first set forth a summary of the current state of the art on experiments in the natural sciences in the philosophy of science. Then, we will present a case study which investigates whether and to what extent there is an analogy between experiments in natural sciences and in cognitive science. On the basis of our findings, we will put forward a metascientific model aimed at describing the structure and workings of experiments in cognitive linguistics.

3.1. Recent views on the nature and limits of experiments in the natural sciences

James Bogen characterises experiments as follows:

> "In experiments, natural or artificial systems are studied in artificial settings designed to enable the investigators to manipulate, monitor, and record their workings, shielded, as much as possible from extraneous influences which would interfere with the production of epistemically useful data." (Bogen 2002: 129)

This quotation indicates that physical experiments are remarkably complex entities. They comprise several ontologically diverse components such as:

- *experimental design*: a comprehensive preliminary description of all facets of the process of experimentation;
- *experimental procedure*: a material procedure where an experimental apparatus is set up, its working is monitored and recorded under controlled circumstances, that is, in an *experimental setting*;
- *a theoretical model of the phenomena investigated*: one has to have at least a rough idea of what one intends to investigate. The problem which the experimenter raises is usually related to one or more imperceptible, low-level theoretical construct(s) (phenomena)[18] that may be relevant in judging hypotheses about high-level theoretical constructs or require theoretical explanation. A detailed theoretical account of the given phenomenon is needed only if the experiment aims at testing hypotheses of a given theory or theories. Previous conceptions can be modified;
- *a theoretical model of the experimental apparatus*: One has to understand the functioning of the apparatus applied insofar as one has to possess explanations about how phenomena are created or separated from the background, which of their properties can be detected with the help of the equipment, and why it can be supposed that the perceptual data produced by the apparatus are stable and reliable. One has to have ideas in advance about which phenomena can be investigated with the help of the experimental apparatus, how perceptual data resulting from the use of the apparatus are related to these phenomena,

[18] For example: the atomic mass of silicone, neutron currents, recessive epistasis, Broca's aphasia.

what the potential sources of "noise" (background effects, idiosyncratic artefacts and other kinds of distorting factors) are, and how they can be ruled out;

- *perceptual data*: data gained by sense perception such as smell, taste, colour, photographs, and, above all, readings of the measurement apparatus, etc.
- *authentication of perceptual data*: The experimenter has to evaluate the outcome of the experimental procedure. He/she has to decide whether the experimental apparatus has been working properly so that perceptual data are stable and reliable; he/she has to check whether sources of noise have been ruled out, or at least their effect can be eliminated with the help of statistical methods;
- *interpretation of perceptual data*: the experimenter has to establish a connection between the perceptual data gained and the phenomena investigated. It has to be decided whether the former are relevant, real and reliable in relation to the latter,[19] and it has to be spelled out what conclusions can be drawn from the former: the perceptual data indicate the presence of the given phenomenon, they indicate its absence, or they require the modification of its supposed properties etc.
- *presentation of experimental results*: since experiments are not private but public affairs aimed at supplying data for scientific theorising, it is not only the results of the experiment which have to be put forward; so must every element of the experimental procedure that is judged *relevant* to the evaluation and acknowledgement of the results. Therefore, the experimenters have to present an *argumentation* that conforms to certain *norms*. It should contain all information that may have any significance when the scientific community have to decide whether the experimental results are reliable and epistemologically useful, that is, whether they can be used for theory testing, explanation, elaboration of new theories etc. To this end, relevant pieces of information have to be selected and arranged into a well-built chain of arguments leading from the previous problems raised through the description of the experimental design and the experimental procedure to the evaluation (authentication and interpretation) of data. Thus, experimental data should be *suitable for integration into the process of scientific theorising*. This subsequent operation may consist either of establishing a link between the experimental data and existing theories of the phenomena at issue (the result of this process may be an explanation of the experimental data, or an analysis of the conflicts between existing theories and the data), and/or presenting a new theory which might be capable of providing an explanation for them.

This brief sketch allows us to reflect upon properties of experiments that are of central importance according to the current literature:

(a) Contrary to the tenets of the standard view of the analytical philosophy of science, experiments cannot be regarded as "black boxes which outputted observation sentences in relatively mysterious ways of next to no philosophical interest" (Bogen 2002: 132). Rather, experiments involve a highly complex network of different kinds of activities, physical objects, argumentation processes, interpretative techniques, background knowledge, methods, norms, etc. which *raise several serious epistemological questions*. The analysis and evaluation of ex-

[19] Cf. Bogen (2002: 135).

periments cannot be reduced to the examination of the end products of the experimentation process – the *whole process* has to be taken into consideration.

(b) Although observation is a necessary component of scientific experiments, *its role is much more modest* than supposed by the standard view. What is perceived is only readings of the experimental apparatus, the smell of a liquid, a photograph taken with the help of a microscope etc. but not the phenomena the researcher is interested in themselves:

> "[…] many different sorts of causal factors play a role in the production of any given bit of data, and the characteristics of such items are heavily dependent on the peculiarities of the particular experimental design, detection device, or data-gathering procedures an investigator employs. Data are, as we shall say, idiosyncratic to particular experimental contexts, and typically cannot occur outside of those contexts. Indeed, the factors involved in the production of data will often be so disparate and numerous, and the details of their interactions so complex, that it will not be possible to construct a theory that would allow us to predict their occurrence or trace in detail how they combine to produce particular items of data. Phenomena, by contrast, are not idiosyncratic to specific experimental contexts. We expect phenomena to have stable, repeatable characteristics which will be detectable by means of a variety of different procedures, which may yield quite different kinds of data." (Bogen & Woodward 1988: 317)

What the researcher intends to give an explanation for is not the outcome of the individual measurements (thus, he/she does not try to explain why he/she read on the display a value of 5.628 at the first measurement and 5.649 at the second etc.) but the link between the results of a series of measurements (a set of perceptual data) and the expected phenomenon.[20] A prerequisite of this is the *authentication* of perceptual data:

> "Noting and reporting of dials – Oxford philosophy's picture of experiment – is nothing. Another kind of observation is what counts: the uncanny ability to pick out what is odd, wrong, instructive or distorted in the antics of one's equipment." (Hacking 1983: 230)

Individual measurements are always influenced by measurement errors. While *random errors* are unpredictable but with statistical methods controllable, *systematic errors* systematically

---

[20] "[…] what we observe are the various particular thermometer readings – the scatter of individual data-points. The mean of these, on which the value for the melting point of lead […] will be based, does not represent a property of any particular data-point. Indeed, there is no reason why *any* observed reading must exactly coincide with this mean value. Moreover, while the mean of the observed measurements has various properties which will […] make it a good estimate of the true value of the melting point, it will not, unless we are lucky, coincide exactly with that value. […] So while the true melting point is certainly *inferred* or *estimated* from observed data, on the basis of a theory of statistical inference and various other assumptions, the sentence 'lead melts at 327.5 ± 0.1 degrees C' – the form that a report of an experimental determination of the melting point of lead might take – does not literally describe what is perceived or observed. […] what a theorist will try to explain is why the true melting point of lead is 327 degrees C. But we need to distinguish […] between this potential explanandum, which is a fact about a phenomenon on our usage, and the data which constitute evidence for this explanandum and which are observed, but which are not themselves potential objects of explanation. It is easy to see that a theory of molecular structure which explains why the melting point of lead is approximately 327 degrees could not possibly explain why the actual data-points occurred. The outcome of any given application of a thermometer to a lead sample depends not only on the melting point of lead, but also on its purity, on the workings of the thermometer, on the way in which it was applied and read, on interactions between the initial temperature of the thermometer and that of the sample, and a variety of other background conditions." (Bogen & Woodward 1988: 308f.; emphasis as in the original)

distort the results. It is very difficult to reveal their presence because they bias every single measurement in the same way, in the same direction and to the same extent. Therefore, they usually cannot be detected by the repetition of the measurement procedure and their effect cannot be eliminated by statistical means. They can be identified only with the help of another apparatus, by an experiment of different type, or by comparison with calculations based on theoretical considerations.

(c) From this it follows that experimental data cannot be equated with perceptual data; the latter are only one of the components of the former. *Experimental data* are not statements about individual observations but about the *link* between a set of observations and phenomena.[21] What lies between them, is the authentication and interpretation of perceptual data. This process is neither an induction from data to a hypothesis nor a deduction from a hypothesis to the data. Instead, it is a cyclic process where the perceptual data are examined, revised, statistically evaluated and brought into relationship with the phenomena investigated.

Since perceptual data are only a list of numerals, a photograph, a smell, a picture seen by looking through a telescope, etc., they have to be *interpreted*. That is, a relationship has to be established to a phenomenon. Phenomena are (low or high level) theoretical constructs. Therefore, researchers with different background knowledge or of different theoretical persuasion may look for different phenomena and with this, for different perceptual data. It may also happen that they judge different aspects of phenomena relevant, or interpret the perceptual data differently insofar as they may find them indicating different phenomena. Consequently,

> "[…] the salience and availability of empirical evidence can be heavily influenced by the investigator's theoretical and ideological commitments, and by factors which are idiosyncratic to the education and training, and research practices which vary with, and within different disciplines." (Bogen 2002: 141)

(d) Although perceptual data may be true with certainty insofar as the researcher may be totally sure that he/she has seen the digit 12.085 on the reader of the experimental apparatus, *experimental data cannot be regarded as certainly true*. First, experimental data are always underdetermined by perceptual data. Although it may be reasonable to think that the phenomenon supposed to be present is one of the causes of the results of the experiment (or vice versa, it may be plausible that the perceptual data indicate the presence of the given phenomenon), the chain of inferences between them is not conclusive and leaves room for other possible interpretations. Second, the resulting explanation does not account for idiosyncratic and unpredictable random errors (which usually remain unidentified) but tries to eliminate their influence; moreover, it may be misguided by systematic errors. Third, as we have seen in (c), the interpretation of perceptual data is *theory-dependent*. Fourth, experimentation is also *practice-dependent* in the sense that the experimental apparatus applied allows for a limited detection of the properties of the investigated phenomenon, and the abilities and skill of the researchers performing experiments may also differ.

---

[21] For example, the statement "The mass spectrometer *X* has shown a value of 27.976 926 532 46 at the first measurement." is a perceptual datum; the statement "The atomic mass of silicone is 28.0854 according to the mass spectrometer *X*" is an experimental datum which comprises a series of measurements and presupposes the authentication and interpretation of the perceptual data.

(e) The experimental design is always necessarily only *partial* in the sense that the researcher cannot identify and rule out in advance all potential sources of error that can bias the outcome of the experiment. Moreover, neither is the repeated experimentation process capable of yielding ultimate and unquestionable results. This means that both the authentication and the interpretation of the data are necessarily partial, too: one cannot be sure that no systematic errors occurred during the experiment; similarly, one cannot be sure that there are no other alternative interpretations and explanations of the perceptual data:

> "Three elements are conjoined in the production of any experimental fact: a *material procedure*, an *instrumental model* and a *phenomenal model*. […] […] in a typical passage of experimental activity, there is *no* apparent relation between the three elements. Incoherence and uncertainty are the hallmarks of experiment, as reported in ethnographic studies of laboratory life. But, at the moment of fact-production, their relation is one of *coherence*. Material procedures and instrumental and phenomenal models hang together and reinforce one another. […] But, following up my remarks that uncertainty is endemic to experimental practice, I want to say that such coherence is itself highly nontrivial." (Pickering 1989: 276ff.; emphasis as in the original)

Therefore, experiments are *open processes* in the sense that, in possession of new pieces of information, they may be continued, modified, or even discarded.

(f) There are no general criteria that would incontestably decide on the acceptability of the outcome of an experiment. Collins formulates this problem as the *experimenter's regress*:

> "What the correct outcome is depends upon whether there are gravity waves hitting the Earth in detectable fluxes. To find this out we build a good gravity wave detector and have a look. But we won't know if we have built a good detector until we have tried it and obtained the correct outcome! But we don't know what the correct outcome is until … and so on *ad infinitum*." (Collins 1985: 84; emphasis as in the original)

The experimenter's regress is mostly broken by referring to *socially accepted norms*. As Kuhn has pointed out, explicit or even only implicitly accepted but in praxis often applied norms determine to a considerable extent what happens in "normal science": paradigms guide the research by prescribing, among other things, how to validate perceptual data. This strategy has, of course, not only advantages but also risks because it may lead to *circularity*.[22] To reduce this danger, Franklin (2002: 3ff., 2009) proposes the following strategies:

- experimental checks and calibration, in which the experimental apparatus reproduces known phenomena;
- reproducing artefacts that are known in advance to be present;
- elimination of plausible sources of error and alternative explanations of the result (the Sherlock Holmes strategy);
- using the results themselves to argue for their validity;[23]

---

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

[23] This strategy is based on the argument that it is highly implausible that malfunction of the experimental apparatus or some background effect could lead to results that fit theoretical predictions to a great extent.

– using an independently well-corroborated theory of the phenomena to explain the results;
– using an apparatus based on a well-corroborated theory;
– using statistical arguments.

Although, as he remarks, "[n]o single one of them, or fixed combination of them, guarantees the validity of an experimental result", they *considerably increase its plausibility*. This also means that the acceptance of experimental results unavoidably contains subjective elements as well, since the comprehensiveness of the validating process of the results cannot be achieved. At certain points, one has to make decisions that remain necessarily *arbitrary* to some extent:

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

(g) Consequently, experiments do not provide us with epistemologically decisive results. They do not lead to certainly true observation statements; therefore, they neither verify nor falsify theories. Rather, their results are only more or less reliable; they are *fallible* and may strengthen or weaken hypotheses of theories to some extent. Despite this, they are *indispensable* tools of scientific theorising.
(h) The presentation of the results of the experiment not only leads to a concise and coherent report on the experiment but also conceals several details of the experimentation process. Therefore, it replaces the original, real event with an edited, selective, informationally reduced picture. As Geoffrey Cantor points out, there is usually a great distance between laboratory notebooks for private usage of the researchers and public reports:

> "Such notebooks not only provide far more detailed accounts of experimental procedures but also indicate the failures, errors and false starts that are not reported in public and those numerous particulars that are deemed unnecessary in a publication.
>
> Yet extant laboratory notebooks also sometimes indicate more interesting mismatches between laboratory practice and published reports. Holton, for example, has drawn attention to Robert Millikan's selection of acceptable results for his oil-drop experiment. During one series of experiments Millikan omitted well over half of his results, retaining data from only 58 drops out of a total of about 140." (Cantor 1989: 159)

Thus, there is a danger that the researcher eliminates relevant information from the published report and important decisions remain outside public control. In research reports, rhetorical tools dominate, since such texts aim to persuade the scientific community of the reliability and relevance of the experimental data gained. This argumentative character of experimental reports is especially salient in didactic contexts. The edition and purification of the raw data and several facets of experiments may lead to the emergence of scientific myths, leading to a *false self-image*:

> "One important function performed by textbooks (and not only textbooks) is to convey the values of the scientific enterprise. […] Such accounts of experiments are deceptive since they appear to deal with reality – both historical reality and the real structure of the physical world. Yet, like all myths and even dreams they are very condensed, invariably glossing over the numerous difficulties (often the immense difficulties) which arose during the construction of the experiment (except to evoke the reader's awe). Likewise, controversy over the experiment and its interpretation are usually suppressed. In the resulting discourse experiments

> emerge as very persuasive devices. They tell the reader the way things are and inculcate the kind of empiricism which philosophers of science have been at pains to undermine." (Cantor 1989: 166)

Thus, one of the most urgent tasks of the philosophy of science is to study the argumentative tools applied in published reports, as well as to find out how to determine which details of the experimental process should be regarded as potentially relevant and which can be omitted without loss of relevant information.

(i) There are manifold connections between theories and experiments. Experiments are not always means of theory testing but they may indicate the existence of phenomena that call for explanation and thus, motivate the elaboration of new theories without relying on some existing theoretical framework of the phenomena discovered.[24] On the other hand, experiments are in several respects *theory-dependent*. First, the design of an experiment involves a theory of the experimental devices applied. Second, the phenomenon investigated has to be explained by a theory. Third, theoretical considerations from diverse disciplines are active in the creation of the link between perceptual data and hypotheses about the phenomena investigated such as statistical tools, models of the background phenomena, an optical theory, investigation of other possible interpretations, calculation of the effects of known distorting factors etc. These considerations may overlap with the given theory aiming at the explanation of the phenomenon investigated to different extents. From this it follows that experimental data are always theory-laden, but this theory-ladenness may concern high-level (that is, very abstract) and specific hypotheses of the given theory, or may be related to rather low-level and non-specific hypotheses. In the latter case, the experimental datum may contribute to the decision between rival theories. Furthermore, even in the case of an overlap between the theory of the phenomena and the other theoretical considerations mentioned, the experimental data are always partially independent from the theory of the investigated phenomenon. Therefore, there may be a *conflict* between the data and hypotheses – that is, experimental data are capable of contradicting theoretical considerations.

According to the current literature on experiments, perceptual data and experimental data, as well as experimental data and hypotheses of theories are usually not connected by deductive inferences:

---

[24] "Many experiments are performed without the guidance of an articulated theoretical framework and aim to discover and explore new phenomena. If by 'theory' we mean a developed and articulated body of knowledge, then the history of science abounds in examples of pre-theoretical observations and experiments. For instance, many electrical phenomena were discovered in the eighteenth century by experiments which had not been guided by any developed theory of electricity. The systematic attempts to detect and stabilize those phenomena were part and parcel of their conceptualization and theoretical understanding […].

To investigate the relationship between experiment and theory one should take into account that 'theory' has a wide scope, extending from vague qualitative hypotheses to precise mathematical constructs. These different kinds of theory influence experimental practice in different ways. A *desideratum* in the philosophy of experiment is to understand the role of various levels of theoretical commitment in the design and implementation of experiments. It is clear, for instance, that theoretical beliefs often help experimentalists to isolate the phenomena they investigate from the ever-present 'noise' and 'provide essential . . . constraints on acceptable data' (Galison 1987: 73)." (Arabatzis 2008: 165f.; emphasis as in the original)

> "[…] often the derivations involve approximations and simplifications and so are not purely deductive. The derivations make use of additional premises, among which are previously established laws, principles, and theoretical results." (Nickles 1989: 307)

From (a)-(i) it follows that *the current literature on experiments sharply rejects the tenets of the standard view of the analytical philosophy of science*. Instead of evaluating only the results of experiments on the basis of abstract philosophical principles alien to everyday research practice, all authors argue for the relevance of every minor detail of the experimentation process. They do not strive for idealised and unrealisable norms but try to reveal the complexity and fallibility of experiments and to find out what difference good and bad praxis makes with the help of detailed case studies, that is, by *studying real experiments*:

> "As a knowledge-producing activity, experiment engages the inchoate, the practical, and the particular. The disorderly, inchoate, and personal character of scientific discovery and the complexity of experimental work needed to elicit meaning from phenomenological disorder have persuaded many that there is nothing philosophically interesting to recover […]. Thus, creative, exploratory, and constructive aspects of experimentation are largely neglected by philosophers of science. Disdain for mundane practice is an obstacle to philosophical understanding of how a language – and the arguments formulated in it – comes to grips both with a material, phenomenologically complex world and with the intellectual and social world of scientists, who are the primary audience for such arguments." (Gooding 2000: 122f.)

At this point, of course, the question emerges whether metascientific reflection on experimentation makes any sense, since, as Galison puts it,

> "The world is far too complex to be parceled into a finite list of all possible backgrounds. Consequently there is no *strictly logical* termination point inherent in the experimental sciences. Nor, given the heterogeneous contexts of experimentation, does it seem productive to search after a universal formula of discovery, or an after-the-fact reconstruction based on an inductive logic." (Galison 1987: 3; emphasis as in the original)

The answer is affirmative. The key point is to *change our view*: experiments should not be conceived of as "fabulous engines harvesting empirical evidence through observation and experimentation, discarding subjective, error ridden chaff, and delivering objective, veridical residues from which to spin threads of knowledge", as Bogen (2002: 128) says. Instead, they should be viewed as a *search for the best fit achievable* between the experimental design, the theory of the experimental apparatus, the process of experimentation, the perceptual data gained, the authentication and interpretation of the latter, the theory of the phenomenon investigated, etc. To find this fit, one has in most cases to turn back to earlier stages of the experimentation process and modify some component. Every component can be revised and the revisions have to be repeated again and again till there is *mutual support* among the constituents:[25]

> "Stable laboratory science arises when theories and laboratory equipment evolve in such a way that they match each other and are mutually self-vindicating." (Hacking 1992: 56)

[25] See also Pickering (1989).

This way of breaking out from the experimenter's regress involves, as we have already mentioned, the *risk of circularity* and may lead to the *experimenter's circle*. This is a real danger, and there are no formal or in every situation mechanically applicable criteria that would allow us to decide whether there is circularity or not.

3.2. Case study 1: Possible analogies between experiments in physics and in cognitive science

In the previous section, we have seen that on the basis of the study of the *praxis* of physical experiments, the current metascientific literature flatly rejects the views summarised in (GR):

– They argue for the inseparability of theory formation (context of discovery) and testing of the theory (context of justification).
– They give up the requirement of strict deductivity. Data and hypotheses about related phenomena, about the link between data and phenomena, among higher-level theoretical hypotheses etc. are seen as being connected by non-deductive inferences.
– Reliability of evidence is not equated with truth and certainty but with truth-candidacy or plausibility.
– Data are regarded as being created and theory-dependent (at least to some extent).

The next task is to find out whether the same holds true of cognitive linguistics as well. Therefore, with the help of a case study about an experiment carried out by Keysar, Shen, Glucksberg and Horton we will examine whether there is an *analogy* between experiments in physics and in cognitive linguistics. In Section 3.2.1, we will present the sketch of the experimental design of Keysar et al. (2000). Then, in Sections 3.2.2-3.2.8, we will try to identify the components found by physical experiments and find out whether these components are burdened by similar epistemological problems. It is important to remark that – as we have seen in the preceding section – they cannot be separated from each other properly, and they do not follow each other in a strict linear order.

3.2.1. Experimental design

Keysar et al. (2000) intended to test whether metaphorical expressions are comprehended by relying on conceptual mappings as Lakoff and Johnson's Conceptual Metaphor Theory (CMT) states. They formulated their conjecture to be tested as follows:

> "We will argue that conceptual mappings are not routinely used when people comprehend conventional expressions. If this is the case, then there would be no role for purported conceptual-level mappings when people comprehend conventional expressions. In contrast, language users might make use of a conceptual mapping when circumstances are appropriate, either by creating a conceptual mapping or by using a preexisting one. […] we explore the roles of novelty and explicitness as conditions that might foster the use of conceptual mappings. Specifically, we expect that people will be more likely to use conceptual mappings for novel, nonconventional than for conventional expressions. Second, explicit mention of a mapping […] might foster use of that mapping if appropriate expressions appear in the text." (Keysar et al. 2000: 579f.)

These hypotheses were tested by presenting people with "scenarios", that is, short texts on a computer screen. The final sentence of every scenario involved a nonconventional expression that was supposed to require a metaphorical mapping, i.e., the use of a conceptual metaphor according to Lakoff and Johnson's theory (target expression). In Experiment 1, there were 4 types of scenarios:

1. *implicit-mapping scenario*: contains conventionalised expressions that can be supposed to belong to the same conceptual metaphor as the target expression;[26]
2. *no-mapping scenario*: conventional instantiations of the supposed mapping are replaced by expressions not related to the given mapping;[27]
3. *explicit-mapping scenario*: in addition to the implicit-mapping scenario, the supposed mapping has been made explicit by being mentioned at the beginning of the text;[28]
4. *literal-meaning scenario*: renders the target expression as literal.[29]

In addition, the experimenters supposed that

> "[i]f a scenario instantiates […] mapping at the conceptual level, then it should facilitate the comprehension of a nonconventional expression that might require the instantiation of the mapping." (Keysar et al. 2000: 580)

From this they concluded that from Lakoff and Johnson's theory it would follow that, first, the target sentences were readily accessible and easier to understand in the case of the implicit-mapping scenario than in the case of the no-mapping scenario; second, explicit mention of the mapping should further facilitate the creation of the given metaphorical mapping. To find out whether this is the case, reading times of the final sentences were measured and compared.

Literal-meaning scenarios had a control function. They were intended to test whether the experimental procedure was capable of detecting relevant differences in comprehension times. Since there is experimental evidence that referential metaphors require more time to be understood than literal referring expressions, literal-meaning scenarios should have the significantly shortest reading times.

---

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

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

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

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

In Experiment 2, the explicit-mapping scenario was replaced by a novel-mapping scenario where the text contained novel metaphorical expressions instead of conventional ones.[30] Here, according to the experimenters, novel expressions should activate the creation of the conceptual mapping at issue, and via this, the comprehension of the target sentence should be faster if Conceptual Metaphor Theory were right.

3.2.2. The experimental procedure

In Experiment 1, 44 undergraduates, all native speakers of American English, took part. 16 item sets were generated, each set for a different conceptual mapping. Besides these sets, the test material included 10 filler scenarios whose final sentence was not metaphorical. Items and fillers were presented in a random order on the computer screen in every case. Item sets and conditions were counterbalanced in each list. In order to check whether participants paid enough attention to the task, they received a comprehension quiz after 8 scenarios. Results of participants who made more than one error were discarded.

The participants were asked to press a button as soon as they comprehended a line. The final sentence appeared not in isolation but simply as the last sentence of the text. The computer registered when the button was pressed after a participant had read a line.

In Experiment 2, 48 undergraduates participated for pay, under the same conditions as with Experiment 1. The same items and fillers were used; the only difference was that explicit-mapping scenarios were changed to novel-mapping scenarios.

The results were evaluated with the help of one-way ANOVA with repeated measures.

3.2.3. Perceptual data

Perceptual data consisted of values gained by measuring the times between pressing the button before and after having read the final sentence of the texts presented. These values were then interpreted as comprehension times of the given target sentence in the context of different scenarios.

3.2.4. Theoretical model of the phenomena investigated

The scenarios were supposed to contain different kinds of metaphorical (or, by contrast, non-metaphorical) expressions. The interpretation of perceptual data involved highly abstract and theory-dependent concepts as well, since the experiment was intended to test one of the central hypotheses of Lakoff and Johnson's theory, namely, the thesis of conceptual metaphors. It is important to remark that this can happen only *indirectly*, through a series of non-deductive inferences, since conceptual metaphors, conceptual mappings, etc. do not have observable properties, nor can a direct link be established between comprehension times and processing mechanisms.

The metaphorical expressions were chosen on the basis of the conceptual system of this theory:

---

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

– Target expressions were created as novel instantiations of conceptual metaphors listed in Lakoff & Johnson (1980).
– In implicit-mapping and explicit-mapping scenarios, metaphorical expressions appeared in the text that can be considered as conventional instantiations of the alleged conceptual metaphor in the target sentence. In explicit-mapping scenarios, the mapping was mentioned overtly.
– Novel-mapping scenarios made use of non-conventional instantiations of the mappings.
– No-mapping scenarios did not contain metaphorical expressions belonging to the mapping supposed to be present in the final sentence.
– Literal scenarios, as opposed to all other scenarios, furthered the literal interpretation of the target expression.

The experimental setting presupposes a complex network of phenomena which are related to perceptual data, high-level theoretical constructs and hypotheses (see Figure 1).

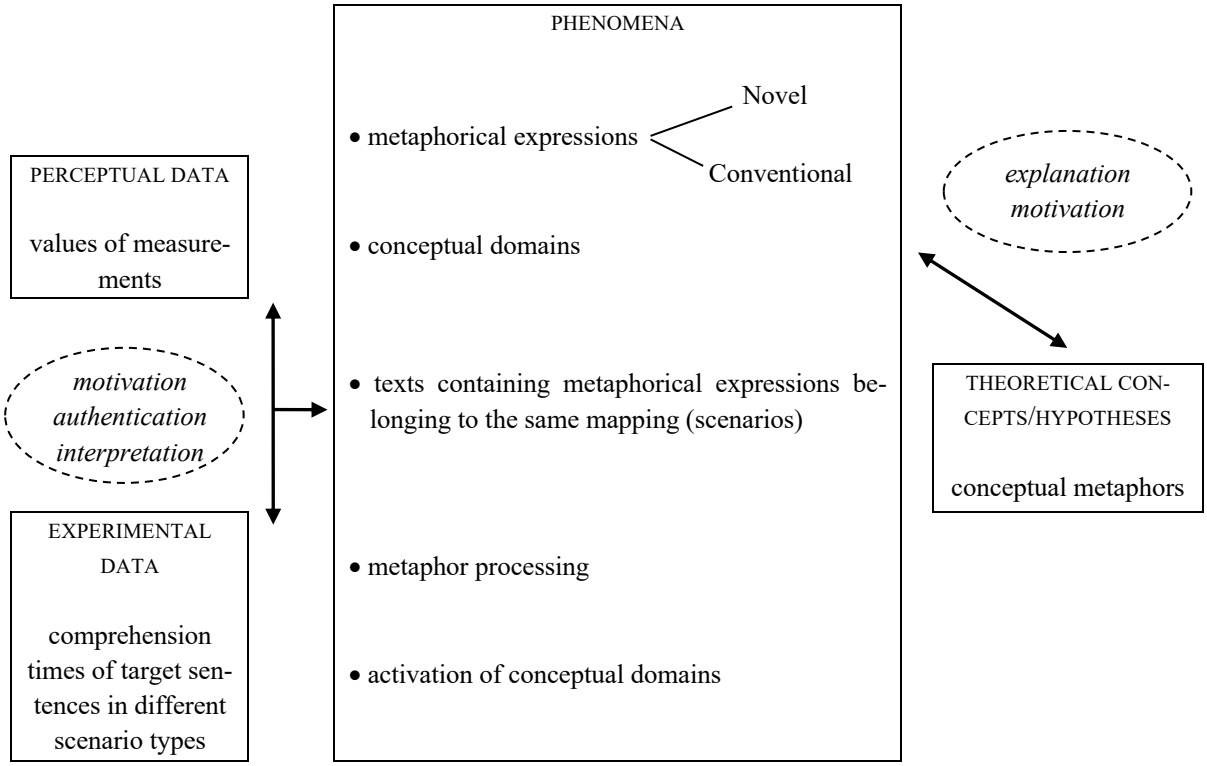

Figure 1. Phenomena and their relationship to data and hypotheses

In addition, it was assumed (and in several cases experimentally checked) that besides the mapping types, all other factors that could influence the comprehension time of the target sentences and lead to differences in the results stemming from the different scenarios could be ruled out. In this way, the authors arrived at the following set of *experimental data*: average comprehension time of sentences containing novel metaphors in implicit-mapping/explicit-mapping/ novel-mapping/no-mapping/literal scenarios.

These experimental data were then linked with further hypotheses of the Conceptual Metaphor Theory. It was assumed that if the thesis of metaphorical mapping in the sense of Lakoff

& Johnson (1980) and Lakoff (1993) holds true, then the comprehension times of novel metaphors in explicit-mapping, implicit-mapping and novel-mapping scenarios are significantly shorter than the comprehension times in no-mapping scenarios. Keysar et al. presumed that significant differences between reading times of the target sentence in no-mapping and mapping-scenarios should be due to the activation of a conceptual metaphor during the comprehension of metaphorical expressions in the text. Since no-mapping scenarios do not contain metaphorical expressions belonging to the given mapping, the processing of the novel metaphor in the final sentence cannot be facilitated this way; in contrast, instances of the conceptual metaphor in the preceding text should ease the processing of the metaphor in the final sentence.

As opposed to this train of thought, the authors raised a *rival hypothesis* as well:

> "Our alternative claim is that we usually *do* 'just talk' about arguments using terms that are also used to talk about war. Put more simply, the words that we use to talk about war and to talk about arguments are polysemous, but systematically related. Just as a word such as depress can be used to talk about either physical depression or emotional depression, words such as win or lose can be used to talk about arguments, wars, gambling, and romances, with no necessary implication that any one of these domains provides the conceptual underpinning for any or all of the others. The bottom line is that conventional expressions can be understood directly, without recourse to underlying conceptual mappings." (Keysar et al. 2000: 578; emphasis as in the original)

They argued that if conventional metaphorical expressions were comprehended not with the help of conceptual metaphors but as categorizations in the sense of the property attribution theory (cf. Glucksberg 2001, 2003; Glucksberg & McGlone 1999; Glucksberg et al. 1992), then there should be no significant differences between the comprehension times in different scenarios – and vice versa; if there were no differences in the reading times in implicit mapping and no-mapping scenarios, this had to be interpreted as experimental data in favour of the property attribution theory. If this were the case, then it would have profound consequences for the interpretation of the outcome of Experiment 2 as well. Namely, in this case, significant differences between reading times of the target sentences in novel-mapping vs. no-mapping scenarios could not be explained by the principle of conceptual metaphors either. Therefore, Keysar et al. raise the following alternative:

> "[…] novel expressions that reflect conceptual mappings between domains do lead readers to either *retrieve or create analogies between those domains.*" (Keysar et al. 2000: 588; emphasis added)

This means that they regard the activation of the source domain of metaphors as part of the processing of novel metaphors; however, they do not consider the mapping between the two conceptual domains involved to be an activation of a stable conceptual metaphor but rather the result of an analogical inference process:

> "Conceptual mappings, then, are not routinely used, but instead may be generated and used from perceived or inferred similarities between domains." (Keysar et al. 2000: 591)

For more on this, see Section 3.2.7.

3.2.5. Theoretical model of the experimental apparatus

The equipment used in the experiments does not seem to be of particular interest from an epistemological point of view. Nevertheless, if we interpret the term 'experimental apparatus' in a somewhat wider sense insofar as other components of the experimental setting such as the choice of the participants and the production of the test materials (scenarios) are regarded as belonging to its scope, then it is easy to see that there are several points needing careful theoretical considerations.

First, several possible sources of noise have been precluded. With the help of further and independent experiments, it was checked

– whether phrases figuring frequently in metaphorical expressions as parts of the target domain (such as *argument*) activate conceptual metaphors even in no-mapping scenarios where they occur without a source domain (for example, *journey*) in non-metaphoric expressions (cf. Keysar et al. 2000: 583);
– whether novel-mapping scenarios contain significantly less conventional metaphorical expressions than implicit-mapping scenarios do (cf. Keysar et al. 2000: 585f.);
– whether in novel-mapping scenarios, the ease of the comprehension is not due to the activation of conceptual mappings but to lexical priming or the text's discourse structure (cf. Keysar et al. 2000: 588ff.).

Second, it was ruled out that participants are linguists or students of linguistics, or that they have any idea on the focus of the experiment, because this could distort the results. Nevertheless, the choice of the participants can be criticised because they should represent the whole of the English speaking population. Although it was checked whether they are native speakers of American English, it can be questioned whether a group of undergraduates is representative of the totality of the language community – or it should have been shown that the population is homogeneous in respect to the use of metaphors.

Third, great attention was paid to the formulation of the text of the scenarios. The conventional metaphorical expressions were chosen from Lakoff & Johnson (1980) with minor editing in order to secure textual flow. The metaphorical expressions were selected in such a way that they belong to the same conceptual metaphor according to the Conceptual Metaphor Theory (CMT) in every scenario.

For further problems, see Section 3.2.7.

3.2.6. Authentication of the perceptual data

Despite the careful considerations mentioned in the previous section, we have to say that the authentication of the experimental results was not satisfactory.

First, it can be questioned whether the perceptual data were stable and reliable. The doubt emerges from the comparison of the results of repetitions of the experiments and the original ones:

a) Experiment 1 and Experiment 2 were almost identical; the only difference was that explicit-mapping scenarios were changed to novel-mapping scenarios. Despite this, there is a huge, and clearly significant, difference between the mean reading times of implicit-mapping scenarios in the two experiments, while with no-mapping scenarios, the difference is apprecia-

ble but may be not significant, and in the literal-mapping condition, the values are almost identical.

b) Thibodeau & Durgin (2008) repeated Experiment 2. The results showed a similar pattern, which could be a strong argument for their reliability. However, all mean reading times were considerably greater than in the original experiment – the mean difference was about 550ms. These discrepancies throw doubt on the reliability of perceptual data gained.

c) Literal scenarios were intended to fulfil a control function. Thus, in Experiment 1, it was emphasised by the authors that the significant difference between the comprehension time of the final sentences in the literal scenarios and in all other scenarios indicates that the experiment is sufficiently sensitive because it is capable of reflecting the difference in processing times between literal expressions and novel metaphors. Keysar et al., however, did not comment on the finding that in Experiment 2, the average of comprehension times of (metaphorical) final sentences in novel-mapping scenarios is almost identical with the mean of comprehension times of (non-metaphorical) final sentences of literal scenarios. This inconsistency needs resolution. See also f) below.

Second, the experimental setting also raises some problems:

d) It should be ruled out that there is any interference between the reading times of whole scenarios of different types and the comprehension time of the target sentences. That is, it should be checked whether there is any considerable difference among the reading times of scenario types, and if this is the case, then the question is whether this influences the reading time of the target sentences. For example, according to Gentner & Bowdle (2008: 117), novel metaphors require more time to be comprehended than conventional ones or literal expressions. Consequently, one has to examine whether the comprehension time of novel-mapping scenarios is longer than that of implicit-mapping scenarios, and the relatively higher comprehension time of novel-mapping scenarios slows down the reading of the target sentence, and the relatively lower reading time of implicit-mapping scenarios accelerates it to some extent.

e) Since the judgement of metaphoricity is subjective and strongly theory-dependent, the choice and categorisation of the metaphorical expressions in the materials may be a controversial issue. In fact, in spite of the author's reference to Lakoff & Johnson (1980), the wording of the scenarios was questioned by many researchers from different points of view.[31] The tenability of these criticisms cannot be judged properly, since Keysar et al's article contains only a part of the materials applied. Nevertheless, examination of the excerpts of the texts presented

[31] "[…] in several cases, the novel and conventional phrasings in the Keysar et al. (2000) stimuli result in different interpretations. We found two kinds of unparallel scenarios. First, there were cases in which the lead-up scenario in the novel version introduced concepts relevant to interpreting the target sentence that were not present in the conventional version. Second, there were cases for which the target sentence may have appeared as a non sequitur following the conventional but not novel version of the lead-up scenario." (Thibodeau & Durgin 2008: 533)

"The experiment makes several assumptions about usage, including the following: 1. that *fertile*, used in the second sentence of the second text, is a novel metaphor; 2. that *weaning*, in the last sentence of each text, is a novel metaphor; 3. that *latest child*, in the last sentence, is potentially ambiguous between the meanings 'a child' and 'a set of experimental findings.' Corpus analyses raised problems with each of the three assumptions […]." (Deignan 2008: 286; see also Gibbs & Lonergan 2007: 78f.)

Although Deignan's first two objections seem to be mistaken since they take into consideration only isolated words instead of phrases, the third one can be considered correct.

in Keysar et al. (2000: 582) and in Thibodeau & Durgin (2008: 525) reinforce Thibodeau and Durgin's concern that the results of the experiment might be unreliable:

– In some cases, metaphorical expressions in the text of a scenario and in the final sentence cannot be regarded as instantiations of the same conceptual metaphor in the sense of Lakoff & Johnson (1980). For example, in the scenario 'love is a patient', the target sentence *You're infected with this disease* should rather belong to the conceptual metaphor BAD FEELINGS ARE ILLNESSES or JEALOUSY IS AN ILLNESS. Moreover, the existence of the conceptual metaphor LOVE IS A PATIENT can be questioned; the expressions *this relationship is on its last legs*, *a strong marriage*, *this relationship is about to flatline* could be interpreted as belonging to the conceptual metaphor RELATIONSHIPS ARE PEOPLE.
– Novel metaphors seem to be more closely related semantically to the target expression than is the case with conventional ones; therefore, as Thibodeau and Durgin also remarked, the text of novel-mapping scenarios is (at least in some cases) more fluent and conceptually more homogeneous.
– Novel-mapping scenarios start – similarly to explicit-mapping ones – with an explicit mentioning of the alleged conceptual metaphor. This may have eased the comprehension of the target expression in contrast to no-mapping or implicit-mapping scenarios due to a semantic priming effect.
– The fluency and conceptual homogeneity of novel-mapping scenarios in comparison to implicit-mapping and no-mapping scenarios may also give rise to semantic priming. Experiment 3 by Keysar et al. tried to rule out this possible source of noise, but it was related only rather indirectly to the problem at issue. Namely, a target word in the last sentence of the novel-mapping contexts was selected on the basis of the votes of 8 participants; then another group of participants had to decide whether these words were English words after having read the text of different types of scenarios. Since there was no significant difference between the reaction times given in the scenarios in this lexical decision task, Keysar et al. concluded that there are no priming effects. It is, however, questionable whether differences among the scenarios could influence the comprehension times of well-known English words. Therefore, without any control of the sensitivity of this method, the result of this experiment cannot be regarded as plausible.
– According to, for example, Bowdle & Gentner (2005: 204ff.) who refer to earlier results as well as their own experiments, comprehension times of metaphors are influenced by familiarity and aptness besides conventionality. These factors should also be accounted for when planning and evaluating the experiments.

f) Thibodeau & Durgin (2008) conducted an experiment similar to Experiment 2 by Keysar et al. The experimental setting was modified at two points. First, the text of the scenarios was rewritten in order to secure textual flow and conceptual fit. That is, the metaphorical expressions were selected in such a way that they can be related to the same conceptual metaphor in the sense of Lakoff & Johnson (1980) in each scenario, but there is no conceptual overlap between the conceptual domains of metaphorical mappings in different scenarios. Second, the filler scenarios were chosen on the basis of different considerations than was the case with the original experiment. Namely, Keysar et al's main motivation was to make sure that "partici-

pants would not anticipate or notice a particular pattern" (Keysar et al. 2000: 583), and in this spirit, their fillers contained neither metaphorical final sentences nor metaphors belonging to the same conceptual domains. With the new version by Thibodeau and Durgin, however, 2 in every 3 filler scenarios did contain metaphorical expressions; moreover, the fillers were intended to "avoid reading strategies that would cause people to skim over metaphors" (Thibodeau & Durgin 2008: 523). Thus, 4 of 10 questions following the fillers asked about metaphors. The outcome of the experiment contradicted the findings of Keysar et al's experiment because there was no significant difference between the comprehension times in the novel-mapping, the implicit-mapping and literal scenarios, while all of them were significantly faster than no-mapping scenarios.

In a further experiment, Thibodeau & Durgin (2008: 529ff.) found that if the novel metaphor in the final sentence belonged to the same metaphor family (metaphorical mapping) as the conventional metaphors in the preceding text (that is, if they were "matched metaphors"), then the final sentence read significantly faster than final sentences involving a novel metaphor from another metaphor family as the preceding text, or when the text of the scenario did not contain metaphors. In this way, they created *new experimental data*: average comprehension time of sentences containing novel metaphors in scenarios using conventional metaphors from the metaphor family of the target sentence vs. average comprehension time of sentences containing novel metaphors in scenarios using conventional metaphors from another metaphor family. Thus, the experiments resulted in a shift in the judgement concerning what data should be regarded as relevant: instead of novelty/conventionality, the key factor seemed to be matchedness/unmatchedness.

Nevertheless, this still does not constitute decisive evidence against Keysar et al's results. First, because of the modification of the fillers and the control questions, the participants might have discovered relatively easily that the experiment focused on the use of metaphorical expressions. Second, it may be the case that the shorter reading times in metaphorical scenarios in comparison to no-mapping scenarios were due to semantic priming.[32] Third, the similarity in reading times of literal targets and metaphorical ones should be accounted for in this case, too. Fourth, Gentner & Boronat's (1992) experiments were in accord with Keysar et al's findings and not with Thibodeau and Durgin's (see also Gentner & Bowdle 2008, Gentner et al. 2001). This is more than a little surprising because Thibodeau and Durgin referred to Gentner's writings many times and argued for the structure mapping theory as a possible explanation of their results. Actually, Gentner & Boronat's (1992) experiments showed a significant difference between comprehension times of novel metaphors after texts containing *novel* metaphors belonging to the *same* metaphorical mapping ("consistent scenarios") on the one hand, and comprehension times of novel metaphors following texts containing *novel* metaphors belonging to *another* mapping ("inconsistent scenarios") on the other. When, however, they used *conventional* metaphors in the text, then the difference in reading times between consistent and inconsistent scenarios disappeared.

[32] The same problems should be eliminated from the third experiment carried out by Thibodeau and Durgin, where comprehension times of final sentences after texts containing metaphorical expressions belonging to the same conceptual metaphor and texts containing metaphorical expressions stemming from different metaphor families were compared.

Although neither the experimental materials used, nor the perceptual data can be found in Gentner's and her colleagues' writings, it seems that these experiments were based on the most thoroughly elaborated experimental design – although they are the oldest among the three series of experiments. First, they used *the most differentiated data-set*: average comprehension times of sentences containing novel metaphors in novel consistent-mapping scenarios, average comprehension times of sentences containing novel metaphors in novel inconsistent-mapping scenarios, average comprehension times of sentences containing novel metaphors in conventional consistent-mapping scenarios, average comprehension times of sentences containing novel metaphors in conventional inconsistent-mapping scenarios, and average comprehension times of sentences containing novel metaphors in literal (non-metaphorical) scenarios. The latter differ from the literal-meaning scenarios used by Keysar et al. and Thibodeau and Durgin insofar as their text contains terms from the source domain (in their literal meaning, without the target domain) of the corresponding metaphorical scenarios, but in the final sentence, the novel metaphor is used in its metaphorical meaning. Thus, literal-meaning scenarios are *controls which seem to be capable of ruling out the effect of semantic priming*.[33] Nevertheless, Gentner and her colleagues' papers present only short excerpts of the stimulus material and no concrete measurement results. Consequently, their contributions cannot be judged properly either.

g) A further important factor is that we are not in possession of the perceptual data, that is, the measurement results. Without the whole data set, it is not possible to check the adequacy of the statistical methods applied by the authors.

At this point, it would be reasonable to scrutinise the texts of the scenarios, and apply a control method frequently used in statistics: namely, the perceptual data should also be evaluated separately for every scenario in order to check whether there are significant differences between the results which might be due to the wording of the particular scenarios.[34] Another important step towards the validation of experimental results would be the repetition of the experiments after the revision of the texts of the scenarios by different researchers and with the participation of subjects representing a wider segment of the population. In this way, further possible shortcomings or malfunctioning of the measurement method could be ruled out. Furthermore, the influence of the semantic priming should be ruled out, and the aptness and familiarity of metaphorical expressions should be taken into account as well. Moreover, the data set should be further differentiated. That is, it should also be investigated whether there is a difference between scenarios making use of novel metaphors related to existing metaphor families (in the text and in the final sentence, respectively) on the one hand, and scenarios containing novel

[33] "In this condition, participants encountered the *terms* from the metaphoric base domain in the passage but not the metaphor itself (until the final test sentence). If the facilitation for the consistent condition over the inconsistent condition were due merely to associative priming, the final sentence should not differ between the consistent condition and the literal control condition." (Gentner & Bowdle 2008: 124; emphasis as in the original)

[34] Such a difference has no significance per se; nevertheless, it can motivate the search for the possible causes of the deviation, and via this, the improvement of the experimental setting and the performing of further experiments.

metaphors connecting two conceptual domains where there are no conventional metaphors instantiating this mapping. Without such and perhaps further revisions of the original experimental setting, the experimental data cannot be regarded as reliable.

3.2.7. Interpretation of the perceptual data

As we have seen in Section 3.2.4 (cf. especially, Figure 1), the theoretical model of the experimental setting involves several low- and high-level theoretical constructs. Consequently, the interpretation of the perceptual data – i.e. establishing the link between them and the phenomena – is clearly theory-laden, in that it involves many hypotheses that cannot be checked empirically in a direct way. Thus, for example, the metaphoricity of expressions, the classification of expressions into metaphor families or metaphorical mappings involve subjective, arbitrary elements stemming from the intuitive judgements of the experimenters, which cannot be completely eliminated.

There is a highly complex, long chain (or rather, system) of hypotheses and inferences establishing a connection between the perceptual data gained and the phenomena. These inferences rely in most cases on premises that are not true with certainty but plausible (only presupposed or partially supported by the perceptual data or other hypotheses). The inferences often also make use of *latent background assumptions* which are only plausible, or even remain unidentified. Thus, they are not capable of securing the truth of their conclusion (although they may make them – under appropriate conditions – plausible). For example, from the observation that in the case of participant *X*, the value 1632 ms was obtained in novel-mapping scenario No. 4, it does not follow conclusively that *X* has applied a mapping from the conceptual domain *journey* to the conceptual domain *argument*. The same is true of the reverse direction: the hypothesis that *X* has applied a mapping from the conceptual domain *journey* to the conceptual domain *argument* is far from being sufficient to explain why the value 1632 ms was gained in novel-mapping scenario No. 4 in the case of participant *X*. Similarly, from the perceptual data one cannot conclude conclusively that the participants applied the same procedure by processing the metaphorical expressions presented. Or, it has not been proved but only presumed that the sentences of the scenarios contain metaphors belonging to the same metaphor family – and the list could be continued.

The statistical tools applied also contribute to the increased abstractness of experimental data in comparison to raw perceptual data. They reduce a series of individual data points to mean values, while isolated extreme data values are omitted. Thus, their application inevitably leads to information loss – although, of course, they lead to new information as well.

To sum up, perceptual data underdetermine not only (theoretical) explanations but the constitution of experimental data as well.

3.2.8. Presentation of the experimental results

The presentation of the experimental results undoubtedly conforms to the generally accepted methodological rules in psycholinguistics. However, it is also in the spirit of these norms that relevant information was eliminated, such as the complete perceptual data set, or the text of the stimulus materials. Without these, the experimental results cannot be judged properly, as we have seen in the previous sections. In contrast, in physics, detailed accounts of the experimental design and raw data sets are often made public.

Since Keysar et al.'s aim was to test one of the central hypotheses of Lakoff and Johnson's theory – the thesis of conceptual metaphors – the experimental results were linked to further high-level, strongly theory-specific concepts and hypotheses. They explained the experimental data gained in Experiments 1 and 2 in such a way that they indicate a fundamental difference in the processing mechanisms of novel and conventional metaphors, respectively. They concluded that while the former rely on mappings between two conceptual domains, the latter are accomplished directly, not as mappings but as categorisations. Explicit mentioning of metaphorical mapping was found to be irrelevant in relation to metaphor processing. On this basis, they rejected the hypothesis of metaphorical mapping on the lines of the Conceptual Metaphor Theory because the latter assumes that novel and conventional metaphors are comprehended in the same way (cf. Keysar et al. 2000: 591f.). As a rival proposal in accord with the experimental data, they offered a mixed explanation: on the one hand, conventional metaphors are processed as categorical statements in the sense of Glucksberg's property attribution theory; on the other hand, novel metaphors result from a cyclic process consisting of the structural mapping of two conceptual domains and a series of analogical inferences, as Gentner's structural mapping model states.

Given that Gentner's 'career of metaphor' hypothesis models the processing of conventional metaphors in a similar way as the structure mapping theory, and Gentner and her colleagues found similar results as Keysar et al., they argue that Gentner's theory is appropriate for accommodating both Keysars' and her and her colleagues' experimental results, too.

Interestingly, Thibodeau and Durgin also interpret their results by referring to Gentner's theory, although they are incompatible with Keysars' and Gentners' findings. The reason for this *inconsistency* might be that according to Gentner's model, the source domain may be activated in the case of conceptual metaphors as well.[35] At this point, the theoretical model should have been improved, and with this, the experimental design should have been developed.

---

[35] Cf.:

"Conventional base terms are polysemous, with the literal and metaphoric meanings semantically linked because of their similarity. Conventional metaphors may therefore be interpreted either as comparisons, by matching the target concept with the literal base concept, or as categorizations, by seeing the target concept as a member of the superordinate metaphoric category named by the base term. This raises an interesting question: How, exactly, are metaphoric categories applied to target concepts during comprehension? We suggest that categorization, be it figurative or literal, relies on the same basic mechanisms as comparison – namely, structural alignment and inference projection. […] there is no reason to believe that the processes involved in categorization are different in kind from those involved in comparison. Both processes involve some kind of alignment of representations to establish commonalities and guide the possible inheritance of further properties. The primary distinction between the two may lie in the kind and degree of inference projection. Although comparison processing entails the projection of inferences, the inference process is highly selective; only those properties connected to the aligned system are likely to be considered for projection. In contrast, categorization involves complete inheritance: Every property true of the base should be projected to the target. Thus, the career of metaphor claim that conventional metaphors may be interpreted as comparisons or as categorizations can be rephrased by saying that such metaphors may be processed as *horizontal alignments* (mappings between representations at roughly the same level of abstraction) or as *vertical alignments* (mappings between representations at different levels of abstraction). There is, however, reason to expect that these two modes of alignment will not be favored equally for conventional metaphors. Let us assume that both meanings of a conventional base term are activated simultaneously during comprehension and that attempts to map each representation to the target concept are made in parallel […]. This

Nevertheless, one should not forget that the perceptual data do not preclude models that assign the source domain an active role in the processing of novel metaphors. Therefore, alternative explanations are possible which may considerably differ from Gentners' view.

A further relevant point is that all three groups of researchers mentioned also make use of non-deductive inferences such as analogy, induction, reduction etc. by establishing a link between the not certainly true but only plausible experimental data and the hypotheses of the preferred or the rival theories. Consequently, they neither verify nor falsify the theories at issue but they make them more or less plausible in comparison with the rival proposals with the help of the experimental results.

The case study presented in this section revealed, among other things, the following *similarities* between experiments in physics and in cognitive linguistics:

– Observation is requisite but its role is by no means as decisive as the standard view of the analytic philosophy of science suggested. Perceptual data have to be authenticated and interpreted.
– The interpretation of data leads inevitably to the theory-ladenness of experimental results.
– Data are evaluated by statistical means in order to eliminate the influence of random errors and to examine whether the data support the hypotheses raised because it is reasonable to ascribe the differences between certain groups of data to factors identified by the hypothesis, or this is not the case and these differences are due to chance.
– The statistical tools not only provide us with new information but reduce the set of information at our disposal in the sense that they substitute individual data points with the mean and some other characteristics of data sets.
– Several potential systematic errors have been excluded by further experiments. Despite this, it is possible that there are others which distort the results; moreover, the control experiments may contain systematic errors, too. Therefore, the experimental design always remains partial.
– Nothing prevents different researchers interpreting the same set of perceptual data differently.
– Experimental data are not true with certainty but only plausible on the basis of the given experiment. Thus, experiments are open processes that can be continued and revised in possession of new data or new considerations.

---

would be akin to parallel process models of idiom comprehension […]. Which of these mappings wins will depend on a number of factors, including the context of the metaphor and the relative salience of each meaning of the base term […]. All else being equal, however, aligning a target with a metaphoric category should be computationally less costly than aligning a target with the corresponding literal base concept. For one thing, metaphoric categories will contain fewer predicates than the literal concepts they were derived from, and a higher proportion of these predicates can be mapped to relevant target concepts. Moreover, assuming that the predicates of metaphoric categories will tend to be more domain general than those of literal base concepts, metaphoric categories should require less rerepresentation when matched with domainspecific predicates in a target concept. In general, then, conventional metaphors will tend to be interpreted as categorizations rather than as comparisons because the former mode of alignment will be completed more rapidly than the latter." (Bowdle & Gentner 2005: 199; emphasis as in the original)

- Results of similar experiments may contradict each other.
- The presentation of the experimental results is fragmentary in the sense that it does not contain details of the experimental process that were judged to be irrelevant. Thus, the "edited" version of the experiment contains only traces of the real process. This may be problematic from two points of view. First, it allows only a partial reconstruction of the experimental procedure. Second, it is the experimenters themselves who decide upon the relevance/irrelevance of events, data or other pieces of information related to the given experiment, which poses the danger of the experimenter's circle.
- The experiments conducted by Gentner & Boroditsky, Keysar et al. and Thibodeau & Durgin, as well as the papers cited in which they analyse the results can be deemed to be stages of a cyclic and prismatic process of successive re-evaluation. Each paper took new points of view into consideration, and tried to revise the experimental setting in order to achieve more reliable results. This process is clearly not linear; neither can it be described as a continuous evolution of the results and theories. Rather, it indicates that previous and already rejected hypotheses or explanations may revive and be improved.

3.2.9. Analogies and differences between experiments in physics and cognitive linguistics

To sum up our considerations presented in Section 3.2, we can conclude that metascientific reflection on the nature and limits of experiments as data sources in linguistics has to be based on the *continuous comprehension and adjustment* of the reflection on the research activities of linguists while working with experiments on the one hand, and insights gained by philosophers of science studying experiments in other disciplines. Results of metascientific reflection on experiments carried out by philosophers of physics, biology, psychology, social sciences, etc. have to be analysed to determine whether there is *analogy* between them and the situation in linguistics.

There are, of course, also certain *differences* between experiments in physics and cognitive linguistics. The most important are perhaps the following:

- Physics divides into experimental and theoretical branches, while in cognitive linguistics, experiments are always presented as arguments for or against a theory or theories; that is, they never appear independently but as parts of scientific theorising, used to argue in favour of the theory at issue or against a rival theory.
- In physics, raw experimental data (perceptual data) are often published; in cognitive linguistics, this is rather exceptional.
- In physics, experiments are almost always repeated; in cognitive linguistics, this occurs only if the results are questioned. Moreover, during the replication the experimental setting applied is often not the same, but only similar.
- In cognitive linguistics, there is usually a strong overlap between theories that are confronted with experimental results and theories applied by the interpretation of the perceptual data. Therefore, the theory-ladenness in linguistics in most cases also means theory-dependence. In physics, however, experimental data usually contain lower-level theoretical concepts.

- In cognitive linguistics, the authentication of perceptual data consists of a checking of the experimental setting and only to a lesser degree that of the experimental apparatus. The importance and role of the latter is considerably greater in physics.
- In physics, the style and content of experimental reports is strongly regulated, impersonal and is strongly focused on the description of the experimental procedure and the experimental data. In cognitive linguistics, in contrast, the role of argumentation in papers presenting experimental results is considerably less regulated, covers a wider spectrum of topics and is clearly more prominent.

These differences indicate that experiments in cognitive linguistics are not identical with physical experiments – there is only a *strong analogy* between them and also between the epistemological problems they raise. This finding underlines the importance of the elaboration of detailed case studies in metalinguistic research, because *research practice* as well as the *self-reflection of cognitive linguists* has to be taken into consideration. Nonetheless, in cognitive linguistics experiments cannot be separated from theory formation, one cannot narrow down the metascientific reflection on experiments to the experimental process itself. Instead, experiments have to be studied as parts of the process of linguistic theorising. From this, a further task arises: we also have to study and model linguistic theorising. This means that methodological guidelines or principles have to be in accord with *a general account of linguistic theorising* which covers not only specific issues related to the treatment of experimental data but comprises the whole process of theory formation.

3.3. The structure of experiments in cognitive linguistics

A very important insight of the current literature on scientific experiments is that neither single experiments nor repetitions of the experimentation process are capable of yielding ultimate and unquestionable results. It is not only the previous considerations and the planning of the experiment which are fallible – the control of the experimental process and the evaluation of the results are to some extent unavoidably uncertain as well. Therefore, experiments are *open processes* in the sense that, in possession of new pieces of information, they may be continued, modified, or even discarded. As we have seen in Section 3.1, according to recent trends in the philosophy of science, the reliability of the outcome of an experiment depends on the reliability of its *components*, as well as the *fit* between them and pre-existing knowledge. To find this fit, one has, in most cases, to *turn back* to earlier stages of the experimentation process and modify some component. Every component can be *revised* and the revisions have to be repeated again and again until there is mutual support among the constituents. See Figure 2.

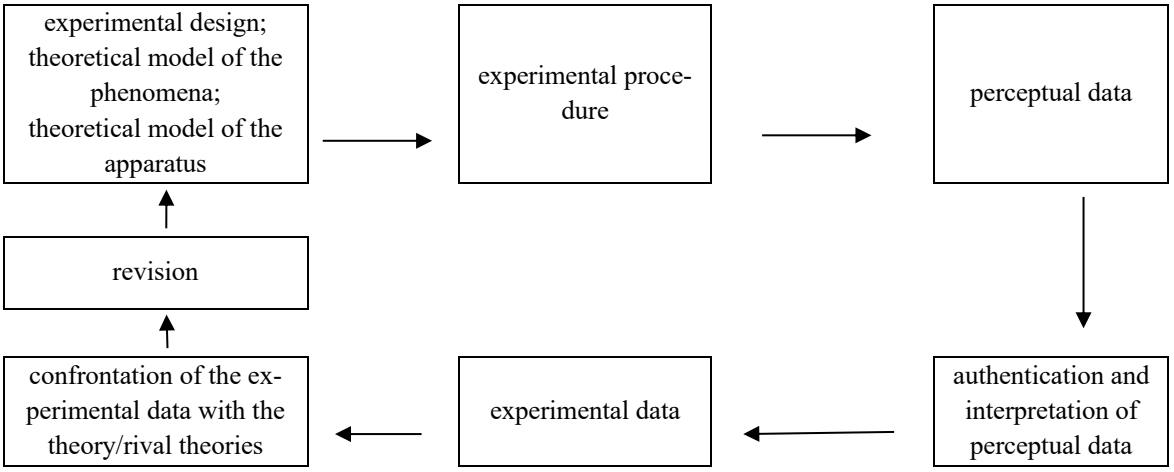

Figure 2. The basic structure of experiments

As we have already mentioned in Section 3.1, the idea of achieving a mutual support among the components of the experiment as an arbiter of the acceptability of the experimental results involves the *risk of circularity* and may lead to the *experimenter's circle*. This problem, as Kertész & Rákosi (2009) has shown, can only be solved *heuristically*. This means that if a process returns to the start in such a way that it leaves it unchanged by failing to re-evaluate the information content at one's disposal, then it is *ineffective* and does not bring one closer to the solution of the problems raised. As opposed to this, with *cyclic processes*, "one indeed returns to 'the same point' but does so at a different cognitive level" (Rescher 1976: 119), since a modified, prismatically re-evaluated, qualitatively new information state is created (see also points *(ix)–(xi)* in Section 4.1 in Kertész & Rákosi 2009, Section 10.4 in Kertész & Rákosi 2012, as well as Rescher 1987). Accordingly, cyclic argumentation is *effective.* From this it follows that if the evaluation of the results and components of the experimentation process systematically ignores potentially relevant data, does not examine alternative interpretations, does not check possible error sources, or leaves unclarified relevant factors which could decrease the reliability of the results, then we are faced with the *experimenter's circle*. If, however, the process is *prismatic* in the sense that one continuously changes the *perspective* from which the pieces of information constituting the context are evaluated (cf. Rescher 1987, 1977; Section 10.5 in Kertész & Rákosi 2012), and *alters our informational state* by extending it with new pieces of information or by revising it, then it is the *experimenter's cycle* – an effective and fruitful enterprise of gaining new information about the world.[36]

Nevertheless, this mutual support does not guarantee the certainty of the results; rather, it is a sign of their *plausibility*. It is reasonable to accept them on the basis of the available information but one should never forget that there may always be systematic errors, other alternative explanations etc. that have not been taken into account. To reduce the amount and impact of the latter and to increase the plausibility of the results of the experiment, one has to *actively seek for possible weak points* – that is, one has to *reflect* on every detail of the experimentation process from its planning to its evaluation with the help of the strategies proposed by Franklin and by elaborating further ones.

[36] For similar views, cf. Nickles (1989), Pickering (1989), and especially in linguistics, Pullum (2007).

In this vein, let us take a closer look at the process of searching for the best fit among the components of the experimental process *from the point of view of the experimenter*.

Hypotheses used in the experimental design, the theoretical model of the phenomena and of the apparatus make up the starting point of the experimental process. They are not true with certainty but they are supported to some extent by theoretical considerations, by earlier experiments, or are simply (reasonable) conjectures. They allow for a rough estimation of the outcome of the experiment. After the experimental procedure, in possession of the perceptual data, this preliminary guess may be strengthened. Nevertheless, it may happen that the perceptual data cannot be interpreted properly, or they seem to be in conflict with the predictions. In such cases, the reliability of the previously accepted hypotheses also has to be revised.

The interpretation and authentication of the perceptual data may also indicate shortcomings in the experimental procedure, in the experimental design, or in the theoretical model of the phenomena or of the apparatus. Therefore, all facets of the experiment conducted have to be re-examined, and, if it seems to be necessary, control experiments have to be carried out, or the experimental design has to be modified and the experiment repeated. Moreover, even the interpretation or the authentication of the perceptual data itself may be faulty and be in need of modification.

From this it follows that revealing the *connections* between the statements capturing different aspects of the experimental procedure and their *analysis,* as well as the *comprehensiveness* of the checks and cross-checks are of crucial importance.

This characterisation of the experimental process will motivate us to raise the hypothesis that *experiments are cyclic processes organised and conducted by an argumentation process* which tries to clarify the relationship among hypotheses of the experimental design, the theoretical model of phenomena, the theoretical model of the experimental apparatus, the theory under test and its rivals, as well as statements describing the events of the experimental procedure, or which capture the results of the interpretation and authentication of perceptual data etc. This motivates us to modify Figure 2 as shown by Figure 3.

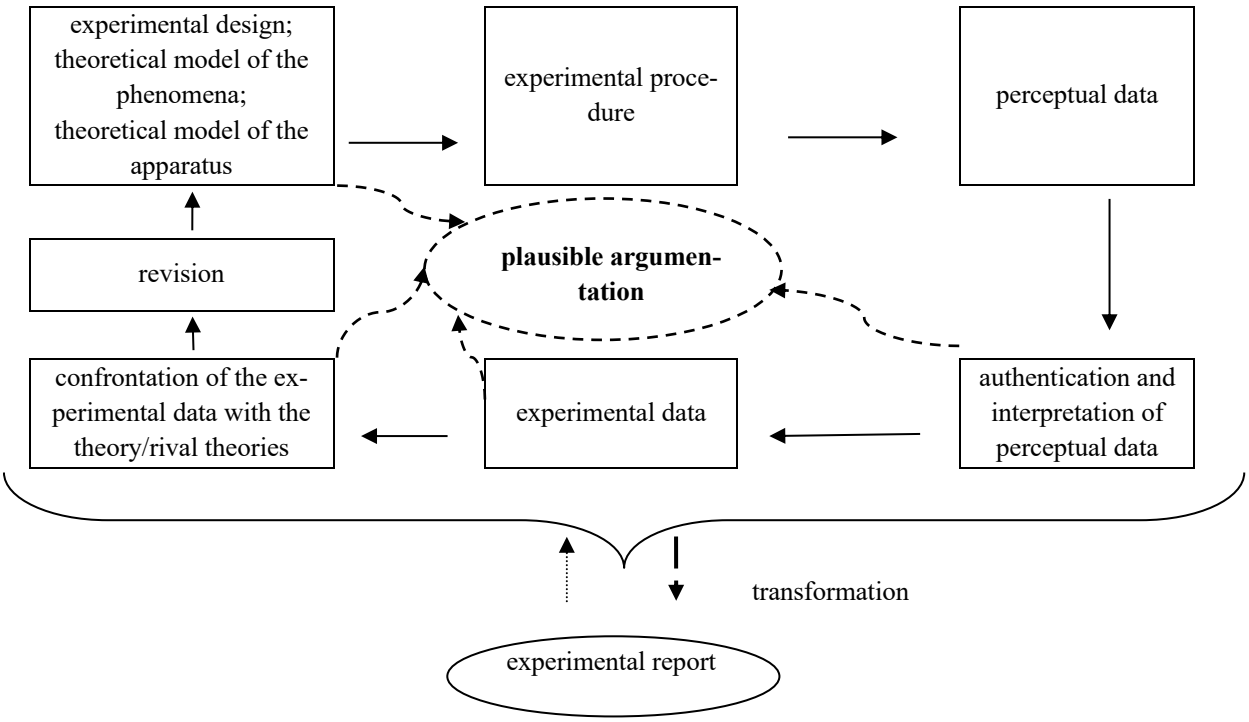

Figure 3. The structure of experiments – extended[37]

The argumentation process organising the conduct and control of the experiment is a central issue in judging the reliability of the experimental results. This argumentation process does not consist of deductive inferences because it takes into consideration the uncertainty of the statements as well. It is not public; rather, it is a private affair of the experimenters. Despite this, it is indirectly influenced by the public norms applicable to experiments. Thus, experimenters have first to convince themselves of the reliability and acceptability of the outcome of the experiment and, after this, they have to persuade the scientific community as well. This means that the researcher has to transform this *private argumentation process* into a *public experimental report*.

If we turn to the *reader's perspective*, we can establish that the evaluation of experimental results can only start from the experimental report, which is an edited, transformed version of the non-public argumentation process. While the latter is part of an ontologically complex process of scientific experiment, the former is purely argumentative. It is a *mixture*: it contains elements or traces of the original argumentation process as well as the argumentative tools needed to make this reduced set of information coherent, comprehensible and persuasive for the reader.

In the next section, we will take a closer look at the peculiarities and function of argumentation in experiments in cognitive linguistics and present tools which allow us to reconstruct it and evaluate its acceptability.

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

4. Argumentative aspects of experiments

Since scientific experiments are *historical products*, the first subsection will be devoted to a brief historical overview of the manifold relationship between experiments and rhetoric/argumentation. It will show that the first view in relation to (RPE), according to which experimental results should be free from any kind of argumentation, is a contingent historical product – one of the rhetorical/argumentative practices that were applied to secure the authority of experiments in science. In the second subsection, we will give an account of the metascientific tools with the help of which the current argumentative practice in experiments in cognitive science can be described and evaluated.

4.1. A brief history of the relationship between rhetoric/argumentation and scientific experiments

The argumentative/rhetorical tools[38] applied in experiments as well as the role they fulfil have undergone several changes during the history of science.[39] This is mainly due to the variety of ways in which science has been practised and reflected upon.

Experiments were first applied in the 17th century when it became clear that pure reasoning, speculation, passive observation, and reference to ancient authorities or to religious dogmas were no longer capable of providing relevant information about nature. Artificial situations were created, but the usability of experiments and the acceptability of the results were fiercely debated:

> "The new 'experimental philosophy' was greeted with scepticism on two different grounds. Its critics pointed out two difficulties with regard to experimentation. First, in contrast to the phenomena that could be observed with the unaided senses, the phenomena created by experiment were neither familiar nor accessible to everyone. Second, it was unclear why the manipulation of nature by means of instruments would reveal, rather than distort, its workings. Those difficulties were two aspects of the same issue, namely the authentication of experimental results; an issue which had to be resolved before experimentation could become a proper foundation for natural philosophy." (Arabatzis 2008: 160)

Simply imagining the processes that might take place under the given circumstances was felt to be unsatisfactory. Thus, the authority of experiments had to be secured by the conduct of the experiment, that is, by the replacement of thought experiments by real ones, as well as by diverse *rhetorical tools*. One method was that scientists listed the names of prominent people who had been present at the experiment. Thus, *the authority of the experimenter and the witnesses* played a decisive role in the appraisal of the experimental results. This was, of course, the application of the earlier rhetorical strategy of reference to authorities to the new method. Another strategy, elaborated by Boyle, was also a rhetorical tool insofar as he gave a vivid and detailed account of every phase of the experimental process in order to make the reader, as Cantor (1989: 163) coins it, a *virtual witness*. This strategy has a medieval counterpart as well:

---

38 In this section, the terms 'rhetorical' and 'argumentative' will be used in a pre-explicative sense.

39 Cf. Cantor (1989: 162ff.), Gooding (2000: 117ff.), Arabatzis (2008: 159ff.).

narratives in medieval chronicles, as Schiffrin (1981: 59) explains, applied the shift to the historical present as an argumentation technique functioning as an "internal evaluation device: it allows the narrator to present events as if they were occurring at that moment, so that the audience can hear for itself what happened, and can interpret for itself the significance of those events for the experience."[40] Nevertheless, the application of this tool was a very important shift in the role of 'witnessing', since the source of the rhetorical power of the experiment was no longer based on authority but on the (real or only conceived) *reproducibility* of the procedure.

A later, far-reaching move was the use of an *impersonal style* by relying on the dominance of mathematical tools, schematisation and formalisation. This was intended to create the impression that the authority of experimental results stems directly from nature, without (subjective) human intervention and interpretation. This style, however, has also led to *fragmentariness*: a state in which several details of the experimental process are dismissed from the experimental report.

Despite this, of course, it was often the case that the arguments presented (calculations, formal/mathematical methods, the experimental design, the interpretation of the results) or the devices applied were criticised. Therefore, the experimental reports were also *extended* to new elements such as the identification of possible systematic errors and the description of the measures taken for their prevention, statistical analyses enabling the elimination of effects of unavoidable sources of noise, etc. These arguments have grown in importance and have come to be regarded as decisive factors in judging the acceptability of experimental results. At the same time, the description of the experimental procedure became more theory-oriented.

Thus, the experimental report is considerably *richer* than the experimental procedure itself was, but, at the same time, it remains strongly schematic and informationally *reduced*. Specifically, the experimenter selects the relevant moves and events which are accounted for in the experimental report; she/he has the privilege of deciding what counts as an accidental, contingent mistake which may remain unmentioned and, on the other hand, what has to be regarded as a systematic error that has to be reported together with its correction. A highly instructive and often cited example of the *gap* between the "real" happenings in the laboratory and their accounting for in the experimental report is Millikan's celebrated, historic oil drop experiment:

> "Yet extant laboratory notebooks also sometimes indicate more interesting mismatches between laboratory practice and published reports. Holton, for example, has drawn attention to Robert Millikan's selection of acceptable results for his oil-drop experiment. During one series of experiments Millikan omitted well over half of his results, retaining data from only 58 drops out of a total of about 140. Against some runs he annotated comments such as '*Beauty*. Publish this surely, *beautiful*', whereas in other cases he dismissed the run with '*Error high* will not use', or some such remark. His reasons for accepting some runs and not others are complex: sometimes parts of this apparatus did not appear to function properly, on other occasions the result was not sufficiently close to the emergent value for *e*, the electronic charge. […] contrary to the manuscript evidence, Millikan announced in his paper: '*It is to be remarked, too, that this is not a selected group of drops but represents all of the drops experimented on during 60 consecutive days*…'." (Cantor 1989: 159; emphasis as in the original)

[40] For more on this, see Nagy C. (2014).

There are, of course, norms – partly formulated explicitly, partly only implicit – governing experiments as well as experimental reports. The fulfilment of the former, however, cannot be checked directly, but only indirectly, with the help of the latter. Nevertheless, as Cantor points out,

> "[…] a laboratory notebook and a published journal article are two very different literary forms, serving different purposes and subject to different conventions. The published version should not be viewed simply as a tidied up version of the laboratory notes, since the former contains many conventional elements that would find no place in the latter. The publication is a retrospective narrative, an impersonal, passive reconstruction which draws attention to those theories, tests and data which are considered appropriate for consumption by the scientific community." (Cantor 1989: 160)

This has significant consequences for the evaluation of experiments as data sources. It is the *scientific community* that decides whether the experimental results are reliable and epistemologically useful, that is, whether they can be used for theory testing, explanation, elaboration of new theories etc. This decision is based not on the analysis of the experimental procedure itself but only on the *judgement of the experimental report produced by the experimenter*. From this it follows that the rhetorical power of the latter is a decisive factor in this case, too.

Although this historical overview is somewhat fragmentary, it clearly shows that the norms related to the acceptability of experiments have undergone several changes. Moreover, there are different norms in different branches of science which are often contested. The same holds true of the rhetorical/argumentative aspects of experimental reports as well. That is, the structure and the rhetorical/argumentative tools applied in experimental reports are *social products* as well.

At this point, of course, the question arises of whether it is possible to elaborate a metascientific model of experiments in cognitive linguistics that is capable of accounting for the relationship between the experimental process and the (argumentative) experimental report. This model has to allow for *an evaluation of the reliability of the experimental process on the basis of the arguments presented in the experimental report*. The question is, however, how to elaborate such a model. As we have already mentioned in Section 1, contemporary philosophy of science rejects the idea of providing general, uniform norms for scientific theorising. In Section 3, we argued that this does not mean that we have to make a start from scratch. We showed that we can rely on the results of metascientific reflection on experiments carried out by philosophers of physics, psychology, social sciences, biology etc. because there is a strong enough *analogy* between them and the situation in linguistics. We will further elaborate on the metascientific model delineated in Section 3.3 by focusing on its argumentative aspects.

4.2. The nature and function of argumentation in experiments in cognitive linguistics

At this point, of course, the question arises of which metascientific tool enables us to reconstruct and evaluate the argumentation process which governs the experimental process and the experimental report, as well as the relationship between them. A purely logical analysis would not suffice because it is formal and is not capable of grasping information related to the

uncertainty and reliability of statements. This motivates the application of the *p-model*, elaborated in Kertész & Rákosi (2012). This theoretical framework might be suitable for fulfilling the task of modelling both linguistic theorising and experimenting. Since the p-model describes scientific theorising as a cyclic and prismatic, heuristic argumentation process, its extension to the metascientific reconstruction of experiments in cognitive linguistics seems to be natural.

4.2.1. The uncertainty of information in experiments: plausible statements

The p-model – following Rescher (1976) – does not interpret scientific hypotheses as propositions but assigns them a structure consisting of an *information content* and a *plausibility value*. The plausibility value of a statement indicates the extent to which its information content is supported, is made reliable by a given *source*, and as a result, to the extent to which we are ready to accept it.

In connection with experiments in cognitive linguistics, the following plausibility rankings may be applied – of course, this list is only a sample of possible rankings and is by no means comprehensive:

$|p|_D = 0$, that is, the statement p has neutral plausibility according to the experimental design abbreviated as D, if the experimental design does not allow the risk of even a rough estimate for the plausibility of the statement p, and neither p nor its negation is supported by D;

$|p|_K = 0.2$, that is, p has low plausibility according to the experimenter's knowledge abbreviated as K, if p is the experimenters' previous, untested and vague conjecture about the outcome of the experiment;

$|p|_{E_1} = 0.4$, that is, p has a rather low plausibility according to an earlier experiment abbreviated as E_1, if p results from an experiment, but some possible sources of noise which may cause systematic errors have not yet been ruled out with the help of control experiments;

$|p|_{E_2} = 0.6$, that is, p has a rather high plausibility according to an experiment E_2, if p results from a well-designed experiment with a thorough authentication of the perceptual data;

$|p|_T = 0.8$, that is, p has a high plausibility according to a theory abbreviated as T, if p is a central, generally accepted hypothesis of the given theory which has already been tested with the help of linguistic, corpus linguistic, experimental etc. investigations;

$|p|_M = 1$, that is, p can be regarded as true with certainty on the basis of a mathematical theory M, if p is a mathematical theorem proven in M.

It has to be stressed that low plausibility values do not mean a statement is improbable but rather that it has a relatively small, limited amount of plausibility (reliability, acceptance). In such cases, the source votes expressly for the given hypothesis. If a source is against a hypothesis then it makes its negation plausible and the given hypothesis *implausible* or even false with certainty; that is, in such cases $0 < |{\sim}h|_S \leq 1$.

The concept of 'plausibility value' allows us to represent and compare the acceptability (reliability) of statements such as previous conjectures, perceptual data, experimental data, hypotheses of linguistic theories, hypotheses about linguistic phenomena, etc. The experimenter's hypothesis about the correctness of the experimental design or about the flawless functioning

of the measuring devices can also be only plausible but not certainly true. From the experimenter's point of view this means that the non-public argumentation process which organises and conducts the experimental process deals with uncertain, fallible pieces of information. From the reader's perspective this means that the experimental report consists of plausible but, in most cases, not certainly true statements. Moreover, the concept of 'plausibility' makes it possible to compare the plausibility value which can be assigned to statements on the basis of the identification of their source on the one hand, and the value which they receive in the experimental report on the other. If the latter values are higher than the former, then this indicates an unwarranted overestimation of the plausibility of certain hypotheses or data and leads to a fallacious argumentation.

4.2.2. Sources of plausibility

We distinguish direct and indirect sources. In the case of *direct sources*, the plausibility of the statement at issue is evaluated with respect to the reliability of its source, as above. *Indirect sources* yield the plausibility value of the given statement on the basis of the plausibility of other statements – that is, via *plausible inferences*. Plausible inferences take into consideration not only the logical structure of the premises and the conclusion but their plausibility values and semantic structure as well. They always rest on a semantic relation: for example, causality, analogy, similarity, sign, necessary or sufficient condition, part-whole relation etc., and are not necessarily deductively valid.

The perfect identification of the direct and indirect sources from which the plausibility of the data and other hypotheses in experimental reports originate makes it possible to check and re-evaluate the plausibility of the statements at issue. Specifically, the reconstruction of the plausible inferences (indirect sources) applied in the experimental report may reveal latent background assumptions that are implausible instead of being plausible or of neutral plausibility. It may happen that an inference relies on a hypothesis that is solely a conjecture but which on closer examination turns out to be implausible or false. In such cases the conclusion loses its plausibility as well – and the same holds true of the inferences that made use of the conclusion of this inference as a premise. This kind of reconstruction may be especially useful for the authentication of the perceptual data as well as for establishing a link between the experimental data and the hypotheses of a theory. In both cases the connection between the perceptual data and the experimental data and between the experimental data and theoretical hypotheses relies mostly on deductively invalid plausible inferences that make use of latent background assumptions.

4.2.3. Conflicting information in experiments: p-inconsistency

An important property of the above concept of plausibility is that it allows a statement to be plausible on the basis of some sources and implausible on the basis of others at the same time. Such cases are called *p-inconsistencies*.

Thus, a hypothesis may be made plausible by an experiment as a source but implausible by another one. Similarly, different theories may judge the acceptability of a given scientific claim differently, or an experiment may refute a prediction, etc. – leading to different cases of p-inconsistency.

The decision between conflicting hypotheses cannot be reduced to the mechanical comparison of their plausibility values. Instead, one has to evaluate statements along with the reliability of the sources making them plausible, their relationship to other statements, to the related methodological norms and so on – that is, the system of relations of the rival hypotheses has to be revealed and compared as a whole. Such constellations are called *the p-context*.

According to the p-model, inconsistencies must not be viewed as fatal failures but indications that either the experiment or the theory at issue (or even both) is in need of some kind of modification. Thus, conflicting experimental results, contradictions between predictions and experimental data, inconsistencies between the hypotheses of a theory and the results of an experiment, and other discrepancies among the components of the experimental process are concomitants of experiments. Nevertheless, there are always several possible causes of a conflict, whose identification may require several attempts. In most cases, inconsistencies are not resolved by simply giving up one of the conflicting statements but more comprehensive revisions are needed that may affect further components of the experiment as well.

4.2.4. Solutions and the resolution of p-inconsistencies

In order to resolve a p-inconsistency, one has to *re-evaluate* the p-context. A *solution of a p-inconsistency* is achieved if a p-context has been arrived at in which (a) the statement in question is unanimously supported or opposed by the sources – that is, it has become either plausible or implausible (or even certainly true or false) on the basis of *all* sources in the given p-context –, or (b) the statements causing inconsistency are represented separately and this separation is systematic and well-motivated.

It is possible, however, that a p-inconsistency has several solutions. This necessitates the introduction of the notion of the *resolution of a p-inconsistency*. This means that one finds a solution of the given p-inconsistency which is, when compared with other solutions, the best on the basis of a particular set of accepted criteria, and according to the information available for us in the given p-context. It may be the case, however, that in a given information state one can only show that for the time being there is no resolution achievable.

It is of vital importance that inconsistencies are not put aside without finding a solution which makes it possible to separate the conflicting statements, at least provisionally. Instead, one has to try to elaborate and compare as many solutions as possible in order to find the best solution available under the given information state.

The reliability of an experiment as a data source is largely determined by careful and strict identification of the inconsistencies among its components, by the number, variety, and comprehensiveness of the investigated solutions, as well as by the choice of the resolution of the conflicts revealed during the experimental process. Since the p-model describes several strategies of inconsistency resolution, its application may contribute to the elaboration and conduct of better experiments in linguistics.

4.2.5. Cyclic revisions in experiments: plausible argumentation

To achieve the solutions or the resolution of a given p-inconsistency, one needs a *heuristic tool* that makes it possible to re-evaluate the p-context and to find and compare the solutions to its problems. This heuristic tool is what we will call *plausible argumentation*. In simple terms, plausible argumentation is the transformation of a problematic p-context into one that is no

longer (or at least, less) problematic. This involves the successive re-evaluation of a problematic p-context by the elaboration of possible solutions to its problems, the evaluation of the alternative solutions and the comparison of the latter. Its aim is the detection of all available solutions and the decision as to which of them is to be accepted as the resolution of the given p-problem.

The above characterisation of plausible argumentation indicates that the argumentation process is basically not linear but *cyclic*, because the re-evaluation of a problematic p-context usually does not lead immediately to an unproblematic one but may raise new problems. This may require the revision of previous decisions, the assessment of other alternatives etc. Therefore, throughout the argumentation process one returns to the problems at issue again and again, and re-evaluates the earlier decisions about the acceptance or rejection of statements, the reliability of the sources, the plausibility values of the statements, the workability of methodological norms, the conclusions previously reached by inferences etc.

The p-model's concept of 'plausible argumentation' allows us to interpret both the argumentation organising and conducting the experimental process and the experimental report as pieces of plausible argumentation. The experimental report should not simply summarise and make public the results of the former but make it possible for the reader to continue the non-public argumentation process. That is, a good experimental report is informative enough to allow the reader to add new argumentation cycles to the non-public argumentation process.

Thus, for example, the reliability of an experiment crucially depends on the question of the extent to which the experimental data may be supposed to be free of systematic errors. In experiments on metaphor processing, by the application of an offline measure, participants might have made use of conscious strategic considerations distorting the results, or semantic priming effects might have led to faulty results, etc. Therefore, when the experimenter suspects or reveals the presence of such a factor, he/she has to carry out control experiments and/or revise the experimental design and start a new cycle of revision. Nevertheless, one cannot check and rule out the presence of every possible systematic error. The set of the factors that might have influenced the outcome of the experiment is always open. From this it follows that even a good experiment may contain errors that can be revealed only later by some other member of the scientific community. Thus, good experiments are characterised not only by the thoroughness of the elimination of possible errors but are also inspirational and motivate the search for more complex explanations of the investigated phenomena. They pave the way for new experiments that take into consideration further factors and for the elaboration of more refined theoretical models.

5. The reliability of single experiments as data sources in cognitive linguistics

In Sections 3 and 4 we presented components of a metascientific model with the help of which the structure of experiments in cognitive linguistics can be reconstructed and their argumentative aspects can be described. In this section we will set these tools to work and show how they might contribute to the analysis and evaluation of experiments. First, we will clarify the principles of the evaluation of experiments in cognitive linguistics and present a detailed guide illustrating the steps of the evaluation process. Then, the elaborated system of criteria will be applied to a series of small-scale case studies. The last subsection will summarise and generalise the moral to be drawn from these case studies.

5.1. Criteria for the evaluation of experiments in cognitive linguistics

As is well-known, experimental research in cognitive linguistics is characterised by a considerable diversity of approaches and experimental methods, as well as contradictory and often controversial experimental results. Raymond W. Gibbs offers a two-step diagnosis of this situation. First, he claims that "psycholinguistic experiments may be […] inherently flawed as a scientific enterprise" (Gibbs 2013: 45). Second, he raises the hypothesis that with the help of his alternative metascientific model, it is possible to "push metaphor scholars closer to thinking and practices seen in more mature scientific disciplines" (Gibbs 2013: 52).

In contrast, in Dirk Geeraerts's view, experiments apply feasible, well-established procedures providing completely reliable experimental results:

> "[...] there is a common, commonly accepted way in psycholinguistics of settling theoretical disputes: experimentation. Given a number of conditions, experimental results decide between competing analyses, and psycholinguists predominantly accept the experimental paradigm as the cornerstone of their discipline." (Geeraerts 2006: 26)

Hasson and Giora (2007) take another route: they provide us with a comprehensive overview of the experimental methods applied in cognitive linguistics, summarising their rationale and identifying their possible weak points. Their list can be profitably complemented with Keenan et al.'s (1990), Haberlandt's (1994) and Kaiser's (2013) considerations. This combined inventory, however, still cannot be regarded as a system of guidelines, mainly due to the circumstance that all three papers focus on the detailed characterisation of the basic hypotheses and working mechanism of the different experimental methods. Therefore, they provide neither a systematic nor an exhaustive typology of errors but discuss the most typical problems related to the different types of experiments.

This disagreement might motivate a twofold strategy. Namely, metascientific reflection on the nature and limits of experiments in cognitive linguistics should be based on the continuous comprehension and adjustment of *insights gained by philosophers of science studying experiments in science* (i.e., a model of scientific experiments in general) on the one hand, and the *reflection on the research activities of linguists while working with experiments* (that is, criteria related to the experimental methods used in linguistics, in particular), on the other. Both components are vital. First, linguists often confuse workable and generally applied norms of

natural sciences with outmoded and untenable tenets of the standard view of the analytical philosophy of science.[41] Second, contemporary philosophy of science does not strive to elaborate universally valid, normative accounts of scientific experimenting. Instead, research practice is studied carefully and closely, and methodological rules or norms are held to be *field-sensitive* and put into a *historical context*.

Experiments involve many potential sources of error and undetected possibilities. Therefore, it is vital to take the fallibility of experiments seriously and search for means which enable us to reduce it. In this section, the list of well-known criteria put forward by cognitive scientists and psycholinguists will be integrated into the metascientific model delineated in Sections 3.3 and 4.2. The proposed system of criteria will be applied to experiments on metaphor processing conducted between the years 1989 and 2004 in order to exemplify their workability.

The key question is, how to decide when an experiment is to some extent reliable as a source and yields plausible (but not certainly true) experimental data and when it is unreliable as a source and is not capable of providing plausible data. A concomitant question is whether the experimental data gained are capable of providing evidence for or against the theory or theories at issue – that is, whether there is a strong enough link between the experimental data and the hypothesis/hypotheses of the theory or rival theories. On the basis of the model presented in Sections 3.3 and 4.2, *the evaluation of experiments in cognitive linguistics* involves the following steps.

1) Reconstruction of the stages of the experimental process in the experimental report. Although the experimental report can only provide an informationally reduced picture of the experimental process, both the accomplishment of the diverse stages of the experimental process and the cyclic returns conducted by the experimenter in order to eliminate problems revealed should be presented in a detailed enough fashion so that the steps taken can be identified and analysed.
2) Re-evaluation of the experimental design. The experimental design should be presented in such a way that the reader is capable of repeating the related thought experiment and checking its validity (including its construct validity, content validity and criterion validity). For example, it should be possible for the reader to check whether the experiment is capable of eliciting participants' natural linguistic behaviour; expectancy effects can be ruled out; semantic priming does not influence participants' performance; participants do not make use of strategic considerations, post-reading checks, or their own implicit theories about the related linguistic phenomena instead of relying on their spontaneous linguistic behaviour, etc.[42]
3) Re-evaluation of the experimental procedure, the authentication and interpretation of the perceptual data. The experimental report usually contains hints at revisions of the original experimental design or the experimental procedure. Thus, the evaluation of the experiment has to examine whether possible error sources have been revealed, and whether their impact on the results has been controlled with the help of control experiments or statistical tools. The interpretation of the perceptual data has to take into consideration, among other things, that there is

[41] Indeed, it must be mentioned that natural scientists are also prone to making the same error.
[42] For details, see Kaiser (2013: 139, 141, 143), Haberlandt (1994: 9, 18), Hasson & Giora (2007: 305, 311, 316), Keenan et al. (1990: 384).

always only an indirect link between the perceptual data obtained and the linguistic phenomena investigated (such as mental processing of metaphorical expressions). Further, the statistical analysis of the perceptual data is a complex and formidable task with many problematic points, pitfalls and alternatives. Therefore, the conduct of statistical control analyses, alternative analyses and meta-analyses is vital. A further important point is checking the reliability (generalizability) of the results.[43]

4) Re-evaluation of the plausibility of the experimental data and their confrontation with the theory/rival theories. Since experiments are not completely reliable data sources, they may produce only plausible results. The strength of the support or counter-evidence they may provide to a hypothesis/theory depends on two things: the plausibility of the experimental datum itself, and the strength of the link between the hypothesis/theory and the experimental data. Thus, for example, it has to be checked whether the plausibility value of the experimental data and other data/hypotheses made use of in the experiment is not overestimated in the experimental report; the experimental data (which result from and are bound to a certain situation) can be generalised; alternative explanations can be ruled out (so that the experimental data support only one of the rival hypotheses/theories), etc.

5) Proposals for the continuation of the experimental process by new cycles. Since the meta-scientific model of experiments presented in Sections 3.3 and 4.2 interprets experiments in cognitive linguistics as open and cyclic processes, the analysis and evaluation of experiments is nothing other than the continuation of the experimental process by new argumentation cycles, and, if possible, the elaboration of proposals for the continuation of the experimental process. Thus, the core of the analysis and evaluation of experiments are *thought experiments*: one tries to imagine whether and how the experiments described in the experimental report took place and what might have happened, whether there might have been problems which could have distorted the results, etc.

6) Conduct of replications or modified versions of the experiment. Thought experiments are, of course, fallible and have their limitations. Thus, while in certain cases such analyses may provide relatively strong counter-arguments (but no ultimate refutations!) which seriously question the reliability of the experiment at issue, in other cases they only indicate weak points and suggest a control experiment or some kind of revision. Similarly, post hoc statistical analyses of the experimental data are not decisive but have to be taken seriously. Consequently, it might be necessary to transform these thought experiments into real experiments: into a repetition of the original experiment or into a revised version of the experiment, and then compare their outcomes. This means that linguists should not only make their experiments replicable, but that *actual replications* are needed either in an unaltered form or following modifications of the original experimental design.

7) Comparison of the experimental data with the results of earlier experiments. Experimental data originating from different experiments cannot be compared mechanically but more sophisticated tools have to be applied – for instance, statistical meta-analyses have to be performed.

[43] I use the term 'reliability' here in the traditional, narrower sense – that is, it refers to the generalizability of the results to other situations.

To sum up, the evaluation of the weight, impact and treatment of the problematic points of experiments requires the analysis and re-evaluation of all details of the given experiment. There are minor flaws that merely *decrease the plausibility* of the affected experimental data, while there are other errors that have to be deemed serious faults that question the usability of the data gained or even *make the experiment unreliable as a data source*. Thus, the evaluation of the experiment can and should be accomplished in such a way that not only is its reliability as a data source judged but possible improvements are proposed which, during further cycles, may lead to the continuation and re-evaluation of the experimental process and result in (more) plausible experimental data. See Figure 4.

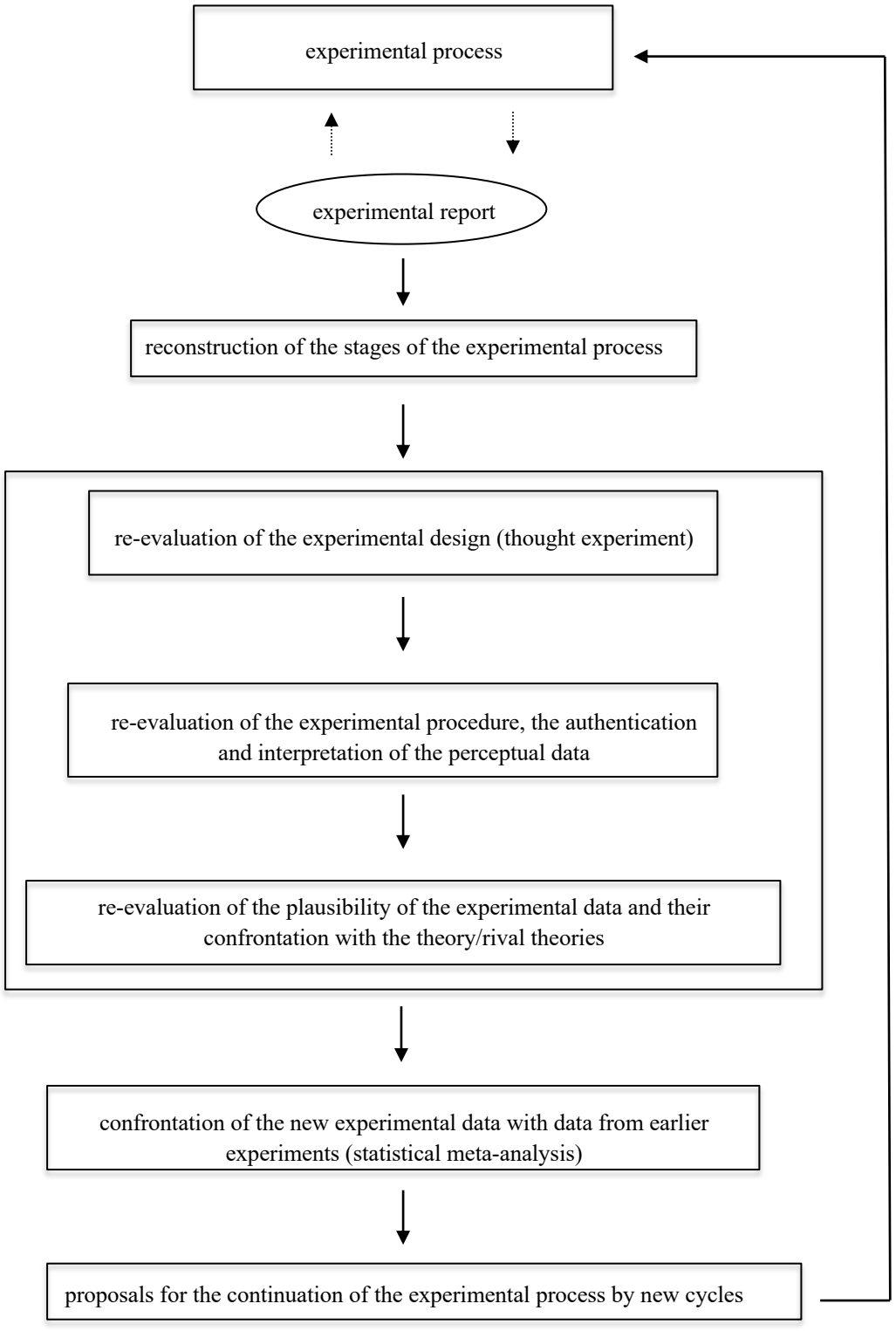

Figure 4. The evaluation of experiments in cognitive linguistics[44]

One might raise the objection that some of the steps proposed do not provide radically new criteria but rather summarise well-known requirements. Clearly, the collection and systema-

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

tization of well-established methodological rules, fruitful practices, insights from the philosophy of science, experiences of scientists working in other fields of research, etc. is inevitable, but clearly not sufficient. What is needed, is to *put them to work*.

Nevertheless, it is important to bear in mind that the model of experiments in cognitive linguistics presented in Sections 3.3 and 4.2 and the criteria of evaluation based on it are not a "Wunderwaffe" solving all problems of linguistic experimenting. Therefore, their application does not provide general methodological rules which could be used in every situation, must not be violated, and would guarantee flawless and totally reliable results. Indeed, although experiments are fallible and can provide only plausible experimental data, this does not mean that the above criteria can be violated without consequences. All possible error sources and problems have to be revealed and examined as thoroughly as possible; no weak point and no infringement of the norms should be concealed or ignored. This does not mean that experiments burdened with problems should be immediately rejected; they have to be given appropriate attention and their possible solutions have to be elaborated and compared – or, if this is not possible on the basis of the information at our disposal, this finding has to be declared.

The task of Section 5.2 will consist of showing the workability of these ideas with the help of the evaluation of experiments on metaphor processing conducted between 1989 and 2004. Since Section 6 deals with the theoretical and practical problems related to replications of experiments in cognitive linguistics, only those experiments will be analysed in this section for which no replication is available yet.

5.2. Case study 2: Analysis and re-evaluation of single experiments on metaphor processing

In the present section, several experiments will be analysed and re-evaluated with the help of the system of criteria presented in the previous section. The aim of the analyses is not to provide a comprehensive overview of the current stand or the history of experiments on metaphor processing but to illustrate the metascientific views advanced. Further, the analyses are not intended 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 would allow us to avoid the errors revealed and/or to increase the reliability of the experiment at issue. Finally, we will summarise the results.

It is important to emphasise that the aim of these analyses is not a 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 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.

5.2.1. 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–).[45] According to the author's hypothesis, if metaphorical interpretation is constructed – in contrast to the traditional view on metaphors, but in harmony with Glucksberg et al's view – in an obligatory, involuntary manner, then decisions 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.[46] 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 had to push a key after having read a line. There were also some quiz questions in order to engage participants' attention. Keysar predicted 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

---

45 For example, one L+/M– story was the following:
"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."

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

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 hypothesis which states 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 again subjects 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, in contrast, it seems to be possible to elaborate and conduct a revised version.

5.2.2. 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 Lakoff and Johnson's Conceptual Metaphor Theory) 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 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 experiments in favour of Conceptual Metaphor Theory (CMT). Namely, they fall short of construct validity and cannot be regarded as reliable data sources.[47] In this case, the thought experiment checking the experimental design is sufficient to reject the idea of possible developments, too.

5.2.3. 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

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

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 (CMT-inconsistency), '1' indicated ambiguity in this respect, while '2' indicated that there is clear reference to the source domain (CMT-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 perceptual data McGlone inferred that participants did not rely on conceptual metaphors as a knowledge source when they interpreted metaphorical expressions.

Evaluation: Like other off-line methods, this experiment yields 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 5.2.2. Nonetheless, 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 on a 7-point scale.[48] With each target metaphor (such as *Dr. Moreland's lecture was a three-course meal for the mind*), 9 possible alternative

[48] Excerpts can be found in Appendix B of McGlone's paper.

metaphors were provided; 3 were CMT-consistent (*Dr. Moreland's lecture was a smorgasbord for the mind* – same CMT source domain), 3 CMT-inconsistent but attributively similar (*Dr. Moreland's lecture was a full tank of gas for the mind* – different CMT 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 base/vehicle[49] – similar in the sense that they belong to the same attributive category – would provide evidence for Glucksberg's Attributive Categorization View (henceforth: ACV).

Evaluation: First, the indirectness of this experimental method is greater than it was with Experiments 1 and 2.[50] 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 Conceptual Metaphor Theory (preference of CMT-consistent metaphors) and the predictions obtained by the Attributive Categorization View are in conflict with the experimental data. In contrast to this, McGlone draws the consequence that the results are in conflict with the CMT but they are consistent with Glucksberg's ACV.

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 base/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 Categorization 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.

---

[49] Glucksberg's ACV uses the term 'vehicle', Gentner's CMT the term 'base'.

[50] 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)

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;[51] 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. Specifically, 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 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 concern native speakers' conscious strategies of metaphor interpretation.

5.2.4. Bowdle & Gentner (1999)

Description: In order to test Gentner's Career of Metaphor Hypothesis (CMH), the authors developed a two-stage experimental design. In the first, study stage, participants saw pairs of novel similes using the same base/vehicle term and they had to fill in a target/topic term in a third example of the same structure.[52] The authors' hypothesis was that priming with novel similes using the same base/vehicle 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 CMH, 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/vehicle 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/vehicle 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,

---

[51] 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).

[52] For example:

An acrobat is like a butterfly.
A figure skater is like a butterfly.
_____ is like a butterfly.

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

Summing-up: Bowdle and Gentner constructed 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.

5.2.5. 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 Glucksberg's ACV, while the latter from Gentner's CMH. Participants had to decide whether the statements presented 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, differing considerably 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 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, 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 metaphors with low-constraint targets/topics, and high-similarity metaphors are often also metaphors with high-constraint targets/topics.[53] Therefore, the experimental data are not capable of discriminating between predictions based on Gentner's CMH and Glucksberg's ACV. 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.

---

[53] 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.*

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.

5.2.6. Gernsbacher, Keysar, Robertson & Werner (2001)

Experiment 1

Description: The authors intended to test the hypothesis that the basic-level meaning of the base/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 target/topic was changed for a member of the basic-level category represented by the base/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 IPAM and the above hypothesis hold, then participants should verify superordinate-level property statements more rapidly after metaphor-prime sentences 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.[54] The explanation may have tempted subjects to interpret their task in such a way that they had 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

[54] "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/)

CMH and Glucksberg's IPAM. For example, if metaphor processing involves structural alignment between the target/topic and the base/vehicle as supposed by Gentner, then mentioning a property of the base/vehicle which cannot be placed in relation to the target/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 base/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 base/vehicle, were used in order to overcome the last problem relating to Experiment 2. This modification, however, leads to another problem: namely, that while the metaphorical prime and the basic-level target both contain the base/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.

5.2.7. Gibbs, Lima & Francozo (2004)

Description: American and Brazilian participants 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.[55] 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.[56] 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 Conceptual Metaphor Theory. 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.

5.3. Re-evaluation the reliability of experiments as data sources in cognitive linguistics

In the previous section, we have seen how the application of the criteria proposed in Subsection 5.1 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 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 *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 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 experiments in cognitive linguistics is their stimulus material. This provides a further argument for the requirement that

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

[56] 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)

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 experiments in cognitive linguistics because they might reveal the aim of the experiment more easily, or rely on some linguistic theory instead of their pure linguistic intuition.
– 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.
– As for the introductory sections of papers dealing with experiments in cognitive metaphor research, it is often the case that authors 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 in many cases presented as unquestionable facts. A correct and balanced presentation of rival theories should be attempted.

These proposals cannot, of course, guarantee that experiments will provide incontestable data for theories about metaphor processing. From the perspective of the metascientific model of experiments as presented in Sections 3.3 and 4.2 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 and these data sources should be banned from cognitive linguistic research. 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.

6. Metascientific modelling of chains of closely related experiments in cognitive linguistics

As we have mentioned in Section 2, experiments have not only an inner life but also *a social life*. That is, one of the sources of the uncertainty of experiments is their relationship to other experiments. This section aims to provide tools for the reconstruction and evaluation of complex structures involving experiments so that we can determine how the results of closely related experiments influence the plausibility of the experimental data. First, we will provide an overview on the situation related to replications in cognitive linguistics and argue for the seriousness and acuteness of the replication problem. Section 6.2 will present the first part of a quasi-historical case study by offering a first concise description of an experiment, its replications, and the related counter-experiments. "Quasi-historical" means that although these experiments are quite old and their results and methods have been heavily criticised (see Section 11 and 15 on this), they are still referred to in current literature on metaphor processing as evidence for or against mainstream theories of metaphor processing. In Section 6.3, the metascientific model of experiments in cognitive linguistics presented in Sections 3 and 4 will be extended in such a way that the relationship between original experiments and their repetitions can be described. We will illustrate the applicability of this model with the help of the second part of the case study Section 6.4, as the extended model will be applied to the replication attempts delineated in Section 6.2.

6.1. Replications in cognitive linguistics

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 it 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)

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

In spite of this, the vast majority of experiments in cognitive linguistics 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 in 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 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* (2015) 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. (2015) and 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: 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 experiments as data sources in cognitive linguistics dubious. Therefore, the common practice should be rethought and new 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)

Therefore, the question of what role replications play in the evaluation of experiments in cognitive linguistics is one of the most significant open questions in the philosophy of linguistics. We do not intend to answer this question in general, but we will rely on an instructive *case study* by analysing various replication attempts conducted within cognitive metaphor research.

6.2. Case study 3, Part 1: An experiment on metaphor processing and its replications

6.2.1. The original experiment: Wolff & Gentner (1992)

Experiment 1: Participants were shown either the target/topic or the base/vehicle 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 Categorization View (ACV) is correct, and metaphor processing starts asymmetrically, by the derivation of a category from the base/vehicle term which is then applied to the target/topic term, then base/vehicle primes should be more effective than target/topic primes. This prediction, however, was not supported by the data obtained: there was no significant difference between base/vehicle and target/topic.

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/vehicle prime *butcher* was presented as the sentence *A something is a butcher*, while if the target/topic 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 difference between 'both' and 'base/vehicle' 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/topic and base/vehicle. There were significant differences between the conditions 'blank' and 'base/vehicle', 'both' and 'base/vehicle', and 'both' and 'target/topic', while no significant difference was detected between 'target/topic' and 'blank', and 'base/vehicle' and 'target/topic', respectively. These results were again found to be incompatible with Glucksberg's ACV but in harmony with Gentner's SMT.

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/vehicles, is a factor that facilitates a processing mode that starts with the base/vehicle term. While the first two experiments used novel metaphors, this experiment employed only bases/vehicles 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/vehicles.

6.2.2. 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 targets/topics were either high-constraint (*lawyer, mind*) or low-constraint (*my brother, life*) and bases/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 experiments – 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 6.2.1, from

the Attributive Categorization View Wolff and Gentner inferred the prediction that base/vehicle primes should be more effective than target/topic primes. In contrast, according to the authors' predictions, high-constraint targets/topics and unambiguous bases/vehicles should be effective primes for metaphor comprehension, while low-constraint targets/topics and ambiguous bases/vehicles should be ineffective or less effective. The experimental data clearly support the latter hypothesis.

6.2.3. 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 SMT, but inconsistent with ACV, since bases/vehicles were not quicker than targets/topic, and metaphors preceded by both primes were faster than bases/vehicles or targets/topics 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/vehicle conventionality and low relational similarity between base/vehicle and target/topic in order to secure ACV maximally advantageous conditions against Gentner's 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 Categorization View, 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/vehicle conventionality, relational similarity and prime type (both, base/vehicle, target/topic, blank). The ISI between prime and metaphor was 0 ms in this experiment. According to the authors' predictions, if the ACV were correct, then there should be a base/vehicle advantage under all conditions. In contrast, from Gentner's newly developed Career of Metaphor Hypothesis it follows that there should be no base/vehicle advantage for low-conventionality metaphors, there should be a base/vehicle 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 ACV, into a more complex version of SMT.

6.2.4. 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 target/topic is a member of the category defined by the base/vehicle. They had to press button 1 for "non-member", 2 for "partial member" and 3 for "full member". According to Glucksberg's ACV, both novel and conventional metaphor-primed items should have higher ratings than literal controls, while Gentner's CMH 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 re-formulated 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.

6.2.5. Interim summary

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 follow-ups 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.

6.3. The relationship between original experiments and replications: Experimental complexes

As we have seen in Sections 1 and 4.1, contemporary philosophy of science does not strive to stipulate generally valid norms for scientific theorising, such as verifiability, falsifiability, etc. Instead, field-sensitive methodological guidelines are elaborated in historical contexts and on the basis of a close and careful study of research practice. This approach fits into this tendency. Its main motivation was to grasp a specific characteristic of experiments in cognitive metaphor research. Namely, in this research field, most papers publishing experimental results involve – in contrast to other branches of science such as physics, medicine, or chemistry – not only one experiment but 3-4 similar experiments, the relationship of which, however, is not clear. They are usually not complementary but rather seem to be improved versions of one another. Despite this, their results are often interpreted in such a way that they reinforce each other and provide converging evidence. If they were regarded in fact as improved versions of each other, then only the last member of such a chain of experiments should be taken into account and made public.

If we summarise the moral of the remarks relating to the experiments presented in Section 6.2, we can reach the conclusion that most replication attempts are *not exact repetitions* but involve some kind of *modification*. Thus, they can be described neither as 'multiple repetitions of the same experiment' nor '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 experiments in cognitive metaphor research has to transgress the boundary of single experiments*. This motivates the elaboration of the concept of the 'experimental complex':

(EC) 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 may involve, among other things,

- *modified (improved) versions* of the original experiment,
- *exact replications* of the original experiment or one of its non-exact replications,
- *control experiments* intended to rule out possible systematic errors in the original experiment or in one of its modifications,
- *counter-experiments* which make the most radical revision to the original experiment by applying a different method (experimental paradigm) to the same stimulus material in order to provide evidence against the research hypothesis at issue,
- a *wider set of perceptual and experimental data*,
- *diverse perspectives* by adherents of different theories,
- *different versions of the research hypothesis*, but also
- *conflicts* emerging from different evaluations of the outcome of the original experiment (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 5 shows, experimental complexes have basically the same cyclic structure as single experiments:

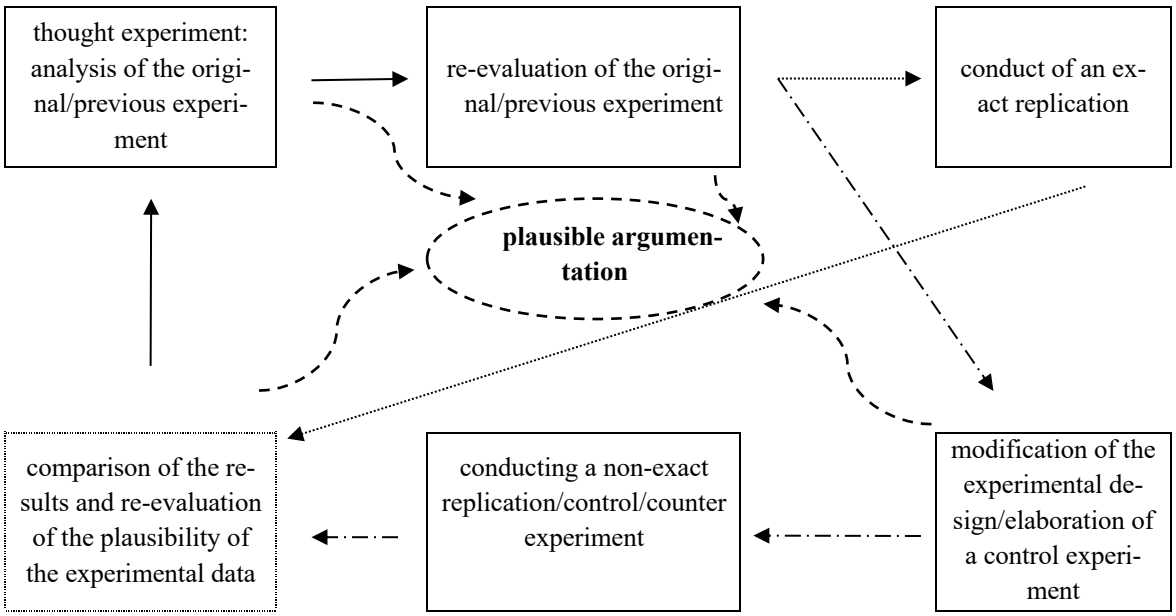

Figure 5. The structure of experimental complexes[57]

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

(LEC) 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).

It is always the limit that provides the *most plausible* experimental data within the given experimental complex, because limits are free of known problems and are also reliable. To that end, however, (LEC) 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:

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

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

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 problems (systematic errors) but also lead to the emergence of new ones. Against this background, one can distinguish between progressive and stagnating non-exact replications:

(PEC) 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 re-evaluations of the original experiment, mostly produce more plausible experimental data, and may bring us closer to a 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.

Nevertheless, it is not the case that every progressive replication produces more plausible experimental data. The reason for this lies in the circumstance that any modification may not only rule out possible (systematic) errors but can also lead to the emergence of new problems, which, in addition, may be more serious than the resolved problem was, or may even turn out to be fatal. Thus, a progressive replication may solve a problem but also induce a dead end at the same time. Moreover, it is not always the case that non-exact replications provide increasingly similar results in the long run: quite often the opposite of this happens and the conflicts deepen and/or multiply.

There are three basic types of scenarios:

1. The experimental complex is convergent. See Figure 6:

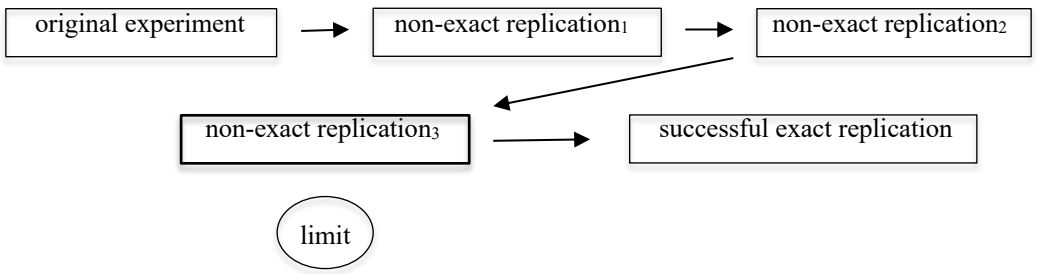

Figure 6. Convergent experimental complex

2. The experimental complex is divergent because the final non-exact replication is not reliable. See Figure 7:

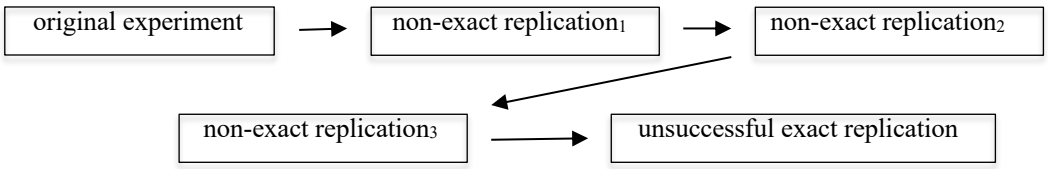

Figure 7. Divergent experimental complex due to unreliability

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 8:

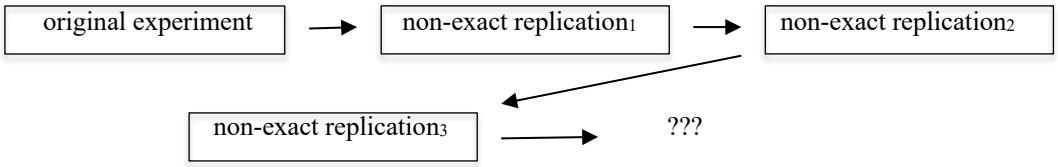

Figure 8. Divergent experimental complex due to unsolved problems

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. From this, however, it would be premature to conclude that replications are ineffective tools of problem-solving. The point is that *effectiveness – in contrast to progressivity – can be judged only in the long run*.

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). Further, experiments belonging to different experimental complexes may be similar enough to provide converging or diverging evidence for a research hypothesis if they rely on different experimental designs but test the same research hypothesis by investigating the relationship between the same variables. We will call such experiments *methodological variants*, since they apply different methods to estimate the strength of relationship between the same variables. The detailed description of such constellations, however, should be the subject of another work.

In the next section, we will reconstruct the experiments briefly presented in Section 6.2 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);[58]
– 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 6.4 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.

6.4. Case study 3, Part 2: Reconstruction and re-evaluation of an experimental complex

As Figure 9 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):

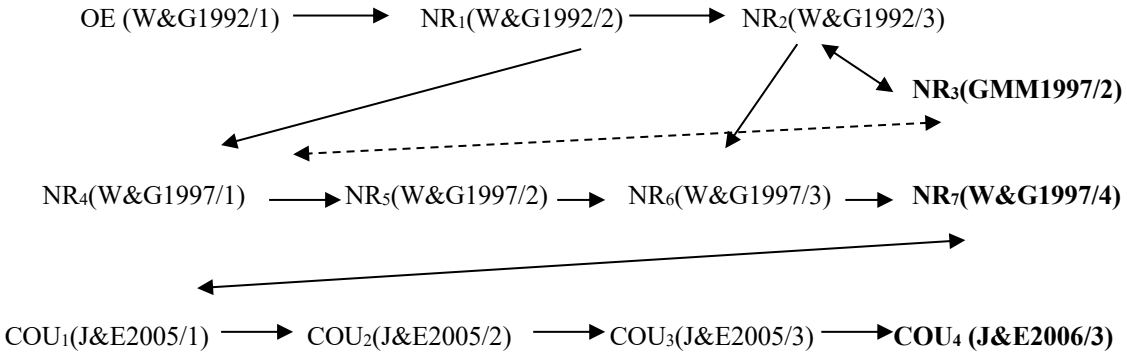

Figure 9. The structure of the experimental complex evolving from Wolff & Gentner (1992)

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 6.2.2), NR_7 (Section 6.2.3), as well as a limit-candidate of a series of counter-experiments: COU_4 (Section 6.2.4). The first step of the re-evaluation of this experimental complex should be the reconstruction and analysis of the three chains of experiments.

[58] See Sections 3.3 and 4.2.

6.4.1. The limit-candidate by Gentner and Wolff

OE (cf. Section 6.2.1): 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 would be a base/vehicle or a target/topic. 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/vehicle 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 6.2.1): 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/vehicle and target/topic is symmetrical, then there should be a significant difference not only between bases/vehicles and blanks but also between targets/topics 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/vehicles and blanks but also between targets/topics 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 6.2.1): 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/vehicle 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/vehicle conventionality is a factor that cannot be balanced, because there are no conventional target/topic 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 6.2.3): Experiment 1 in Gentner & Wolff (1997) is a progressive non-exact replication of NR_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 6.2.3): In Experiment 2 of Gentner & Wolff (1997), 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/topics or bases/vehicles 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: 232).

NR_6 (cf. Section 6.2.3): 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/vehicle 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 experiments 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/vehicles 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/vehicles 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 6.2.3): 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:[59]

	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. Overview of the re-evaluation of the limit-candidate by Wolff & Gentner

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.

[59] 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.

6.4.2. The limit-candidate by Glucksberg, McGlone & Manfredi

NR_3 (cf. Section 6.2.2): The progressivity of the non-exact replication of the original experiment OE_1 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,[60] the aim of the experiment was not masked.

Problem 15: Both high-constraint targets/topics and unambiguous bases/vehicles are less susceptible to causing false starts than low-constraint targets/topics and ambiguous bases/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 Interactive Property Attribution Model it would follow that unambiguous bases/vehicles should be faster than high-constraint targets/topics. Second, the experimental data obtained seem to be consistent with Gentner's Structure Mapping Theory, since both high constraint targets/topics and unambiguous bases/vehicles offer fewer properties for matching and both may facilitate the projection of candidate inferences from base/vehicle to target/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:

	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. Overview of the re-evaluation of the limit-candidate by Glucksberg et al.

6.4.3. Counter-experiments by Jones and Estes

COU_1 (cf. Section 6.2.4): 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, however, questioned by the following problems:

[60] 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 stage* or the common expression *Some lawyers are sharks*." (Glucksberg et al., 1997: 61)

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.[61] 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.[62] While the metaphors had a nominal structure and stated that the target/topic is a kind of the base/vehicle, inviting a 2- or 3-rating, the literal counterparts interpret target/topic and base/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 IPAM states that metaphors are interpreted in such a way that the target/topic belongs to an *ad hoc* category, and the base/vehicle is a typical member of this category. It is not required, however, that this *ad hoc* category is that of the base/vehicle *per se*; instead, the base/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 6.2.4): 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 6.2.4): 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

---

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

[62] 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)

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 6.2.4): 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,[63] 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 bases/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:

	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. Overview of the re-evaluation of the counter-experiments by Jones & Estes

Moreover, Problem 17 is a problem which calls the plausibility of the data obtained from these experiments into question.

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

6.4.4. Interim summary

The reconstruction of the three chains of experiments shows that the experimental data originating from experiments by the same researchers have, in most cases, become more plausible.[64] 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.

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

7. Conclusions: The reliability of experiments and experimental complexes as data sources in cognitive linguistics and a possible resolution to (RPE)

The moral of our considerations related to the *inner life* of experiments in cognitive metaphor research in Section 5 can be generalised as follows:

- The structure of experiments in cognitive linguistics basically *corresponds to the structure of experiments in the natural sciences*, although the role and impact of certain stages diverge.
- Experiments in cognitive linguistics should be viewed as *cyclic open processes*. This means that the plausibility of the statements related to different stages of the experimental process should be re-evaluated again and again during the elaboration and conduct of the experiment, as well as during its evaluation.
- The analysis and evaluation of experiments is nothing other than *the continuation of the experimental process by new plausible argumentation cycles*, and, if possible, the elaboration of *proposals for its continuation by new experimental cycles*.
- The conduct of the proposed revised version(s) of the original experiment and the comparison of the results obtained may lead to *more elaborated experiments* and *more reliable experimental data*.
- This also means that it is not only the "discovery phase" (the experimental process) that contains justification but the "justification phase" (the evaluation of experiments) also involves discovery.

On the basis of our considerations as well as the case study in Section 6, we can extend the above set of methodological guidelines on the reliability of experiments and data sources to the *social life* of experiments by clarifying what role replications play in the evaluation of experiments in cognitive linguistics:

- 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.
- 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.
- 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 an even stronger piece of evidence than one using the same stimulus material.
- 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.
- 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 results. Of course, 'multiple determinations of experimental results' (cf. Section 6.3.2), that is, experiments belonging to other experimental complexes, may increase the plausibility of the results, too.

– 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.
– Through the replications, revisions and improvements, experiments should become *collective works of a research field* and not private affairs of single minds.

As we have seen in Sections 3.3 and 4.2, experiments have a *dual argumentative structure*. The experimental process is organised and conducted by *a non-public* plausible argumentation process that is then transformed into the experimental report, that is, a *public* piece of plausible argumentation. This transformation can be regarded as acceptable if it does not change the plausibility value of the data, hypotheses and other statements of the original, non-public argumentation. This, of course, cannot be checked. Nonetheless, if the experimental report contains all information that might be relevant for the evaluation of the steps of the experimental process, then the reader may be in a position to reconstruct and re-evaluate the non-public argumentation process to the greatest extent possible, and compare it to the public argumentation process in the experimental report. In this way, the reader can be made *a virtual participant of the creation, analysis and evaluation of the experimental data*. This means that he/she has to be equipped to reconstruct not only the experimental procedure but the interpretation and authentication of the perceptual data, the significant steps of the argumentation process organising the elaboration of the experimental results, as well as the relationship between the experiments, their exact and non-exact replications, the related control- and counter-experiments, and their impact on the plausibility of the experimental data.

To sum up, we have argued for the idea that (RPE) can be solved with the help of an argumentation theoretical analysis of the experimental report. The model presented in Sections 3.3, 4.2 and 6.3 provides tools for, among others

(a) revealing the *sources* from which the plausibility of the data, hypotheses, and background assumptions used in the experimental process and in the experimental report originate. In this way, the points where plausibility values enter the argumentation process can be identified and their reliability re-evaluated;
(b) representing the *acceptability* of statements as plausibility values. In this way, it can be determined which sources make the statements in experimental reports (background assumptions, predictions, perceptual data, experimental data, hypotheses of theories, conclusions etc.) plausible or implausible, and to what extent;
(c) *determining the plausibility value of conclusions of inferences* with premises that are not true with certainty but only plausible to some extent. Thus, not only the impact of direct sources on the plausibility of the statements can be represented, but the impact of plausible or implausible statements on each other as well. In this way, the strength of support

provided by experimental data as evidence for/against (rival) theories can be determined and compared;

(d) *comparing and summarising* the plausibility value of hypotheses stemming from different sources. Therefore, the dynamism of the change in the plausibility of data and hypotheses in complex chains of related experiments can be accounted for;

(e) *reconstructing the relationship between members of complex chains of related experiments* by identifying the function they fulfil (for instance, original experiment, non-exact replication providing convergent evidence, counter-experiment yielding diverging evidence, etc.). Thus, it is possible to re-evaluate the correctness and efficacy of the plausible argumentation presented in the experimental report by re-evaluating the plausibility of the experimental data as well as the resolutions of the conflicts which emerged between experiments.

Against this background, the following resolution of the rhetorical paradox of experiments presents itself:

The resolution of (RPE):

The reliability of experiments as data sources in cognitive linguistics is directly proportional to the effectiveness of the plausible argumentation process and inversely proportional to the number and weight of the unsolved problems.

This also means that the first standpoint mentioned in Section 2, according to which one has to strive for pure objectivity and stick to the hard facts, is right insofar as the experimental report has to allow access to all relevant details of the experimental process. This, however, does not lead to full objectivity and certainty (although it may reduce uncertainty to a considerable extent) but makes the re-evaluation of the results possible and influences their quality. Further, argumentation is not prohibited in the experimental reports. On the contrary: the re-evaluation of the experiments is an argumentative gesture – at least to the point at which one replicates the experiment in an unvaried or in a modified version. We also agree with the second standpoint mentioned in Section 2, i.e. we agree with the practice of psycholinguistic research insofar as experiments in cognitive linguistics have to be embedded into the context of current linguistic theories. As the case studies illustrate, this is far from being a trivial task.

PART II. THE TREATMENT OF INCONSISTENCIES RELATED TO EXPERIMENTS IN COGNITIVE LINGUISTICS

8. Introduction: The Paradox of Problem-Solving Efficacy (PPSE) in cognitive linguistics

Replications are regarded as inevitable means of securing the reliability of scientific experiments. Despite this, most replications in cognitive metaphor research are – as we have seen in Section 6 – not exact repetitions, but modified or refined versions of another experiment. Sometimes different but closely related experimental designs are used to test the same research hypothesis.

Non-exact replications are conducted in order to rule out possible systematic errors, or to test a more differentiated research hypothesis. Closely related experimental designs, which we called 'methodological variants' in Section 6.3, make use of different techniques but aim at investigating the same research hypothesis. If there is harmony between the results of the original experiment, its non-exact replication(s) and/or methodological variants, then the results are mostly evaluated as *reinforcing* the original research hypothesis. If, however, the revised version of the original experiment or the methodological variant is carried out by adherents of a rival theory, then the experimental data gained are usually found to *conflict* with the original results and prompt the rejection of the research hypothesis. This seems to lead to the Paradox of Problem-Solving Efficacy:

(PPSE) Non-exact explications and methodological variants are
- (a) *effective tools of problem-solving* in cognitive linguistics because by resolving problems they lead to more plausible experimental results; and they are also
- (b) *ineffective tools of problem-solving* because they trigger cumulative contradictions among different replications and methodological variants of an experiment and lead to the emergence of new problems.

Part II of the book intends to propose a possible resolution of The Paradox of Problem-Solving Efficacy. First, as a metascientific background, Section 9 provides a concise overview of the views on inconsistency in the philosophy of science and in theoretical linguistics. Section 10 puts forward a method of inconsistency resolution based on the metascientific model of experimental complexes presented in Section 6.3. Its feasibility will be tested with the help of a case study on replication attempts conducted within cognitive metaphor research. Section 11 offers another possibility insofar as we will apply the tools of statistical meta-analysis to the resolution of inconsistencies. On the basis of our findings, Section 12 will try to generalise the results and provide a solution to (PPSE).

9. Inconsistency in the philosophy of science and in theoretical linguistics

The first step of our line of reasoning will be a survey of the main standpoints pertaining to inconsistencies already discussed, both in the general philosophy of science and the philosophy of linguistics. Thus, this section presents the most influential approaches which have been put forward in relation to the emergence, treatment and function of inconsistencies in the philosophy of science and in theoretical linguistics, as well as the p-model's inconsistency resolution strategies.

9.1. Inconsistency in the philosophy of science

9.1.1. The standard view of inconsistencies in the philosophy of science

The principle of non-contradiction is one of the cornerstones of classical logics since Aristotle:

> "[…] *the same attribute cannot at the same time belong and not belong to the same subject and in the same respect*; we must presuppose, to guard against dialectical objections, any further qualifications which might be added. *This, then, is the most certain of all principles* […]. For it is impossible for any one to believe the same thing to be and not to be […]. For what a man says, he does not necessarily believe; and if it is impossible that contrary attributes should belong at the same time to the same subject (the usual qualifications must be presupposed in this premise too), and if an opinion which contradicts another is contrary to it, obviously it is impossible for the same man at the same time to believe the same thing to be and not to be; for if a man were mistaken on this point he would have contrary opinions at the same time. *It is for this reason that all who are carrying out a demonstration reduce it to this as an ultimate belief; for this is naturally the starting-point even for all the other axioms*." (Aristotle: Metaphysics, Book 4/3; emphasis added)

Modern (Fregean) logic kept the principle of non-contradiction as one of its basic principles. Since the standard view of the analytical philosophy of science interpreted scientific theories as sets of hypotheses that can be represented as axiomatic systems (calculi), it imposed the requirement of strict inconsistency-intolerance on scientific theories:

> "[…] the importance of the requirement of consistency will be appreciated if one realizes that a self-contradictory system is *uninformative*. It is so because any conclusion we please can be derived from it. Thus no statement is singled out, either as incompatible or as derivable, since all are derivable. A consistent system, on the other hand, divides the set of all possible statements into two: those which it contradicts and those with which it is compatible. (Among the latter are the conclusions which can be derived from it.) This is why *consistency is the most general requirement for a system, whether empirical or non-empirical, if it is to be of any use at all*." (Popper 1959: 92; emphasis added)

Accordingly, researchers should immediately give up a theory if it is burdened with an inconsistency. In particular, it is assumed by the standard view of the analytical philosophy of science that when a hypothesis is in conflict with some empirical datum (for example, result of an experiment or observation), then it is always the hypothesis that has to be immediately rejected and the theory should be revised. This precept is based on the assumption that scientific theories either have an empirical basis consisting of a previously secured data set or, at least,

there is consensus within the scientific community about the types and range of the useable data which can be treated as irrevisable facts in the given inquiry.

9.1.2. Break with the standard view in the philosophy of science

The *historical turn* in the philosophy of science brought a radical re-evaluation of the role of inconsistencies in science. According to Kuhn (1962), there is no direct connection between theory-change and inconsistencies. On the contrary: theories often coexist with counter-examples. It is only the emergence of *a series of serious, irresolvable and irritating anomalies* that forces researchers to seek the radical revision of their theory and leads to the elaboration of a new paradigm. Nevertheless, a paradigm is given up not by individual researchers but always by a scientific community. Moreover, paradigm change may happen only if there is already a rival paradigm that seems to be more promising because it can explain some of the most stubborn anomalies and has a huge heuristic potential.

It is, however, often the case that the new theory has several counter-examples from the outset. If the community of researchers thinks that there is hope that these counter-examples may be eliminated at a later stage of the development of the theory, then these inconsistencies are temporarily tolerated. From this picture of science, it follows that there are no methodological rules governing the treatment of *individual inconsistencies*. In Kuhn's view, it is a social, historical and contingent fact whether a given anomaly is tolerated or is regarded as a major disorder of the given theory.

Lakatos (1978) agrees with the idea that counter-examples are neither fatal nor equal and tries to elaborate a methodology that makes it possible to *provide guidelines* for dealing with inconsistencies in science.

A substantial point of his argumentation is that he breaks with the tenet of the standard view according to which it is possible to draw a sharp dividing line between theoretical claims and empirical (observational, experimental, etc.) statements:

> "Propositions can only be derived from other propositions, they cannot be derived from facts: One cannot prove statements from experiences. […] If factual propositions are unprovable then they are fallible. If they are fallible then clashes between theories and factual propositions are not 'falsifications' but merely **inconsistencies**. […] *Thus, we cannot prove theories and we cannot disprove them either*." (Lakatos 1978: 16; cursive emphasis as in the original, bold emphasis added)

Popper, the father of falsificationism argued that the consensus of researchers and previous clarification provide a firm ground for a demarcation between 'base statements' and hypotheses. Thus, still within the boundaries the standard view, he thought that in practical terms falsification could be deemed to be decisive insofar as a single piece of counter-evidence is capable of falsifying a hypothesis. According to Lakatos, in contrast, conflicts between "empirical facts" and theories are *conflicts between theories* – the theory proposed and a theory used to interpret experimental results, observations, etc.:

> "It is not that we propose a theory and Nature may sound *no*; rather, we propose a maze of theories, and Nature may shout *inconsistent*." (Lakatos 1978: 45; emphasis as in the original)

Therefore, since both "empirical data" and "hypotheses of the theory" are fallible theory-burdened statements, it is quite natural that researchers often save their theory from refutation by discrediting the counter-evidence instead. They reveal faults, for example, in the interpretation of the observed "facts" or in the theory of the equipment used. Nevertheless, the question emerges exactly when, under what circumstances this strategy is rational and when it counts as an irrational move. Lakatos (1978) proposes a demarcation criterion according to which an anomaly is only *fatal* for theory *T* if there is another theory *T'* that (a) is not in conflict with this piece of evidence, (b) is capable of explaining all observations/experimental results etc. that *T* could, and (c) makes predictions which are novel facts (that is, which were either forbidden by *T* or to which *T* cannot be applied). In such cases, *T* has to be given up, i.e., it has to be regarded as falsified and has to be replaced by *T*', which we do not require to be free from inconsistencies.

The impact of inconsistencies in scientific theorising is further reduced in the Lakatosian model on the basis of the following considerations:

– It is always possible to save a hypothesis of a theory from falsification if some other part of the theory or the interpretation of the counter-evidence is modified in such a way that the conflict disappears ("negative heuristics"). Nevertheless, not all such moves are equal. Conflict resolutions are called by Lakatos (1978) 'progressive problem-shifts' if as a result of the modification, new predictions are also made that can be corroborated (checked successfully). Conflict resolutions failing to provide new corroborated predictions are dubbed 'degenerative problem-shifts'.
– Usually, there are many theory-versions that count as progressive problem-shifts. This means that in most cases the researcher can raise a wide range of at least partially incompatible possibilities of the modification of the theory. One may give up or refine less important hypotheses or extend the theory by new auxiliary hypotheses in such a way that the modification or extension leads to (of course different) novel predictions. Thus, the question emerges of how to resolve the inconsistency between these rivals.
– Lakatos raises the hypothesis that scientific research is basically guided not by the treatment of anomalies but by a "positive heuristics":

> "The positive heuristic […] saves the scientist from becoming confused by the ocean of anomalies. The positive heuristic sets out a programme which lists a chain of ever more complicated *models* simulating reality: The scientist's attention is riveted on building his models following instructions which are laid down in the positive part of his programme. He ignores the *actual* counterexamples, the available 'data'." (Lakatos 1978: 50; emphasis as in the original)

– Theories do not have to be consistent; they can co-exist with anomalies if their heuristic potential is high and there is no better theory available yet.

Apparently, the Lakatosian model is considerably closer to the practice of scientific research than the tenets of the standard view. Despite this, it is problematic from several points of view:

– As Nickles (2000) notices, Lakatos loosens his criterion that only progressive problem-shifts are rational. To wit, he argues that it may be fruitful to be bound to stick to a degenerative research program because it can revive and become progressive again. One has merely to keep in mind that this is a risky decision. This means, however, that Lakatos's proposal cannot be used as a methodological guideline, because it can be judged only in the long run whether a decision was rational or not.
– Lakatos does not provide means for the treatment of inconsistencies between rival theories that are not modifications of each other but diverge to a considerable extent. In such cases, it may happen that one theory solves an inconsistency included in the other theory but is burdened by others, or both theories are capable of explaining empirical phenomena that cannot be explained with the help of the other. Therefore, it is not clear which of them is better.
– While distancing himself from falsificationism, Lakatos (1978) seems to separate an autonomous and self-controlling "positive heuristics" and a totally reactive "negative heuristics" from each other and subscribe a leading role expressly to the former:

> "Science can grow without any 'refutations' leading the way. […] The problem fever of science is raised by proliferation of rival theories rather than counterexamples or anomalies." (Lakatos 1978: 36)

This would mean that in the progressive phase of their development, theories do not react to their "context" but follow their own route and, consequently, there is no connection between the emergence of rival theories and anomalies or anomalies and the development of theories. That is, while Kuhn (1962) emphasised the irritating effect of the proliferation of resistant anomalies, Lakatos is of the opinion that

> "[w]hich problems scientists working in powerful research programmes rationally choose is determined by the positive heuristic of the programme rather than by psychologically worrying (or technologically urgent) anomalies. The anomalies are listed but shoved aside in the hope that they will turn, in due course, into corroborations of the programme. Only those scientists have to rivet their attention on anomalies who are either engaged in trial-and-error exercises […] or who work in a degenerating phase of a research programme when the positive heuristic ran out of steam." (Lakatos 1978: 52)

In contrast, Larry Laudan puts *problems* – and among them, inconsistencies – at the centre of his model of science and tries to reveal their multifaceted contribution to the development of science in a more balanced and comprehensive way. The central role of problems in science might seem to be a platitude at first glance but it is not:

> "The literature of the methodology of science offers us neither a taxonomy of the types of scientific problems, nor any acceptable method of grading their relative importance. It is noticeably silent about what the criteria are for an adequate solution to a problem. It does not recognize there are degrees of adequacy in problem solution, some solutions being better and richer than others." (Laudan 1977: 13f.)

According to Laudan (1977), there are two basic types of problems: empirical problems and conceptual problems. Both groups can be further subdivided into subgroups.

As regards *empirical problems*, besides solved problems there are also unsolved problems and anomalous problems. A problem can be regarded as *solved* if there is a theory that provides

results that are good approximations of the experimental results or observations at issue. Thus, solutions do not have to be true and exact explanations of facts. Moreover, solutions can be rejected not only by rival theories but also at later stages of research by followers of the given theory. *Unsolved problems* are problems for which no solution has been proposed yet. They are often vague: it is not always clear whether they belong to the given research field and concern real phenomena. Therefore, they do not constitute serious counter-evidence against a theory. *Anomalous problems* cannot be solved within a particular theory while at least one of the latter's rivals is capable of solving it. They are, in Laudan's view, of vital importance in scientific research and belong to the driving forces of science. They motivate adherents of a given theory to turn them into a solved problem. Nevertheless, the presence of anomalous problems does not involve the rejection of a theory, because the experimental results can be faulty, too. Nor does anomaly mean in every case inconsistency of empirical data with the theory. For instance, the inability of a theory to provide an explanation of observations that can be accounted for by rival theories counts as a serious anomaly as well (Laudan 1977: 29).

Empirical problems are of *different weights*. Solved problems always represent a challenge to the rival theories, because they are anomalies for them. Problems related to central concepts or basic hypotheses of a theory are also attributed greater significance. The level of generality of a problem is an important factor, too. Of course, the significance of a problem may also decrease or vanish. A decisive factor in the judgement of the importance of an anomaly is "the competitive state of play between that theory and its competitors" (Laudan 1977: 38), as well as "age and its demonstrated resistance to solution by a particular theory" (Laudan 1977: 39). Therefore, the question is always how big the failure or partial success of a theory in connection with a problem in comparison is with that of its rivals.

The second type of problems are *conceptual problems*. The most acute (but relatively rare) version of *internal conceptual problems* is the *logical inconsistency* (self-contradictory character) of a theory. Laudan (1977: 49) enumerates three strategies that can be applied in such cases: abandoning the rules of inference, localisation of the inconsistency, and refusing to accept the theory. Nevertheless, he does not provide a detailed analysis of the working mechanism and applicability of these methods. *Problems arising from conceptual ambiguity or circularity* are more frequent internal conceptual problems. Ambiguity is, in Laudan's view, not completely eliminable from science, but, of course, while mildly vague definitions may be useful in certain situations, acute vagueness makes a theory useless. *External conceptual problems*, that is, conflicts or tension with other, highly appreciated theories are more important from a historical point of view than internal ones. The most acute (but not the most frequent) form of external conceptual problem is *inconsistency*. As Laudan (1977: 56) emphasises, inconsistency between two theories makes both of them weaker and its resolution requires the thorough examination and comparison of all possibilities: giving up the first theory, rejection the second theory, or giving up both. In all cases, a new theory is also needed which is capable of superseding the rejected one(s). A second subtype of external conceptual problems is when a new theory is not inconsistent with a related earlier theory but makes it *implausible*. Thirdly, it may happen that a theory is elaborated to reinforce another theory but is not capable of fulfilling this task and is *merely compatible* with it instead of providing support for it (Laudan 1977: 53).

The weight of conceptual problems is determined by the *tension* caused by the conflict between the two theories, the *acceptability* of the conflicting theory, the overlap between the conflicting theories (that is, whether they are complementary or competitors), and the age and history of the given problem.

Laudan is of the opinion that *problem-solving efficiency* is the most important factor in the evaluation and comparison of theories. However, as he indicates, this task is very complicated because the solution of a problem may generate other (or even more serious) problems (Laudan 1977: 67).

Laudan clearly realised the need for the elaboration of a comprehensive typology of problems, and contributed to the break with the tenets of the standard view with very deep and important revelations. Although he has paved the way for the investigation of inconsistencies in the context of a more general account of science as a problem-solving activity, he offers rather starting points as a detailed model. That is, he developed a model of science at a high level of abstraction, which is relatively close to the practice of scientific research due to the numerous historical examples but does not fulfil its aim of providing methodological guidelines. This particularly concerns inconsistencies, because he seems to underestimate the role of inconsistencies in science.

9.1.3. New approaches to inconsistency in the philosophy of science

The insights of Lakatos, Laudan and others ushered in a radically new era in the philosophy of science. According to Nickles (2002: 2), a key point in the break with the standard view is that *local problem-solving* has come into sharp focus instead of theories interpreted as hypothesis systems. Scientific theorising is now understood as a non-monotonic, self-corrective *process*. Against this background, inconsistencies are regarded no longer as fatal failures or individual faults but as everyday concomitants of inquiry. From this it follows that inconsistency is seen as a *more significant* but *less serious* foundational problem:

> "[…] we are left with the task of better understanding how inconsistency and neighbouring kinds of incompatibility are tamed in scientific practice and the corresponding task of better modelling idealized practice in the form of inconsistency-tolerant logics and methodologies." (Nickles 2002: 2)

The radicalness of this change of view is clear. First, the relevance of logic in relation to inconsistencies in science is still acknowledged but new, *non-classical logics* have been elaborated which make it possible to tolerate (certain) inconsistencies. Second, the starting point of the new treatment of inconsistencies should be the *practice* of scientific research. Nevertheless, it is by no means an easy task because

> "[…] many scientists move back and forth between (or among) the various stances on these issues, depending upon their problem situation and the audience. In other words, their stance is not consistent even at the metalevel, insofar as 'consistent' implies fixed." (Nickles 2002: 2f.)

That is, the elaboration of useable metascientific tools for the treatment of inconsistencies is an immensely difficult task because of the missing or faulty self-reflection of scientists as well as the presence of the remnants of the standard view. Since the standard view interpreted theories as deductively arranged hypothesis systems, it was unavoidable that all inconsistencies

were deemed fatal and that consistency had to be regarded as one of the most important criteria that cannot be violated in order to satisfy some other criterion against scientific theories. As Nickles (2002) shows, according to recent approaches to the philosophy of science, this concept of theories must be given up. Theories should be viewed and explained *dynamically* and in an *evolutionary sense* and not statically. This means that one should concentrate on the whole process and not only on the product of the process of scientific theorising. The practical side, the skills, practices, the application and modification of *models* and *problem-solving activities* have to be pushed to the foreground, and the *cognitive aspects* of scientific theorising should be paid considerable attention:

> "It seems fair to describe the shift from theories to models as a shift from an absolutist, 'God's eye' account of science (or its ultimate goal) to a more pragmatic, perspectival, human-centered account, and, correspondingly, a shift from a high logical account to a more informal and rhetorical account." (Nickles 2002: 19)

This should not mean that striving for consistency would cease and every inconsistency should be tolerated. Rather, its weight and role in scientific theorising should be rethought:

> "Although consistency remains a strong desideratum, and justifiably so, consistency increasingly becomes a regulative ideal for the 'final' products of research rather than something to be imposed rigidly from the start on the process of research." (Nickles 2002: 20)

> "Today, it is generally recognised that almost *all* scientific theories at some point in their development were *either inconsistent or incompatible* with other accepted findings (empirical or theoretical). A growing number of scholars moreover recognises that inconsistencies *need not be disastrous for good reasoning*." (Meheus 2002: VII; emphasis added)

Basically, in last decade the principle of non-contradiction in relation to scientific inquiry has been re-evaluated in the following respects:

- Contradictions are not monolithic, rather, there are *different kinds* of contradiction.
- Inconsistencies may differ with respect to their *structure*.
- *Not all* contradictions are *harmful*.
- Different kinds of contradiction may have different *functions* in scientific theorising.
- New systems of logic have been elaborated which allow for certain kinds of contradiction without being exposed to logical chaos. Such systems are called *paraconsistent logics*.

This radical change of view, however, leads to a series of new questions (cf. Nickles 2002: 19f.):

- What does consistency/inconsistency mean if we interpret theories not as static systems of hypotheses but as dynamic, self-correcting processes?
- When is it fruitful to insist on consistency and when is it heuristically rewarding to tolerate a contradiction?
- How many types of inconsistency are there?

- In which cases (if at all) can be the result (final state) of the research process be inconsistent?
- What is the role of inconsistencies in the regulation of the process of scientific theorising interpreted as problem-solving process?
- How can be inconsistency weighed up against other violations of other constraints?
- What are the techniques of temporal inconsistency-toleration?
- What are the techniques of permanent inconsistency-toleration?

Although several attempts have been made to solve them, they are still perceived as open questions. Despite this, Nickles thinks that there are some arguments indicating that the efforts at the elaboration of a new methodology will be rewarded with a more efficient and useable scientific practice. First, several other unrealistic elements of the standard view, such as the requirement for the totally "objective", theory-free, intersubjective and certainly true evidential basis of scientific inquiry or the representation of scientific theories as deductive systems has also gradually loosened and been given up. Second, there is an evolutionary argument: if inconsistencies are tolerated at least during the research process instead of being eliminated at once, then a wider spectrum of possible theory-variants are developed and can be compared:

> "The more kinds of incompatibility represented in the initial population, the more severe the competition, and the higher the probability of arriving at a good solution sooner rather than later. […] By contrast, in standard symbolic logic once you have one inconsistency you have them all, and it makes little sense to speak of diversity. […] This point applies not only to research communities but also to each individual problem solver, at least at the subconscious level, where our brains are engaged in a kind of evolutionary exploration of the problem space." (Nickles 2002: 28)

Third, if one does not compel oneself to consistency from the beginning and at every step of the research process but also enables inconsistencies and lower grades of plausibility, then consistency can be achieved in the long run, as a result of a comprehensive, complex process that has examined and compared a rich variety of possible alternatives:

> "When won under risk of inconsistency, consistency of results becomes a genuine achievement and is itself a mark of robustness. […] Accordingly, methodology should treat this sort of consistency as an aim, as a desired feature of the refined conclusions of research, rather than as a prior condition of research itself." (Nickles 2002: 22)

These considerations create a need for the elaboration of *inconsistency-tolerant logics* and methodological rules pertaining to their application in scientific theorising.

9.2. Inconsistency in theoretical linguistics

9.2.1. The standard view of inconsistencies in linguistics

Hjelmslev's 'empirical principle' clearly shows that for structuralists, the principle of non-contradiction and the related tenets of the standard view of the analytical philosophy of science

were the most important requirements formulated in relation to linguistic theories, which cannot be counterbalanced by the fulfilment of any other criteria:

> "The description *shall be free of contradiction (self-consistent)*, exhaustive, and as simple as possible. The requirement of freedom from contradiction *takes precedence* over the requirement of exhaustive description. The requirement of exhaustive description *takes precedence* over the requirement of simplicity." (Hjelmslev 1969: 11; emphasis added)

Chomsky and the generativists accepted this view for decades by stipulating "explanatory adequacy" as the first requirement theories of grammar should fulfil. Corpus linguists adopted a similar stance insofar as they regarded Popperian falsificationism as a basic methodological rule. According to this tenet, a single counter-example is sufficient to falsify a hypothesis. This criterion, however, was felt to be too strict in practice. Therefore, both generativists and corpus linguists tried to replace "strong falsificationism" with weaker norms.

9.2.2. Weak falsificationism in corpus linguistics

Several corpus linguists shared the view that rules of language have to be interpreted not as strict prescriptions that do not allow exceptions but rather as statistical tendencies. Thus, according to *weak falsificationism*, rare occurrences are not sufficient to falsify a hypothesis. As Penke & Rosenbach (2004: 483) remark, however, it is not clear how to distinguish exactly such rare occurrences from counter-examples that have to be regarded to be falsifying evidence against a hypothesis. That is, the problem is the question of how to interpret the idea of "weak falsification" in quantitative terms. For example, it is not clear how many counter-examples refute a hypothesised linguistic rule and how many exceptions can be tolerated.

9.2.3. Linguistics in a "Galilean style"

An important reason why strong falsificationism turned out to be a too strict criterion for theories in generative linguistics was that the proposed models of grammar *overgeneralised*. In phonology, for example, attempts made at the formal restriction of possible types of alternations obstinately failed:

> "[…] the alternations permitted by every formal model unfortunately also include alternations that are both unattested and thought to be unlikely. *There were always counterexamples*." (Archangeli 1997: 25; emphasis added)

Generative grammarians tried to react to this phenomenon by proposing a series of new, more specific rules and constraints. This led, of course, to an increased number of hypotheses. According to Archangeli's (1997) diagnosis, however, a second problem in the generative phonology of the 70s and 80s was the overcomplication of the proposed grammars. For example, the characterisation of alternations or the set of constraints related to different levels of linguistic representation seemed to be hopelessly and unrealistically complex.

As for syntax, the situation was similar in Archangeli's view. A great many empty terminal nodes and conditions or principles which should rule out ungrammatical structures and were supposed to be exceptionless were introduced. This, however, made syntactic theories too

complicated. Chomsky reacted to this situation with a gradual re-evaluation of the tolerability of counter-examples:

> "It is not necessary to achieve descriptive adequacy before raising questions of explanatory adequacy. On the contrary, the crucial questions, the questions that have the greatest bearing on our concept of language and on descriptive practice as well, are almost always those involving explanatory adequacy with respect to particular aspects of language structure." (Chomsky 1965: 36).

At the end of this re-evaluation process we find Chomsky's radical proposal for pursuing linguistics in a "Galilean style":

> "Apparent counterexamples and unexplained phenomena should be carefully noted, but it is often rational to put them aside pending further study when principles of a certain degree of explanatory power are at stake. How to make such judgements is not at all obvious: there are no clear criteria for doing so. […] But this contingency of rational inquiry should be no more disturbing in the study of language than it is in the natural sciences." (Chomsky 1980: 2)

This strategy resembles Lakatos' model of positive and negative heuristics (cf. Section 9.1.2), since Chomsky seems to subscribe a greater importance to the development of a model of grammar than to the avoidance of inconsistencies. The idea of *the temporary ignorance of counter-examples* retains the view that exceptions are *failures* – although they are no longer regarded as fatal but only as hindrances or difficulties, which have to be overcome in future. According to Chomsky, in most cases inconsistencies may be put aside in the hope that later developments of the theory will solve them. Nevertheless, as the quotation witnesses, counter-examples are deemed by him to be foreign bodies in the actual phase of theory formation, and have to be practically ignored for a shorter or longer time. According to his view, this neglect is all the more justified, as exceptions are *disturbing factors*, which may divert the process of linguistic theorising away from its correct direction. Furthermore, as the quotation makes explicit, where the limits of inconsistency-tolerance lie has not been made clear. That is, Chomsky does not clarify under what circumstances the application of this strategy is legitimate and when not.

9.2.4. Inconsistencies as stimulators of further research in linguistics

In harmony with the views presented in Section 9.1.3, some authors have indicated that inconsistencies should not be evaluated negatively in linguistics, either. Rather, it should be acknowledged that they *play a vital role in the development of theories*. In this vein, Kepser & Reis (2005: 3) emphasise that contradictions resulting from the diversity of data may be *fruitful* because striving for their resolution plays a central role in scientific progress:

> "Evidence involving different domains of data will shed different, but altogether more, light on the issues under investigation, be it that the various findings support each other, help with correct interpretation, or by contradicting each other, lead to factors of influence so far overlooked." (Kepser & Reis 2005)

Similarly, it is illuminating to compare Chomsky's cited formulation on the one hand and Penke and Rosenbach's interpretation of its essence on the other:

> "According to Chomsky it is legitimate to ignore certain data to gain a deeper understanding of the principles governing the system under investigation. [...] In all these cases, *the apparent counter-evidence* was not taken to refute a theory, but *stimulated further research that resulted in the discovery of principles so far unknown, thus enhancing our understanding of the phenomena under study*." (Penke & Rosenbach 2004: 484; emphasis added)

On this view, contradictions are, at least in certain cases, no mistakes at the outset. Rather, they have to be considered one of the major driving forces of the development of linguistic theories.

9.3. Problem-solving strategies of the p-model

As we have seen in Section 4.2.3, the p-model by Kertész & Rákosi interprets inconsistencies not as purely formal issues but as conflicts resulting from opposing evaluations of the plausibility of statements. The resolution of *p-inconsistencies* cannot be reduced to the comparison of the two conflicting statements in isolation; one has to re-evaluate the whole p-context. As already mentioned in Section 4.2.5, the decision as to whether the argumentation process can be terminated and the resolution of the p-problems is achieved is not absolute and not incontrovertible. The reason for this is firstly, that because of *practical limits*, the cyclic re-evaluation cannot be complete, cannot take into consideration every piece of information and cannot examine every possibility, but has to remain partial. The second reason is that in most cases, attempts at the solution of the initial problems lead to the emergence of new problems which should be solved – and so on ad infinitum. Third, the rival solutions obtained as results of the argumentation cycles conducted are partial, too, since they have been elaborated with the help of diverse heuristics. Moreover, the comparison of the rival solutions cannot be reduced to the mechanical comparison of the plausibility of their hypotheses but has to rely on a series of criteria. One must examine which solution is, *as a whole*, the least p-problematic p-context, which solution contains the highest number of hypotheses with a high plausibility value, which solution is the most comprehensive etc. However, this is usually a difficult and complex task because there may be conflicts among the evaluations obtained: one solution can be optimal in respect to one criterion but not as satisfactory with others.

Therefore, it is of vital importance to find *problem-solving strategies* which may lead to effective and reliable decisions. Such heuristics make it possible to elaborate and compare a fair number of p-context versions and arrive at a well-founded resolution of the starting p-problem even though the fulfilment of these tasks can be only partial.

An important subgroup of heuristics consists of *strategies for the treatment of p-inconsistency*. Basically, one can follow three strategies:

– **The Contrastive Strategy**. The essence of this strategy is that it treats the p-context versions containing the contradictory statements as *rival alternatives*, compares them and strives for a decision between them.
– **The Exclusive Strategy**. This strategy is the continuation of the Contrastive Strategy in cases when a decision has been reached between the rival p-context versions elaborated. It fulfils a kind of control function insofar as it examines whether the p-context version

chosen can provide an explanation for all phenomena that could be explained by the rejected p-context version. This is important because the explanatory power of the resolution should be as high as possible; therefore, information loss resulting from the rejection of p-context versions should be avoided or at least minimised.

– **The Combinative Strategy**. It may be the case that one wants to keep both rival p-context versions because they illuminate some phenomenon from different points of view which are equally important and cannot be given up. Therefore, the two p-context versions are no longer deemed to be rivals but co-existing alternatives which have to be maintained *simultaneously*. The task is to elaborate both p-context versions, and make them as comprehensive and as free from p-problems as possible. Nevertheless, the separation of the two p-context versions has to be *systematic* and *well-motivated*.

The successful application of the Contrastive and Exclusive Strategies clearly eliminates the given inconsistency in such a way that it remains within the boundaries of classical logic. The Combinative Strategy, in contrast, is different: we have to transgress the boundaries of classical two-valued logic. For details, see Kertész & Rákosi (2013).

10. Inconsistency resolution and cyclic re-evaluation in relation to experiments in cognitive linguistics

In this section, our aim is to describe the emergence and treatment of inconsistencies in relation to experiments on metaphor processing by integrating the metascientific model of experimental complexes as presented in Section 6.3 and the p-model's strategies of inconsistency resolution as briefly summarised in Section 9.3. As a case study, we will analyse three experiments by Keysar, Shen, Glucksberg & Horton, Glucksberg, McGlone & Manfredi and Bowdle & Gentner and their non-exact replications.

10.1. Case study 4, Part 1: Three experiments on metaphor processing and their replications

First, we present a first concise description of the original experiments and the replication attempts.

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

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

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

1. *implicit-mapping scenario*: contains conventionalised expressions supposed to belong to the same conceptual metaphor as the target expression (which was always the final sentence of the scenario);[65]
2. *no-mapping scenario*: conventional instantiations of the supposed mapping are replaced by expressions not related to the given mapping;[66]
3. *explicit-mapping scenario*: in addition to the implicit-mapping scenario, the supposed mapping has been made explicit by being mentioned at the beginning of the text;[67]
4. *literal-meaning scenario*: renders the target expression as literal.[68]

---

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

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

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

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

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

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

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

B) Replication: Thibodeau & Durgin (2008)

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

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

Experiment 3: In a reading times experiment, there were 3 types of scenarios. In the related metaphor scenarios, the target sentence contained a novel metaphor instantiating the same metaphor family as the conventional metaphors in the previous text. In the unrelated metaphor scenarios, the target sentence and the previous text made use of different metaphor families. Non-metaphor scenarios used literal sentences. The authors found that in the related metaphor scenarios, the final sentence read significantly faster than the final sentences of the unrelated

scenarios, or in the non-metaphor scenarios. This also means that the experiments resulted in a shift in the judgement concerning what data should be regarded as relevant: instead of novelty/conventionality, the key factor seemed to be matchedness/unmatchedness.

C) Commentary

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

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

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

The authors intended to provide empirical evidence for the claim that metaphors are, in harmony with the Interactive Property Attribution Model, nonreversible. The stimulus material consisted of 24 metaphors, their corresponding similes and 12 literal similarity statements, each of them in original-order, in noun-reversed and noun-phrase reversed versions.[69] Participants had to evaluate the meaningfulness of the sentences on a 0-7 scale,[70] and, in the case of ratings 1-7, they were asked to write a paraphrase of the sentence as well. The paraphrases were analysed by two independent judges. The authors found that both reversed metaphors and metaphoric comparisons obtained significantly lower meaningfulness ratings than their original counterparts, while with literal comparisons, there was no such difference. Only a few reversed metaphoric statements were equivalent in meaning with the original-order statement; most reversed metaphoric statements were explicitly or implicitly re-reversed, and some were interpreted with new grounds.

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

The first modification to Glucksberg et al's (1997) first experiment pertains to the stimulus material: the set of the target metaphors and similes was extended from 24 to 52, and literal similes were omitted. The authors also modified the research hypothesis as follows: (a) if the traditional comparison theory of metaphors holds, then metaphors are converted into similes and interpreted as comparisons; thus, reversing targets/topics and bases/vehicles should decrease the comprehensibility of metaphors and similes to a slight but equal degree; (b) if Glucksberg's IPAM is correct, then non-literal similes are interpreted, similarly to metaphors, as category statements; thus, both metaphors and similes should be irreversible; (c) if the authors' "distinct statements" view holds, then metaphors function like category claims and similes like similarity claims; thus, reversal should affect metaphors more strongly than similes. The analysis of the paraphrases was conducted in two steps. First, a judge examined the original order items and identified the most frequent interpretations. As the second step, the reversed order paraphrases were classified by two further judges in such a way that they compared the

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

[70] 0 = makes no sense; 7 = makes perfect sense.

reversals to the most frequent original versions, without knowing whether they were presented as metaphors or similes. In contrast to Glucksberg et al., who found that both metaphors and (metaphorical) similes received significantly lower values when reversed, Chiappe, Kennedy and Smykowski came to the conclusion that reversion affected metaphors to a greater extent than similes. The results of the paraphrase analyses were considerably different from the earlier findings, too. Namely, reversed similes were accepted to a greater extent than metaphors, and most reversed items (metaphors and similes alike) were equivalent in meaning to their original counterparts. Further, re-reversal was more frequently applied for metaphors than for similes.

C) Second replication: Campbell & Katz (2006)

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

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

D) Commentary

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

10.1.3. Bowdle & Gentner (2005) and its replications

A) The original experiment: Bowdle & Gentner (2005)

Experiment 1: Participants had to indicate on a 10-point scale whether a certain idiom sounds more natural or sensible in metaphor form or in simile form. On the basis of pre-tests, the stimulus material consisted of 64 items: 32 figurative statements in both the comparison (simile) form and the categorization (metaphor) form,[71] 16 literal comparison statements[72] and 16 literal categorization statements.[73] Half of the figuratives were conventional, the other half were novel; similarly, the figuratives were either abstract or concrete. According to Gentner's career of metaphor hypothesis, novel metaphors are processed as comparisons, while conventionality results in a shift to another mode of processing, namely, categorisation. The experimental data were found to be in harmony with the predictions of the career of metaphor hypothesis, as conventional figurative statements were more acceptable in categorization form than novel figuratives. No main effect of concreteness was found, but there was an unpredicted interaction between concreteness and conventionality.

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

Experiment 3: Experiments 1 and 2 do not touch upon the claim of Gentner's Career of Metaphor Hypothesis that the shift in the processing mode of metaphors occurs gradually, as a by-product of the repetitions of the comparison process. That is, during the repeated derivation or activation of the same abstract, domain-general meaning of the base/vehicle term, this meaning becomes lexicalised and added as a secondary sense to the base/vehicle term. To test this part of Gentner's theory, the authors developed a two-stage experimental design. In the first, study stage, participants saw pairs of novel similes using the same base/vehicle term and they had to fill in a target/topic term in a third example of the same structure.[74] The authors' hypothesis

---

71 For example: *Friendship is like a wine* vs. *Friendship is a wine*.

72 For example: *An encyclopedia is like a dictionary*.

73 For example: *Pepper is a spice*.

74 For example:
An acrobat is like a butterfly.
A figure skater is like a butterfly.

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

B) Replication: Jones & Estes (2006)

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

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

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

_____ is like a butterfly.

C) Commentary

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

10.2. Strategies of inconsistency resolution related to experimental complexes

The definition of experimental complexes in Section 6.3 does not exclude cases in which within an experimental complex, two chains of non-exact replications (or non-exact replications and counter-experiments) lead to conflicting results. A conflict with results of experiments belonging to some other experimental complex may occur, too. These contradictions cannot be resolved simply by a mechanical comparison of the plausibility value of the results of the last member of the chains of experiments. Most frequently, it is not the current state of the cyclic process of re-evaluation that is decisive but *the assessment of future prospects*.

Therefore, the process of inconsistency resolution in cognitive linguistics involves the following three stages:

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

   The **first strategy** consists of a separate continuation of the chains of experiments by conducting further non-exact replications, counter- or control experiments, comparing the results and taking a decision. An analogue of this method was called the "*Contrastive Strategy*" in Section 9.3. There are three basic situations:

   – If the elaboration of further non-exact replications of one of the chains terminates and leads to a limit of the experimental complex in the sense of (LEC),[75] while the other chain comes to a dead-end, then the conflict can be resolved in such a way that the limit is kept, while the rival chain is rejected. Clearly, the elaboration of the first chain of experiments was an effective problem-solving process, while the second one was ineffective.

[75] Cf. Section 6.3.

– If no limit can be achieved by continuing all chains, then the experimental complex is not capable of reaching a limit and the problem-solving process is ineffective.
– It may also occur that both chains of experiments evolving from the same original experiment lead to a limit. In such cases, it would not be reasonable to give up either of them. Thus, this inconsistency has to be (at least temporarily, in the given informational state) tolerated by the application of the second strategy.

A **second strategy** is based on the elaboration and conduct of further experiments involving *a refinement of the research hypothesis and experimental design in such a way that all factors found relevant so far are taken into consideration*. The analogue of this method was called the "*Combinative Strategy*" in Section 9.3. This method might make it possible to resolve contradictions between experiments conducted by researchers committed to rival approaches by *integrating* their results with the help of paraconsistent logic.

In the next section, we will apply this model to the experiments briefly presented in Section 10.1.

10.3. Case study 4, Part 2: Reconstruction and re-evaluation of the problem-solving process

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

A) The structure of the experimental complex

The experimental complex evolving from Experiment 1 in Keysar et al. (2000) involves one exact and three non-exact replications, and a control experiment. See Figure 10.[76]

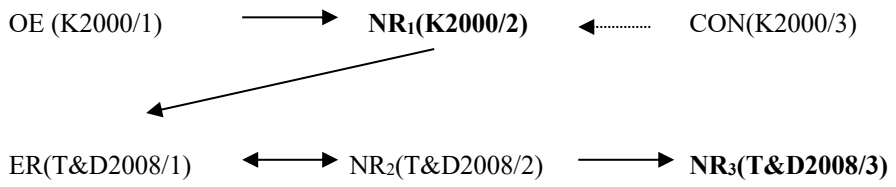

Figure 10. The structure of the experimental complex evolving from Keysar et al. (2000)

Two chains of experiments can be identified:
– NR_1 is an improved version of the original experiment, while CON is a related control experiment;
– ER is an exact replication of NR_1, while NR_2 and NR_3 are non-exact replications of ER, leading to a conflicting result.

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

There are two limit-candidates: NR_1 (Section 10.1.1A), and NR_3 (Section 10.1.1B). Thus, we have to reconstruct and re-evaluate two chains of experiments.

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

OE (cf. Section 10.1.1A):
The first task is the identification of the problematic points of the original experiment:[77]

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

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

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

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

NR_1 (cf. Section 10.1.1A): Experiment 2 in Keysar et al. is a progressive non-exact replication because it deals with Problem 3 insofar as it extends the stimulus material with novel metaphors. Nevertheless, it leaves Problems 1, 2 and 4 open, and leads to the emergence of some new problems, too:

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

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

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

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

[77] See also Section 3.2.6.

CON (cf. Section 10.1.1A): Experiment 3 is a control experiment aimed at providing a solution to Problem 8, whose efficiency, however, can be questioned:

Problem 9: At least in the excerpts presented in the experimental report, the target words were semantically clearly less related to the text of the scenario than other expressions of the target sentence. For example, in the case of the "ideas are people" scenario, the target word was "weaning", while the target sentence also contained the word "child", which was semantically related to "fertile", "giving birth".

Table 4 summarises the current state of the re-evaluation process.[78]

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

Table 4. Overview of the re-evaluation of the limit-candidate by Keysar et al. (2000)

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

ER (cf. Section 10.1.1B): The exact replication of NR_1 leads to a similar pattern of results. Of course, Problems 1-8 emerge here, too.

NR_2 (cf. Section 10.1.1B): Experiment 2 by Thibodeau and Durgin can be considered a progressive non-exact repetition because it provides a solution to Problems 1-5, and 7. Despite this, the results seem to be burdened by the following systematic errors:

Problem 10: Problem 8 has become even more severe than it was with Keysar et al.'s (2000) experiments.[79] That is, there were semantically related words in the target sentences and the texts of the scenarios with the novel metaphor, the conventional metaphor and the literal target scenarios, while this was not the case with the non-metaphoric scenarios.[80]

Problem 11: The filler scenarios were chosen on the basis of other considerations than was the case with the original experiment. Specifically, Keysar et al's main motivation was to make sure that "participants would not anticipate or notice a particular

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

[79] For a more detailed analysis, see Section 3.2.

[80] For example:
IDEAS ARE FOOD
Target sentence: *Otherwise, they give him **indigestion**.*
Novel: *David has a hard time **ingesting** new ideas. He has to **gnaw** on them for days.*
Conventional: *David has a hard time **swallowing** new ideas. He has to **stew** them over for days.*
Non-metaphor: *David takes a while to fully understand new ideas. He has to think about them for days.*
Literal-reading: *David has weak **stomach**. He has to take his time when **eating meals**.*

pattern" (Keysar et al. 2000: 583), and in this spirit, their fillers contained neither metaphorical final sentences nor metaphors belonging to the same conceptual domains. With the new version by Thibodeau & Durgin, however, 2 of every 3 filler scenarios did contain metaphorical expressions; moreover, the fillers were intended to "avoid reading strategies that would cause people to skim over metaphors" (Thibodeau & Durgin 2008: 523). Thus, 4 of 10 questions following the fillers asked explicitly about metaphors. Therefore, participants might have discovered relatively easily that the experiment focused on the use of metaphorical expressions.

Problem 12: As the authors diagnosed, the similar reading speed of the target sentences in the novel and conventional scenarios might be due to the circumstance that participants expected a metaphorical sentence after a text which also contained metaphors – independently of whether or not these metaphors belong to the same metaphorical mapping.

NR_3 (cf. Section 10.1.1B): Experiment 3 by Thibodeau and Durgin aimed to solve Problem 12 and make it possible to reject an alternative explanation of the results. The practice and filler tasks were modified, too, so that they no longer asked about metaphors explicitly. Thus, NE_3 is a progressive non-exact replication. It is, however, not a limit, because a variant of Problem 10 emerges again:

Problem 13: It cannot be ruled out that the significantly shorter reading time of the consistent metaphorical scenarios was the result of semantic (lexical) priming.[81]

Table 5 shows the reconstruction of this chain of experiments.

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

Table 5. Overview of the re-evaluation of the limit-candidate by Thibodeau & Durgin (2008)

D) Comparison of the problem-solving processes

On the basis of our reconstruction, the decision of Thibodeau and Durgin regarding the rejection of NR_1 despite the successful exact replication, has become completely reasonable. Namely, both NR_1 and ER are burdened with problems which could not be eliminated. Therefore, NR_1 cannot be regarded as the limit of this experimental complex. Our analyses motivate

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

a similar verdict with the limit-candidate NR_3. From this it follows that the conflict between Keysar et al's and Thibodeau and Durgin's results cannot be resolved on the basis of the information at our disposal at this point of the process of re-evaluation. Although Thibodeau and Durgin's results are more plausible, it would be erroneous to apply the first (contrastive) strategy and terminate the problem-solving process at this point. Moreover, there are experiments by Gentner and her colleagues (see, above all, Gentner & Boronat 1992) whose results are in harmony with Keysar et al's findings.

E) Determination of the direction of the continuation of the cyclic process of re-evaluation

The next question is, of course, how the problem-solving process should proceed. Since there were two factors (conventionality, matchedness) which seemed to be relevant, the first choice could be the application of the second (combinative) strategy[82] insofar as an experimental design should be elaborated that takes both of them into account and helps us to compare their contribution to metaphor processing. The persistent emergence of semantic priming effects, however, seriously questions the viability of this endeavour.

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

A) The structure of the experimental complex

This experimental complex consists of an original experiment, two non-exact replications, and a counter-experiment. See Figure 11.

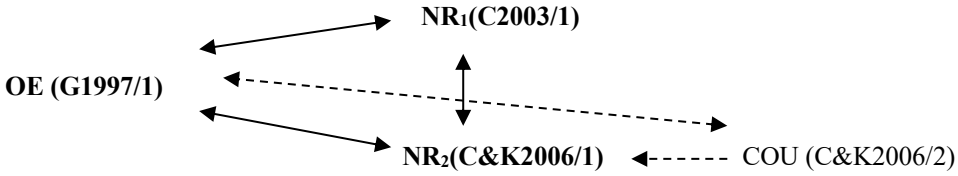

Figure 11. The structure of the experimental complex evolving from Glucksberg et al. (1997)

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

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

OE (cf. Section 10.1.2A): The first task to be undertaken is the identification of the problematic points of the original experiment.

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

[82] See Section 10.2.

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

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

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

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

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

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

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

NR_1 (cf. Section 10.1.2B): The progressivity of this non-exact replication is due to three factors: the extension of the set of metaphors in the stimulus material, suggesting a more elaborated research hypothesis (and providing a partial solution to Problem 7), as well as the solution of Problem 4 by applying independent scorers blind to the aim and structure of the experiment. Problems 1, 2, 3, and 5, in contrast, remained open, and also new problems emerged:

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

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

Problem 10: NR_1 seems to make use of rather novel metaphors, while OE contained both conventional and novel metaphors. Thus, the role of conventionality is not reflected upon.

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

NR_2 (cf. Section 10.1.2C): This non-exact replication is clearly progressive, because at several points the experimental design was re-thought and modifications were made, such as the addi-

tion of the contextually embedded versions and the refinement of the coding system. Thus, NR_2 provides at least a partial solution to Problems 2, 3, 4, 5, 8, 9 and 10. The author's attempt to resolve Problems 2 and 3, however, has also led to a new problem:

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

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

Problem 12: Since participants had to press a button after reading each word, this might have distorted their normal reading habits.
Problem 13: The negative outcome of the experiment (no reliable differences were found) motivates a control experiment in order to check whether this method is sensitive enough to detect relevant differences.

E) Comparison of the problem-solving processes

Table 6 shows the emergence and solution of problems in this experimental complex.

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

Table 6. Overview of the re-evaluation of the experimental complex evolving from Glucksberg et al. (1997)

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

F) Determination of the direction of the continuation of the cyclic process of re-evaluation

Since NR_2 proved to be the most refined version of the original experiment, the most promising decision might be to improve it further, i.e., to use the first (contrastive) strategy.[83] The follow-

[83] See Section 10.2.

ing points should receive special emphasis during the elaboration of a new non-exact replication:

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

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

A) The structure of the experimental complex

This experimental complex involves an original experiment and a related control experiment, as well as two non-exact replications of the original experiment and a non-exact replication of the control experiment. Experiment 3 by Jones and Estes is not included because it belongs to another experimental complex. That is, it is neither the non-exact replication of NR_2 or CON_2, nor a counter- or control experiment to any of the experiments. See Figure 12.

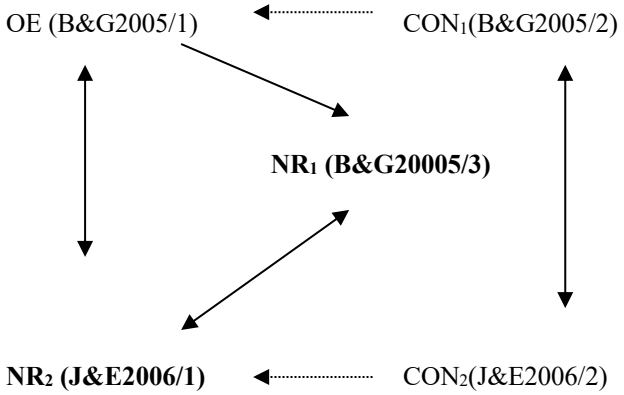

Figure 12. The structure of the experimental complex evolving from Bowdle & Gentner (2005)

In this case, there are two limit-candidates: NR_1 and NR_2.

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

OE (cf. Section 10.1.3A): Regarding the *identification of the problematic points of the original experiment*, the following list can be compiled:

Problem 1: The number of participants was very low, since there were only 16 subjects.
Problem 2: In the pre-tests, a small group of subjects had to evaluate the conventionality and abstractness of a huge number of items, i.e. 100 figurative statements.

Problem 3: The high number of items and the invariance in the task might have led to unnatural linguistic behaviour and the use of conscious strategies.

Problem 4: Neither the stimulus material nor the results of the pre-tests can be found in the experimental report.

Problem 5: Grammatical form preferences do not necessarily mirror processing differences. It might be the case that conventional figuratives are preferred as metaphors, due to the higher frequency and familiarity of these forms.

Problem 6: There was an unpredicted interaction between conventionality and concreteness. Thus, the research hypothesis and the predictions seem to be incomplete because they leave the role of concreteness/abstractness unclarified.

CON_1 (cf. Section 10.1.3A): Experiment 2 in Bowdle & Gentner (2005) is a control experiment. Although it provides a solution to Problem 5 by the application of an online method, it also raises new problems:

Problem 7: Since participants knew they had to provide an interpretation, response times might have been not pure comprehension times but might have been lengthened if a participant had already tried to formulate an interpretation. Therefore, the ease of formulation of an interpretation might have influenced the comprehension times.

Problem 8: The role of aptness, as raised, for example, in Chiappe, Kennedy & Smykowski (2003) and Jones & Estes (2005), was only investigated in a (very thorough) post-hoc test.

NR_1 (cf. Section 10.1.3A): Experiment 3 in Bowdle & Gentner (2005) is a non-exact replication of Experiment 1. Its progressivity is due to the involvement of further elements of the theory into the tested hypothesis and experimental design. Problem 1 was solved by recruiting a higher number of participants, but Problems 2-5 emerge in this case, again. There were two further problems, as well:

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

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

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

NR_2 (cf. Section 10.1.3B): Experiment 1 by Jones & Estes (2006) is a progressive non-exact replication of OE due to the addition of a new, potentially relevant factor (aptness) to the tested hypothesis, as well as the solution of Problems 1, 4 and 8, and a partial solution to Problem 2. Two new problems have arisen:

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

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

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

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

CON_2 (cf. Section 10.1.3B): Experiment 2 was a control experiment for NE_2, and, at the same time, a non-exact replication of CON_1 by Bowdle & Gentner (2005). Its progressivity is mainly due to the same factors as was the case with NR_2. Nonetheless, it also inherited problems from CON_1 and NR_2, and a new problem emerged, too:

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

Table 7 visualises the problem-solving process in this experimental complex.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
OE	**E**	**E**	**E**	**E**	**E**	**E**									
CON_1	**O**		**O**	**O**	**S**		**E**	**E**							
NR_1	**S**	**O**	**O**	**O**	**O**				**E**	**E**					
NR_2	S	**P**	O	**S**	O			**S**			**E**	**E**	**E**	**E**	
CON_2	S	**P**	O	S	S		**O**	**S**			**O**				**E**

Table 7. Overview of the re-evaluation of the experimental complex evolving from Bowdle & Gentner (2005)

D) Comparison of the problem-solving processes

As Table 7's visualisation of the upshot of the re-evaluation process we conducted shows, both limit-candidates are burdened with problems. Thus, although the pair NR_2 and CON_2 provides more plausible experimental data, as with the other two experimental complexes, no well-founded decision can be made regarding the conflict between the two series of experiments.

This motivates again the extension of this experimental complex with more refined versions of the experiments and control experiments.

E) Determination of the direction of the continuation of the cyclic process of re-evaluation

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

10.3.4. Interim summary

The application of the metascientific model we proposed suggests that it is the identification and resolution of problems which is one of the major forces of experimental work into metaphor processing. We have also seen that the problem-solving process is highly complex, so that its progressivity (that is, the solution of single problems) cannot be decisive in the evaluation of the efficacy of the whole process but the number and weight of all problems burdening experiments belonging to an experimental complex have to be taken into account, too. Further, not only do the completed steps of the problem-solving process have to be reconstructed and re-evaluated, future prospects also have to be discovered and compared. That is, the re-evaluation of experiments is not a static snapshot of the current state of the experiments at issue but a dynamic analysis of the development of the relationship among a series of related experiments and the problem-solving process. This method paves the way for designing and conducting new, more refined experiments which may produce more plausible experimental data.

11. Inconsistency resolution and statistical meta-analysis in relation to experiments in cognitive linguistics

The model with the central concept of 'experimental complex' as presented in Section 10 is, of course, only one of the possible tools for dealing with inconsistencies in relation to experiments in cognitive linguistics. In this section, we will check another route: the application of statistical meta-analysis to the problem of diverging evidence in cognitive metaphor research. We will exemplify the workability of this method with the help of a case study on experiments dealing with the impact of aptness, conventionality and familiarity on metaphor processing.

11.1. Case study 5, Part 1: Experiments on the impact of aptness, conventionality and familiarity on metaphor processing

The role of conventionality, familiarity and aptness is highly controversial in recent research into metaphor processing. A major reason for this situation is that experimental results vary in perplexing ways. This might motivate the use of *statistical meta-analysis*. Meta-analytic tools make it possible to reveal hidden and valuable information in the experiments conducted over the past decades compare and synthesise these pieces of information, and identify reliable starting points for future research. To be specific, statistical meta-analysis might allow us to

– calculate and compare the effect of conventionality, familiarity and aptness on other variables such as grammatical form preference, comprehension latencies and comprehensibility ratings on the basis of the totality of experiments conducted so far, in such a way that the impact of errors in the individual experiments may be counterbalanced,
– investigate whether the experiments indicate the same true effect size (that is, they are similar enough) or there are subgroups among them, and
– if there are experiments which indicate a small, a moderate or a large effect, to examine what kinds of other differences there are among these groups.

Statistical meta-analysis is, however, no silver bullet. Thus, it is not capable of dealing with conceptual issues and cannot eliminate the systematic errors burdening all or most of the experiments involved. Since the conceptual-theoretical background of the role of conventionality, familiarity and aptness in metaphor processing is an area of dissension in the literature, too, the question arises whether we need a tabula rasa and must search for a new beginning, or, on the basis of careful considerations and with caution, earlier experimental results can be re-analysed and synthetized. In this section, we will argue for the following hypothesis:

With the help of statistical meta-analysis, it is possible to
– provide more reliable and accurate estimates of the impact of aptness, conventionality and familiarity on metaphor processing on the basis of experiments conducted so far;
– identify subgroups among the experiments on the basis of exact criteria;
– check whether factors such the experimental design, the control experiments, or the range of metaphors in the stimulus materials have influenced the outcome of the experiments;

– under certain circumstances, re-interpret apparently diverging evidence as converging evidence;
– make a better-founded decision between rival theories of metaphor processing;
– put forward proposals for the revision of the methodology applied in this research field;
– propose new directions of research so that theoretical-conceptual problems can be resolved in future.

This section will provide an overview of the definitions of the concepts of 'conventionality', 'familiarity' and 'aptness' in the literature, as well as the problems related to their operationalization. The task of the last subsection will be to check the conditions and provide guidelines for the application of meta-analytic tools.

11.1.1. Explications of the concepts of 'conventionality', 'familiarity' and 'aptness'

There are several different explications of the concepts of 'conventionality', 'familiarity' and 'aptness' in the literature. According to Bowdle and Gentner, 'familiarity' and 'conventionality' can be distinguished as follows:

> "*Familiarity* is a property of an entire expression, and a familiar metaphor or simile involves a particular target-base pairing that has been encountered before. *Conventionality*, in contrast, is determined primarily by the base term of an expression: Conventional metaphors and similes contain base terms that have become polysemous because of repeated and consistent figurative use. Because of this, conventional figurative expressions can be either familiar or unfamiliar, depending on the target term that has been paired with the base." (Bowdle & Gentner 2005: 204)

That is,

> "*Conventionality* is the strength of association between a metaphor vehicle and its figurative meaning" (Jones & Estes 2006: 19; emphasis added)

Jones and Estes argue for a clear differentiation between 'conventionality' and 'aptness'. As regards the latter,

> "*[a]ptness* is the extent to which the vehicle's figurative meaning expresses an important feature of the topic […]. For a metaphor to be apt, two conditions must be met. First, the vehicle term must have a salient property for attribution. […] The second necessary condition of aptness, then, is that the salient property of the vehicle must be relevant to the topic. […] If the property implied by the vehicle is irrelevant to the topic, then the metaphor will be less apt […]. Thus it is the interaction between topic and vehicle that is critical for aptness." (Jones & Estes 2006: 19; emphasis added)

Gentner & Bowdle (2008: 122), however, cast doubt on the usability of the concept 'aptness', because, on their view, it is a by-product of the evaluation of metaphors and plays no role in figurative language processing.

In contrast, according to Thibodeau & Durgin (2011: 11), the concept of 'conventionality' is unusable, because "the construct cannot be defined for vehicles independent of topics". These researchers argue for the application of the concept of 'familiarity' instead, which takes

into consideration both the targets/topics and the bases/vehicles and the relationship between them.

11.1.2. The operationalization of the three concepts

Experiments on metaphor processing usually collect participants' ratings with the help of norming studies in order to determine the stimulus items' conventionality, familiarity or aptness. This method, however, was judged to be problematic from several points of view in the literature.

A) In general

Thibodeau, Sikos and Durgin summarize the findings of two experiments they conducted as follows:

> "In Experiment 1, we showed that a context manipulation affected how fluently people processed metaphors. In Experiment 2, we showed that the same context manipulation affected ratings of the comprehensibility, familiarity, aptness, surprisingness, and metaphoricity of the target metaphors." (Thibodeau et al. 2018: 769)

On the basis of these results, they raise the general concern that the outcome of ratings tasks should not be used as independent variables mirroring autonomous concepts, but should be regarded as *dependent variables*, more precisely, as *measures of processing fluency*.

B) Specifically

Aptness: From the results of two experiments, Thibodeau and Durgin (2011: 10f.) concluded that participants' ratings do not mirror the concept of 'aptness' as defined in Section 11.1.1. Namely, they found that *the number* of the relevant (applicable, mapped) features were *positively correlated* with aptness ratings, and, more interestingly, the number of the irrelevant (non-applicable, unmapped) features were *strongly negatively correlated* with published ratings of aptness. Further, a model making use of predictors for both relevant and irrelevant feature counts was markedly better than a model taking into consideration only the number of the relevant features. From these findings the authors concluded that participants also seem to take into consideration the *number* of features of the base/vehicle term which are *relevant* and, more importantly, those which are *irrelevant* to the characterisation of the target/topic. Moreover, they also revealed a correlation between aptness ratings and corpus frequency. They interpreted this result in such a way that subjective ratings of aptness exhibit the degree of *felt familiarity*.

From a different angle, Gentner and Bowdle raise the objection that aptness ratings strongly correlate with other dimensions of metaphor processing such as "relationality […], ease of interpretation, degree of metaphoricity, imagery, subjective familiarity, and the number of alternative interpretations possible […], as well as with ease of comprehension […]" (Gentner & Bowdle 2008: 122).

Familiarity: According to Thibodeau & Durgin (2011: 11), instead of subjective ratings, corpus frequency should be used as a more direct and objective measure of metaphor familiarity. Dulcinati et al. (2014: 77) provide detailed guidelines for the application of Google searches

as a possible operationalization of this concept. Bambini et al. (2014), however, suggests that familiarity is subjective in the sense that

> "[f]amiliarity reflects how often a subject has been exposed to a particular statement either in written or oral form. It does not overlap with frequency, as an item may be frequent on a lexical database but unfamiliar to a single individual. Familiarity is thus best defined as frequency of experience or 'felt familiarity'." (Bambini et al. 2014: 3)

From this concise overview we can conclude that the operationalization of the three concepts is quite controversial in the literature.

11.1.3. Guidelines for subsequent meta-analyses

As we have seen, all aspects of empirical research into the effects of conventionality, familiarity and aptness on metaphor processing are strongly controversial in the literature. There seem to be no fixed points. This could mean that a new beginning is inevitable and previous work in this field has to be discarded. The question is, of course, whether such a fatal decision is reasonable and unavoidable, or whether there is a chance of saving and using the results accumulated so far.

Statistical meta-analysis might allow us to take a middle course between these two extremes. This is because it makes it possible to analyse experimental results from a new perspective in order to motivate and find reliable starting points for future research on the basis of the comparison and synthesis of the information lying hidden and unrevealed in the experiments conducted over past decades.

Nevertheless, against the background of the problems mentioned in this section, it is anything but self-evident that statistical meta-analysis is apposite to this issue. Therefore, we first have to lay down the conditions for the application of meta-analytic tools. Thus, on the basis of the criticisms discussed in Subsections 11.1.1-2, we put forward the following *guidelines for subsequent meta-analyses*:

– **On the usability of ratings in general**: The stimulus material of all experiments involved contains context-free metaphorical sentences. Further, from a single experimental result indicating that the context is capable of influencing both processing fluency and the familiarity and aptness of metaphorical expressions, it would be premature to conclude that the latter are not autonomous concepts.
  Guidelines: We will deem participants' ratings as relevant and legitimate data, whose interpretation and confrontation with the hypotheses of the rival theories, however, needs to be re-thought.
– **On the concepts 'conventionality', 'familiarity' and 'aptness'**: Basically, we will suppose that all three concepts rely on and mirror the subjective judgements of language users.
  Guidelines: 'Conventionality' will be defined as *the strength of association between a metaphor base/vehicle and its figurative meaning for an individual*. Similarly, 'familiarity' is *the frequency of experience or 'felt familiarity' of a base/vehicle and target/topic pair*. Thirdly, 'aptness' is the felt applicability of the salient features of the base/vehicle to the target/topic – that is, the proportion of *the number of salient features of the base/vehicle*

term which seem to be important and relevant to the characterisation of the target/topic to those which are deemed to be inapplicable.

– **On the operationalization of the three concepts**: The experiments conducted so far have applied different methods of the operationalization of the three concepts. These methods mirror participants' individual judgements only indirectly, because no experiment has made use of the conventionality/familiarity/aptness ratings of the participants of the main experiment. Instead, they have relied on separate groups of subjects recruited from the same population, or applied corpus linguistic methods by supposing that corpus frequency corresponds to the judgements of the whole population and via this, to the narrower population taking part in the experiments.
  Guidelines: We have to pay close attention to the methods of ratings collection, and within this, the instructions participants received. It has to be examined whether differences in the methods and instructions might have influenced participants' behaviour.
– **On the relationship between conventionality/familiarity/aptness ratings and theories of metaphor processing**: Both the predictions which can be drawn from the two rival theories of metaphor processing and their confrontation with the experimental data (form preferences/ratings/latencies) have to be re-thought.[84]

11.2. Basic ideas and concepts of statistical meta-analysis

This section provides a concise overview of the most important ideas and concepts of meta-analysis. The reader is recommended to skim through this section in order to become acquainted with the theoretical background and come back for a short consultation if needed in the course of the application of meta-analytic tools in Section 11.3.

11.2.1. The aim of statistical meta-analysis

Meta-analysis attempts to accumulate all available pieces of information so that the shortcomings of individual experiments can be counterbalanced, and more robust results can be obtained. As Geoff Cumming puts it,

> "Meta-analytic thinking is estimation thinking that considers any result in the context of past and potential future results on the same question. It focuses on the cumulation of evidence over studies." (Cumming 2012: 9)

Statistical meta-analysis is *the application of statistical thinking and of statistical tools at a meta-level.* The objects of this meta-level analysis are the results of a series of experiments as data points. Its aim is to estimate the strength of the relationship between two (or more) variables. Hence, it works with *effect sizes*, first at the level of the individual experiments and then at the level of their synthesis. There are several types of effect size (Pearson's correlation coefficient, Cohen's d, odds ratio, raw difference of means, risk ratio, Cramér's V, etc.), which can be converted into each other.

[84] See Section 15 on this.

According to Borenstein et al. (2009: 297ff.), focusing on the effect sizes is considerably more instructive than the use of p-values, because it also provides information about the *magnitude of the effect*. That is, a higher effect size indicates a stronger relationship between the variables. Moreover, if we calculate confidence intervals for them, then they also reveal whether the result is statistically significant. In this way, we may obtain information about

– the magnitude of the effect (distance from the null-value);
– the direction of the effect (positive vs. negative, showing an effect in the predicted or the opposite direction);
– the precision of the effect estimate (width of the confidence interval).

The application of effect sizes also makes it possible to compare and synthesize the outcome of a set of similar experiments. Thus, for example,

– there may be a considerable overlap among their confidence intervals (or one of them may completely contain the other one), indicating a harmony among the results of the different experiments;
– the confidence intervals may be totally distinct, pointing to a case of heterogeneity;
– between these two extremes, there may be a small overlap among the confidence intervals, suggesting the compatibility of the results;
– even if one of the confidence intervals includes the null value (indicating a non-significant result) while the other confidence interval is above the null value, the two experiments' results may be compatible or even in harmony.

Therefore, statistical meta-analysis is a possible tool of conflict resolution. It allows us to calculate a summary effect size by taking into consideration the effect size of the individual experiments, their precision (confidence intervals) and size (number of participants).

11.2.2. The selection of experiments included in the meta-analysis

The first step of a meta-analysis is the selection of the experiments. The decisive point is that in order to be combinable all experiments have to test the same research hypothesis, or their research hypotheses have to share a common core. This means that all experiments should provide information about the relationship between two variables, so that the strength of this relationship is determinable in each case.

In our case, we divided experiments which produced experimental data about the effect of conventionality, familiarity and aptness on metaphor processing into three groups. We investigated separately experiments dealing with the relationship between grammatical form preference, comprehensibility ratings and comprehension latencies and the three factors mentioned.

11.2.3. The choice and calculation of the effect size of the experiments

With the help of the CMA software, effect sizes and their 95% confidence intervals can be computed from more than 100 summary data types, but there are also several online effect size calculators such as this one: https://www.psychometrica.de/effect_size.html. Reliance on the summary data presented in the experimental reports is not a compulsory step of meta-analysis

but often a necessity, because we do not usually have access to the data sets. Nonetheless, if the data sheets are made available by the researchers, it is better (i.e. will result in more precise effect size values) to make use of the raw data than to rely on the summary data as published in the research papers.

In our case, the choice of the effect size was straightforward, because many relevant studies provided correlation coefficients in their results section. Thus, we could determine in each case the strength of the correlation between the variables of conventionality/familiarity/aptness and grammatical form preference ratings/comprehensibility ratings/comprehension latencies from the experimental data available in the papers. Mostly, the mean and standard deviation of the ratings/latencies of the two groups (for example, low apt vs. high apt) could be used to calculate the correlation coefficient.

11.2.4. Synthesis of the effect sizes

Basically, the summary effect size is calculated as a weighted mean of the experiments' effect sizes. There are two methods to combine the effect sizes of individual experiments: the fixed-effect model and the random-effect model. Following Borenstein et al.'s (2009: Part 3) characterisation, the two methods can be described as follows.

The *fixed-effect model* should be applied if the experiments to be combined made use of the same design, their participants share all relevant characteristics which might influence their performance, they were performed within a relatively short time frame by the same researchers in the same laboratory, etc. If all circumstances are practically identical in each case, then we can suppose that the experiments have the same true (underlying) effect size, and any difference between the values in the individual studies is due solely to sampling error. Thus, fixed-effect models offer an estimation of the common (underlying, true) effect size. *Random-effect models*, in contrast, can be applied if, despite their important similarities, there are also substantial differences among the experiments. In fact, in the great majority of cases, we have to assume that the experiments differ from each other regarding their underlying (true) effect size. Our task is to estimate the mean of the distribution of the true effect sizes, which has to take into consideration, besides the within-study error, the between-study variation, as well.

Since with a fixed-effect model, all experiments provide information about the same true effect size, greater importance (weight) should be attached to larger experiments when calculating the summary effect size. As for random-effect models, every experiment contributes to the summary effect size from a different point of view. Thus, smaller experiments should receive a somewhat greater importance than in the fixed-effect case, and, conversely, the impact of larger studies should be moderated in comparison to the fixed-effect models. This can be achieved in such a way that the weights assigned to the experiments involve the between-studies variance, too.

In our case, the application of random-effect models is undoubtedly the right choice, because there were considerable differences in the stimulus materials used, the instructions participants received, and the range and characteristics of participants. Furthermore, the experiments were conducted by different researchers in different laboratories at different time points.

11.2.5. The prediction interval

The prediction interval provides us information about the *dispersion* of the effect sizes. That is, it tells us whether *a new experiment* will probably have a true effect size falling between certain limits. Or to put it differently, the 95% prediction interval tells us in which range the true effect size of the whole population could be found in 95% of the cases. This interval is always wider than the confidence interval of the summary effect, since the latter shows us where the true *mean effect size of a series of experiments* will fall in 95% of the cases.

11.2.6. Consistency of the effect sizes

The consistency of the (true) effect sizes can also be investigated.[85] The *Q* statistic describes the *total amount of the observed between-study variance*. This total dispersion has to be compared with the expected value of this variance, that is, with its value calculated when supposing that the true effect sizes were identical in all experiments. This latter value is simply the degree of freedom (*df*). The difference between the total variance and its expected value gives the *excess dispersion of the effect sizes*, i.e. the real heterogeneity of the effect sizes. In relation to this, the first important information is *whether Q is significantly different from its expected value*. The second relevant issue is an estimate of the between-study standard variation of the true effects, denoted as T^2, computed from the excess dispersion in the true effect sizes – or more intuitively, *T* is the estimate of *the standard deviation in the true effects*. The third useful indicator is the *ratio of the excess dispersion (Q – df) and the observed between-study variance* (*Q*). This is the I^2 statistic. The higher its value, the more real variance there is within the observed variance, and the less dispersion due to random error. A high I^2 value indicates that if all experiments were conducted by a huge number of participants, then the observed variance would barely decrease, because the sampling error is small and the larger part of observed variance is real. In such cases, it is advisable to conduct subgroup analyses or meta-regression in order to find out whether there are subgroups among the studies indicating some methodological or other differences, or subgroups among participants which behave differently.

11.2.7. Publication bias

Meta-analysis also includes tools for the estimation of possible publication bias. Publication bias often results from the circumstance that experiments showing a significant result are more likely to be published than those indicating an insignificant result. Since experiments with a small number of participants produce significant results only if the effect size is large, they might remain unpublished more easily due to their low power.

There are several methods for checking publication bias. Their power might, however, be low with small numbers of experiments.

One method to check whether smaller studies with negative outcomes have been neglected is to examine the disposition of studies around the mean effect size. Large, medium-sized and smaller experiments alike should be located symmetrically on the two sides of the mean effect size. We can visualise this with the help of *funnel plots*. A funnel plot is a special scatter plot. It shows the standard error of the effect sizes as a measure of the experiments' size or precision on the vertical axis in such a way that the larger, more precise studies are towards the top, and

[85] See Borenstein et al. (2009: Part 4) and Borenstein et al. (2017) for more on this topic.

the smaller/less precise experiments are at the bottom. If there is publication bias, then there will be an asymmetry in the case of small studies so that the number of experiments showing a positive result will be greater than those producing a negative result. Funnel plots also provide us with valuable information about heterogeneity: a triangle indicates an area within which 95% of the experiments should be found. Experiments plotted outside of this area indicate the presence of heterogeneity.

Duval and Tweedie's *Trim and Fill* method allows us to estimate the true effect size correcting for publication bias. To this end, the list of experiments is supplemented by fictional smaller experiments so that the symmetry is restored, and the summary effect size is re-calculated and compared to its original value. Nonetheless, it is important to bear in mind that this method can be applied reliably to at least 10 experiments.

Egger's test indicates a bias if it produces a significant result, although its power might be low with small numbers of experiments.

Another possibility is to conduct a *cumulative meta-analysis*. For a cumulative analysis, the experiments are ordered by their size. We start with the largest experiment, then we add the experiments one by one towards the smaller ones, and at each step we calculate the summary effect size. In this way, we can check whether the summary effect size changes if we take into consideration the smaller experiments.[86]

11.3. Case study 5, Part 2: Meta-analysis as a tool of inconsistency resolution

In this section, we will apply the tools of statistical meta-analysis briefly presented in Section 11.2 to experiments testing the impact of conventionality, familiarity and aptness on three aspects of figurative language use: grammatical form preference, comprehension latencies and comprehensibility.

11.3.1. Grammatical form preference

A) Conventionality

Most experiments dealing with the impact of conventionality on grammatical form preference have relied on a prior experiment in which a separate group of participants rated the conventionality of the base/vehicle term, while one experiment applied a post hoc control experiment. As for the main experiment, there were four types. In the first type (*grammatical form preference ratings, GFPR*), participants were asked to indicate whether they prefer (i.e. feel to be more natural or sensible) a figurative statement in metaphor or simile form by using a rating scale. In a subtype of GFPR, the conventionalization process was speeded up in such a way that as a pre-task, participants had to read novel similes using the same base/vehicle term paired with several different base/vehicle terms (*in vitro conventionalization, IVC*). A second type of design (*interpretation predication check, IPC*) collected interpretations of the figuratives from participants and divided them into two groups on the basis of "whether the description was applied to the target/topic term alone (target/topic-only predications) or to both the target/topic

[86] For a more detailed description of these methods, as well as for further tests, see Borenstein et al. (2009: Section 30).

term and the base/vehicle term (double predications)" (Bowdle & Gentner 2005: 205). Double predications indicate that the figurative statement at issue was comprehended as comparison (that is, as a simile), while single predication suggests that the figurative was seen as categorization (that is, as a metaphor). In the third type of experiments (*category membership ratings, CMR*), category membership ratings were collected and evaluated. That is, participants had to judge to what extent the target/topic is a member of a category named after the base/vehicle. The fourth type was a figurative statement production task (*figurative statement production, FSP*). Participants had to create a figurative statement after seeing a target/topic term and a property on the screen, that is, they had to find the base/vehicle term which best ascribes the property at issue to the target/topic. They were encouraged to choose between a metaphor and a simile form in each case.

Table 8 in Appendix 1 summarises the experimental data from 14 experiments on the basis of which the correlation coefficients between conventionality and grammatical form preference can be calculated.

The CMA software computes the correlation coefficients of each experiment, their confidence intervals, *Z*-value, *p*-value and weight, as well as the summary effect size. Since the experiments were conducted by different researchers making use of different methodologies, the application of a random-effects model (see Section 11.2.4) is clearly advisable. See Figure 13.

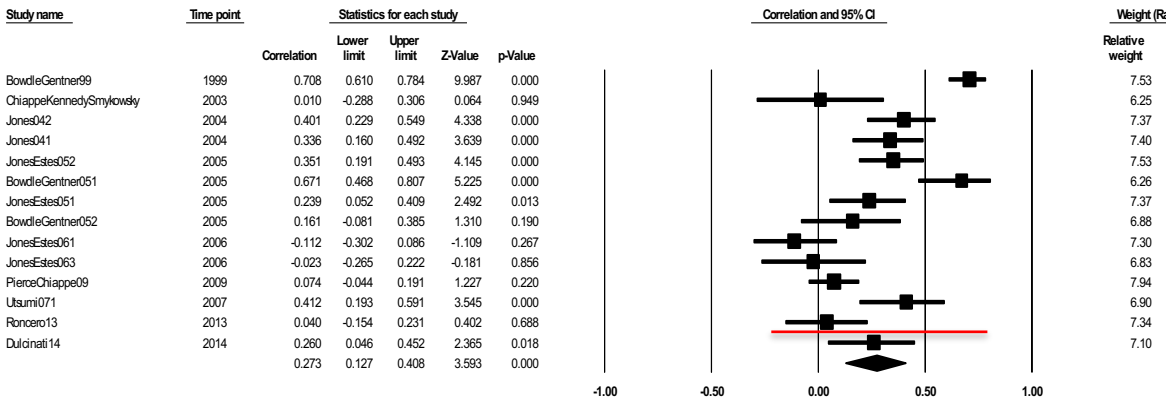

Figure 13. Random-effects model of grammatical form preference with conventionality as a decisive factor

The first thing which catches the eye is that there is no overlap among all the confidence intervals of the individual experiments. Despite this, the majority of the confidence intervals partially cover each other. A second impression is that instead of the binary significant vs. non-significant division, we can compare the outcome of the experiments with each other and characterise their relationships in a more detailed and precise manner.

The summary effect size is $r = 0.273$ with a rather narrow 95% confidence interval of [0.127; 0.408]. This indicates a rather weak but clearly positive correlation between base/vehicle conventionality and grammatical form preference. That is, the totality of the experiments taken into consideration provides evidence for the hypothesis that conventionality is a relevant factor in relation to grammatical form preferences with a relatively high accuracy – at least, if we accept the background assumption that 'conventionality' has to be interpreted as subjective

judgements on base/vehicle conventionality, and conventionality ratings mirror this concept reliably.

The prediction interval (cf. Section 11.2.5) is as wide as [-0.321; 0.713], as indicated by the red line in Figure 13. This means that the true effect size for any similar experiment will fall into this range in 95% of cases, provided that the true effect sizes are normally distributed (while the true mean effect size will fall into the confidence interval in 95% of cases). This prediction interval provides an inconclusive picture insofar as one cannot predict whether a similar experiment would indicate any effect of the metaphorical frame – a small reversed effect, no effect or a large effect are all equally possible.

As for possible heterogeneity (see Section 11.2.6), the total amount of the observed between-study variance, $Q = 105.278$, is significantly different from its expected value, $df(Q) = 13$. The standard deviation of the true effect sizes is $T = 0.264$. The value of the I^2 statistic is 87.652, i.e. almost 88% of the observed variance is real variance. To put it differently, if all experiments were conducted by a huge number of participants (so that there were no sampling errors), then the observed variance would only decrease by 12%. From these pieces of information we can conclude that there is a considerable amount of heterogeneity in our data, the majority of which cannot be due to sampling error. This means that we should try to reveal the causes of this heterogeneity by performing subgroup analyses. If we return to Figure 13 and examine the confidence intervals of the experiments, we can see that there are three experiments (Bowdle & Gentner 1999, Bowdle & Gentner 2005, Experiment 1, and Jones & Estes 2006, Experiment 1) whose effect sizes' confidence intervals do not overlap with the confidence interval of the summary effect size. We might try to eliminate at least the two experiments conducted by Bowdle & Gentner, because their confidence intervals do not, or only slightly, overlap with the confidence intervals of the other experiments. As a consequence, the Q-statistic drops to 36.21 (which is still significantly different from its expected value of 11), with an I^2 of 69.622. It is, however, not completely clear why these experiments are outliers. The removal of Bowdle & Gentner (1999) could be justified by reference to the application of the in vitro conventionalization technique; their other experiment, however, did not use this method. A further idea could be to conduct a subgroup analysis by authors as a grouping variable. This procedure does not produce useable results, either, because there is a considerable amount of within-group variance both among the experiments conducted by Bowdle and Gentner and among those conducted by other researchers.

The barrenness of these two attempts could motivate a change of perspective insofar as we might try the opposite route. Namely, on the basis of their effect sizes, the 14 experiments can be easily divided in 3 distinct groups. See Table 9.

group	below average effect size	average effect size	above average effect size
experiments	ChiappeKennedySmykowsky2003 JonesEstes2006/1 JonesEstes2006/3 PierceChiappe2009/1 Roncero2013	Jones 2004/1 Jones 2004/2 BowdleGentner2005/2 JonesEstes2005/1 JonesEstes2005/2 Utsumi2007/1 Dulcinati2014	BowdleGentner1999 BowdleGentner2005/1
summary effect size	0.022 [-0.06; 0.103]	0.317 [0.246; 0.384]	0.699 [0.614; 0.768]
within groups variance	2.668 (p = 0.615)	4.837 (p = 0.565)	0.151 (p = 0.698)
between groups variance	97.623		
design	3 x GFPR, 1 x FSP, 1 x CMR	2 x GFPR, 4 x CMR, 1 x IPC	1 x GFPR, 1 x GFPR + IVC
control of conventionality	1 x rating how conventional it is to use a word to convey the most common interpretation 2 x detailed explanations + rating how conventional it is to use the concept to represent the given property 1 x rating how common it is to use the base/vehicle to convey candidate properties derived by the authors 1 x rating how common it is to use the base/vehicle to convey the most frequently generated property	3 x rating how conventional it is to use the base/vehicle term to convey the most common interpretation 4 x rating how conventional or familiar the canonical metaphoric meaning as an alternative sense of the base/vehicle term is	1 x rating how conventional or familiar the canonical metaphoric meaning as an alternative sense of the base/vehicle term is 1 x – (conventionalization process = repeated experiences of the base/vehicle term)
range of metaphors	3 x all, 1 x examples from other papers + Google search	4 x high similarity, 3 x all	2 x all

Table 9. Three possible relevant factors in the three groups of experiments on grammatical form preference with conventionality as a decisive factor

It is important to realise that this grouping does not conform to a significant vs. non-significant division. Thus, Experiment 2 of Bowdle & Gentner (2005) produced an insignificant result (since its confidence interval includes the 0 value); despite this, it belongs to the group of average effect sizes.

As Table 9 shows, three kinds of factors were investigated as to whether they might make it possible to separate the three groups from each other: the applied experimental design, the formulation of the task in the control experiments on conventionality, and the range of the metaphors included in the stimulus materials. None of these, however, seems to be decisive. This means that the true effect size should be stable against variations in these three factors, and the differences among the groups could be due to some other factor. It is possible, for

example, that peculiarities of the stimulus materials are responsible for the heterogeneity of the effect sizes. This motivates a close inspection and comparative analysis of the stimulus materials used in the experiments. The problem is, however, that these were not included in the research papers in each case.

Nonetheless, there is an important caveat: none of the experiments have been replicated so far. Therefore, it might be the case that if we conducted all experiments again, they would yield different results, and, as a consequence, different groups among them. This scenario cannot be ruled out – what is more, against the background of the methodological-theoretical criticism discussed in Section 11.1, this is a quite strong possibility.

Finally, we have to check whether there is publication bias (cf. Section 11.2.7). Duval and Tweedie's trim and fill method indicates two missing, medium-sized studies to the *right* of the mean (black dots). The extension of the set of experiments with the missing ones yields a slightly higher summary effect size, as indicated by the black rhombus below. See Figure 14.

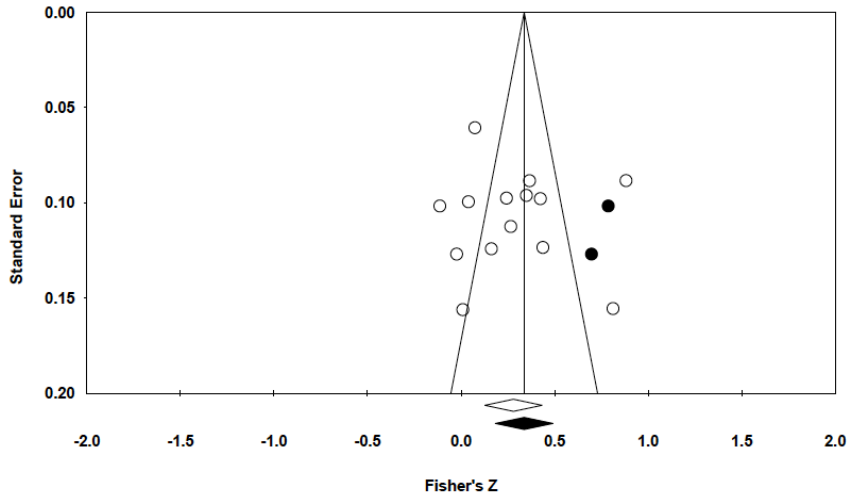

Figure 14. Funnel plot for grammatical form preference with conventionality as a decisive factor

From this we may conclude that there is a slight bias in our results. This does not result, however, from missing small non-significant experiments (as is also implied by the non-significant result of Egger's test) but rather, from missing average sized experiments producing higher effect sizes. A cumulative meta-analysis reinforces this interpretation insofar as it does not show any clear tendency; the smallest experiments are even farther from the null-value than the biggest ones. See Figure 15.

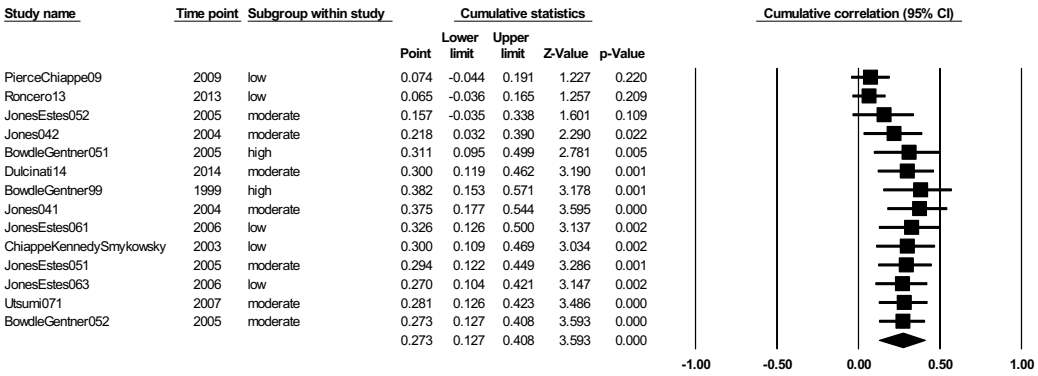

Figure 15. Cumulative forest plot for grammatical form preference with conventionality as a decisive factor

Taking the results of the different methods together, one cannot rule out the possibility that the slight bias they yield is due to the high amount of heterogeneity we detected.

B) Aptness

A second series of experiments was designed to check whether it is aptness that determines grammatical form preference. See Table 10 in Appendix 1 for the relevant experimental data.

Figure 16 presents the results of a random-effects meta-analysis.

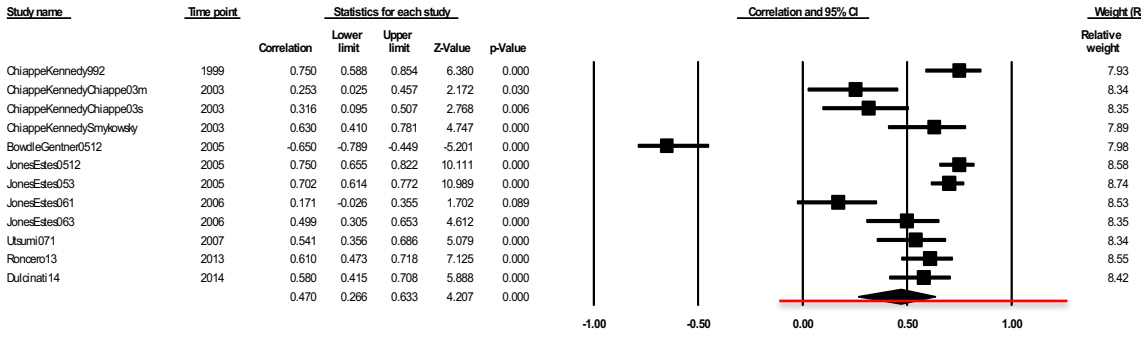

Figure 16. Random-effects model of grammatical form preference with aptness as a decisive factor

The summary effect size is substantially higher than in the previous case: 0.47 with a 95% confidence interval of [0.266; 0.633], indicating that aptness exerts a stronger influence than conventionality.

The prediction interval is signally wide at [-0.402; 0.895], as indicated by the red line in Figure 16. Consequently, the true effect size for any similar experiment will fall into this range in 95% of cases, provided that the true effect sizes are normally distributed. This means that a similar experiment could yield almost anything, from a moderate reverse effect to a very large effect of the metaphorical frame.

As for the consistency of the effect sizes, the Q-value was 150.738, significantly different from its expected value df(Q) = 11. Therefore, there is a huge amount of heterogeneity. The standard deviation of the true effect sizes is $T = 0.166$. The value of the I^2 statistic is 92.703, i.e. about 93% of the observed variance is real and does not result from sampling error. As Figure 16 shows, there is an especially extreme outlier: Bowdle & Gentner (2005), Experi-

ments 1-2, which, in contrast to all other experiments, indicate a reverse effect; moreover, their confidence interval does not overlap with that of the summary effect size or those of the other experiments. Therefore, it seems to be reasonable to omit this study. If we exclude this outlier from the random-effects analysis, the summary effect size increases to 0.551 with a 95% confidence interval of [0.424; 0.658]. The prediction interval reduces to [-0.002; 0.846], which is still very wide and practically uninformative because it only rules out a reverse effect. The total amount of the observed between-study variance, Q, reduces to 65.329, although this value is significantly different from its expected value, df(Q) = 10; T = 0.254, I^2 = 84.693. That is, as was the case with conventionality, if all experiments were conducted by a huge number of participants (so that there were no sampling errors), then the observed variance would barely decrease. The question is, of course, what the cause of this finding might be. A grouping on the basis of the researchers is clearly pointless. If we conduct a subgroup analysis on the basis of the effect sizes as in the previous case, then the following groups present themselves. See Table 11.

group	below average effect size	average effect size	above average effect size
experiments	ChiappeKenndyChiappe2003/sim ChiappeKenndyChiappe2003/met JonesEstes2006/1	ChiappeKennedySmykowsky2003 JonesEstes2006/3 Utsumi2007/1 Roncero2013 Dulcinati2014	ChiappeKennedy1999/2 JonesEstes2005/3 JonesEstes2005/1-2
summary effect size	0.239 [0.117; 0.355]	0.572 [0.499; 0.638]	0.726 [0.669; 0.775]
within group variance	1.011 (p = 0.603)	1.579 (p = 0.813)	0.807 (p = 0.668)
between groups variance	61.932		

Table 11. Three groups of experiments on grammatical form preference with aptness as a decisive factor

Here again, only one experiment in the low group produced an insignificant result, the other two were significant. The three factors of experimental design, the formulation of the task in the control experiments on aptness, and the range of the metaphors included in the stimulus materials did not influence the effect size of the experiments.

There is no publication bias according to Duval and Tweedie's trim and fill model, and this is reinforced by a non-significant Egger-test.

C) Familiarity

Table 12 in Appendix 1 shows the data pertaining to familiarity as a possibly relevant factor.

As Figure 17 indicates, the summary effect size is 0.393 with a 95% confidence interval of [0.215; 0.546].

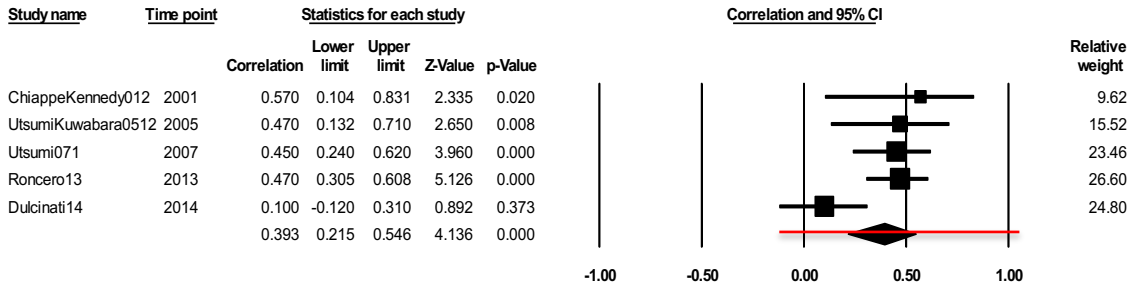

Figure 17. Random-effects model of grammatical form preference with familiarity as a decisive factor

The prediction interval is as wide as [-0.203; 0.777]. From these results we may conclude that the strength of the effect of familiarity is between those of conventionality and aptness. Here again, we have an outlier: Dulcinati (2014) is the only experiment which produced a correlation coefficient near to 0, although its confidence interval overlaps with that of the others. Thus, it is no wonder that the *Q*-statistic is significantly different from its expected value (9.918 vs. 4), $p = 0.042$ and, as the I^2 value of 59.670 indicates, almost 60% of the observed variance is real. The standard deviation of the true effect sizes is $T = 0.166$. These data point towards the hypothesis that the experiments do not share a common true effect size. As the total amount of variance of the four experiments with a relatively higher effect size in Table 13 indicates, these experiments are in harmony with each other.

group	**below average**	**above average**
experiments	Dulcinati2014	ChiappeKennedy2001/3 UtsumiKuwabara2005/1-2 Utsumi2007/1 Roncero2013/1
summary effect size	0.100 [-0.120; 0.310]	0.470 [0.358; 0.569]
within group variance	0	0.289
between groups variance	9.630	

Table 13. Two groups of experiments on grammatical form preference with familiarity as a decisive factor

The summaries show only one substantial difference between the two groups. Namely, while the experiments conducted by Chiappe and Kennedy, Utsumi and Kuwabara, and Utsumi and Roncero relied on participants' familiarity ratings, Dulcinati et al. applied a Google search instead. This explanation, however, contradicts the findings of Thibodeau and Durgin (2011), who found a strong correlation between familiarity ratings and frequency counts based on Google searches. Therefore, further experiments are needed to resolve this conflict.

11.3.2. Comprehension latencies

A) Familiarity

Chronologically, it was familiarity whose impact on comprehension latencies was first checked with the help of experiments. Table 14 in Appendix 1 summarises the most important characteristics of 14 related experiments.

As Figure 18 reveals, the great majority of the confidence intervals overlap, suggesting that there should be no heterogeneity in the results.

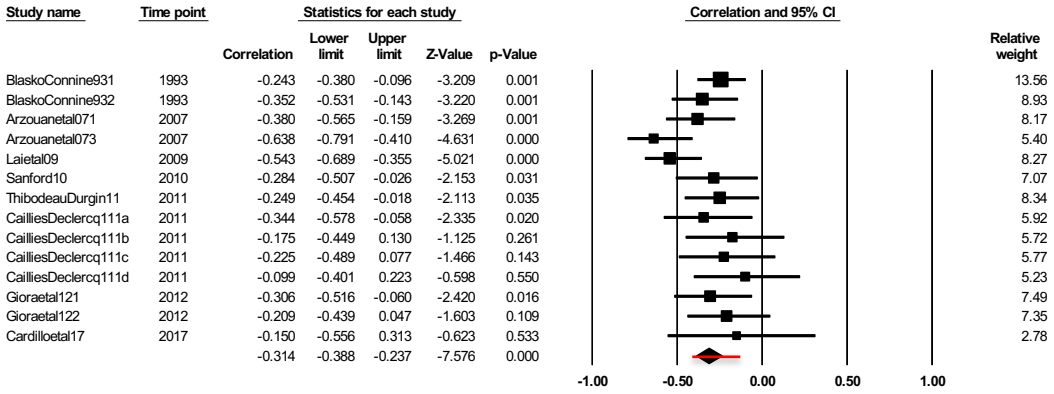

Figure 18. Random-effects model of comprehension latencies with familiarity as a decisive factor

The experiments together produce an effect size of -0.314 very precisely, since the 95% confidence interval is as narrow as [-0.388; -0.237]. The prediction interval at [-0.494; -0.109] is quite narrow, too, indicating that a future experiment should yield a small to moderate reverse effect of the metaphorical frame on the comprehension latencies. The *Q*-statistic reinforces our impression that the results of the experiments are in harmony with each other, since its value of 18.733 is not significantly different from the expected 3, $p = 0.132$. The standard deviation of the true effect sizes, $T = 0.087$, is low. An I^2 value of 30.604 indicates that only 30% of the observed variance is real, and 70% is due to random error. From these data a very important conclusion can be drawn. Namely, the experiments above seemed to constitute diverging evidence in the sense that 5 of the 14 studies produced insignificant results, while 9 produced significant ones. In the absence of heterogeneity, however, this conflict can be *resolved*: the outcome of the experiments can be interpreted as an instance of *converging evidence* for the summary effect size, i.e., a small-moderate effect.

According to Duval and Tweedie's trim and fill model, three small experiments are missing from the left side. See Figure 19.

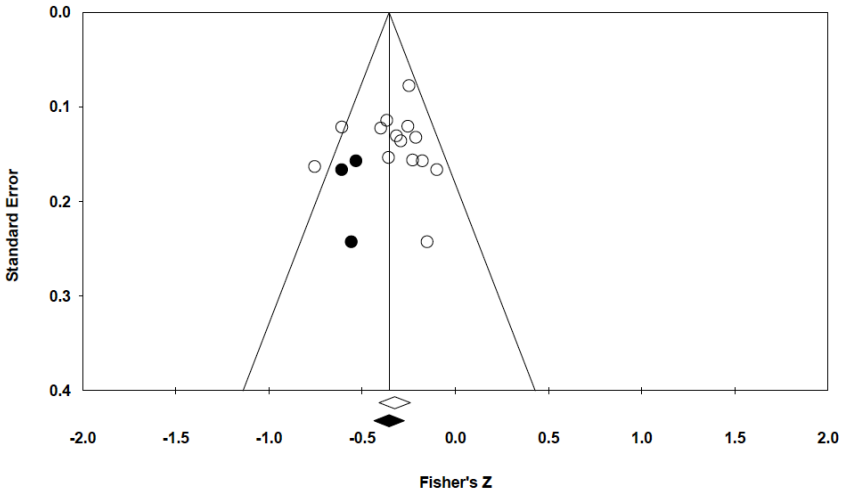

Figure 19. Funnel plot for comprehension latencies with familiarity as a decisive factor

Nonetheless, the adjusted values are similar to the observed values. Egger's test is not significant, $p = 0.934$, suggesting that there is no bias. Since the power of this test is weak, a cumulative meta-analysis seems to be advisable. Figure 20 also shows that there is no sign of any publication bias.

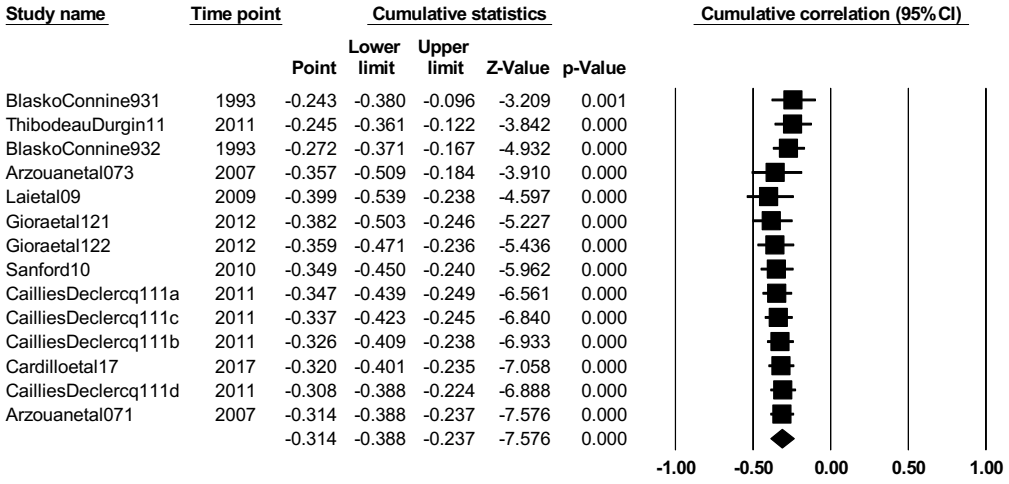

Figure 20. Cumulative analysis for comprehension latencies with familiarity as a decisive factor

The upshot of the tests presented is that there is no publication bias.

B) Aptness

The second factor which had been regarded as relevant by some researchers is aptness. Consult Table 15 in Appendix 1 for the details.

The confidence intervals, as shown in Figure 21, overlap only slightly.

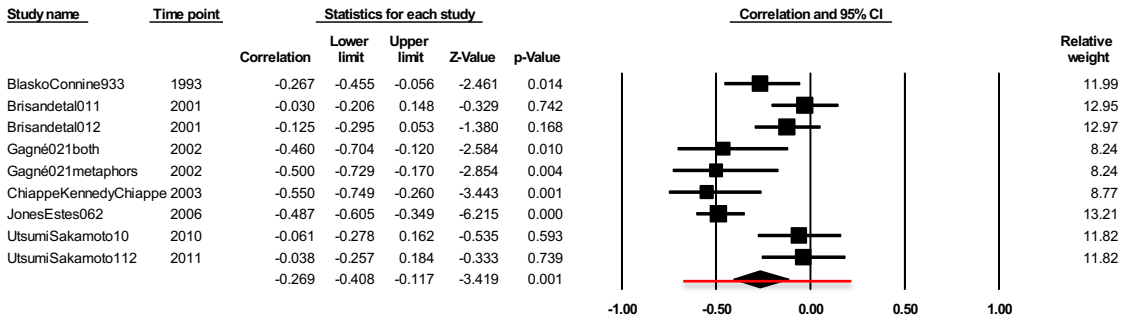

Figure 21. Random-effects model of comprehension latencies with aptness as a decisive factor

The summary effect size lies at -0.269 [-0.408; -0.117], indicating a small reverse effect. The prediction interval of [-0.662; 0.24] is indeterminate insofar as it allows a remarkable reverse effect but also a small effect. Since the confidence intervals do not overlap in each case, there should be some amount of heterogeneity. In fact, the Q-statistic (32.961) is significantly different from its expected value, df(Q) = 8. The I^2 value of 75.729 indicates that almost 75% of the observed variance is real, and only 25% is due to random error. The standard deviation of the true effect sizes is $T = 0.201$. This suggests that the experiments do not share a common true effect size and motivates a subgroup analysis. If we create three groups on the basis of the

effect sizes in such a way that the first group consists of experiments with an effect size between -0.2 and -0.3, the second group of experiments with effect sizes close to -0.5, and a third group of experiments with effect sizes close to 0, then we obtain three homogenous groups. See Table 16 for an overview.

group	below average effect size	average effect size	above average effect size
experiments	Brisand2001/1 Brisand2001/2 UtsumiSakamoto2010 UtsumiSakamoto2011/2	BlaskoConnine1993/3	Gagné2002/1both Gagné2002/1met ChiappeKennedy Chiappe2003 JonesEstes2006/2
summary effect size	-0.067 [-0.164; 0.032]	-0.267 [-0.455; -0.056]	-0.495 [-0.588; -0.389]
within group variance	0.643 (p = 0.887)	0 (p = 1)	0.248 (p = 0.969)
between groups variance	32.071		

Table 16. Three groups of experiments on comprehension latencies with aptness as a decisive factor

As with the previous cases, neither the experimental design, nor the formulation of the task nor the range of metaphors seems to influence the effect sizes.

As for publication bias, for the application of Duval and Tweedie's trim and fill model we would need at least one further experiment. The low power of Egger's test makes its application questionable in this case, too.

C) Conventionality

Table 17 in Appendix 1 summarises the most important data pertaining to the relevant experiments. Figure 22 presents how statistical meta-analysis makes the comparison and combination of these results possible.

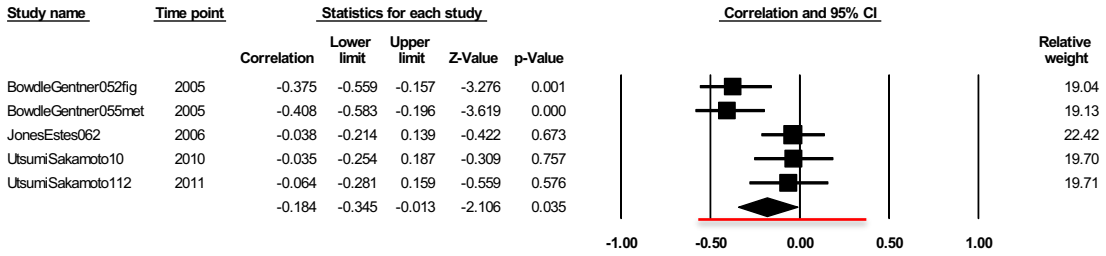

Figure 22. Random-effects model of comprehension latencies with conventionality as a decisive factor

The summary effect size is -0.184 with a 95% confidence interval of [-0.345; -0.013], indicating a small reverse effect of conventionality on comprehension times. The prediction interval is much wider at [-0.653; 0.387]. All effect sizes are below the null-value, but they can easily be divided into two groups: the experiments conducted by Bowdle and Gentner produced a correlation coefficient close to -0.4, while the other three experiments show an effect size only

slightly below 0. The total amount of the observed between-study variance, i.e. the *Q*-statistic (12.621), is significantly different from its expected value of 4, $p = 0.013$. The I^2 value of 68.307 indicates that about 68% of the observed variance is real, and only a third is due to random error. The standard deviation of the true effect sizes is $T = 0.162$. In this case, a subgroup analysis with the authors as a variable seems to be a quite natural choice. See Table 18.

group	**below average**	**above average**
experiments	JonesEstes2006/2 UtsumiSakamoto2010 UtsumiSakamoto2011/2	BowdleGentner2005/2figuratives BowdleGentner2005/2metaphors
summary effect size	-0.045 [-0.162; 0.074]	-0.392 [-0.523; -0.243]
within group variance	0.040 ($p = 0.980$)	0.051 ($p = 0$)
between groups variance	12.530	

Table 18. Two groups of experiments on comprehension latencies with conventionality as a decisive factor

As regards publication bias, we have a too small number of experiments at our disposal to check this.

11.3.3. Comprehensibility ratings

A) Familiarity

Table 19 in Appendix 1 presents the relevant data pertaining to the relationship between comprehensibility ratings and familiarity. Thibodeau et al. (2016, 2018) was – after thorough consideration – dropped from the analyses. The reason for this decision was that the instructions for evaluating the familiarity of metaphors were not formulated clearly enough, as the uncertainty of the researchers also shows: they labelled the same factor 'conventionality' in Thibodeau et al. (2016) but 'familiarity' in Thibodeau et al. (2018).[87]

According to Figure 23, all experiments produced a positive correlation, but there seem to be subgroups.

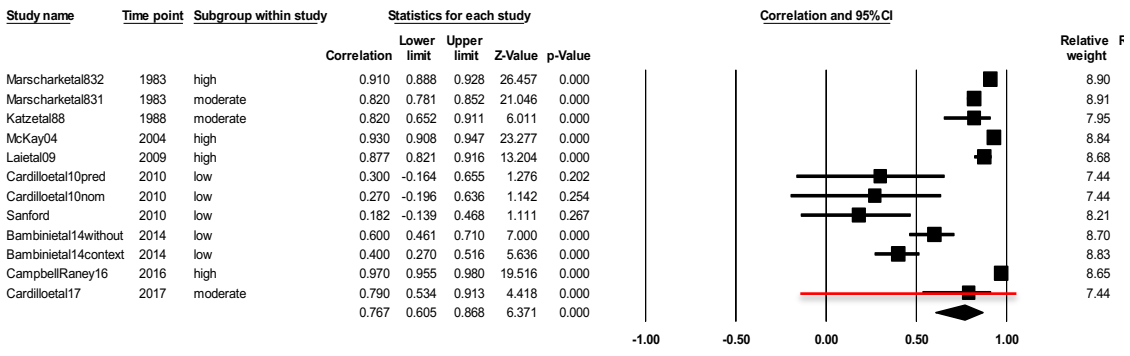

Figure 23. Random-effects model of comprehensibility ratings with familiarity as a decisive factor

[87] The instructions asked participants to focus on the base/vehicle term by capitalizing it; but they had to judge its "conventionality" not in isolation but in a metaphorical context in the same sentence (and in a supporting/not supporting metaphorical/literal wider context). It is also not clear on the basis of a comparison of the instructions and the excerpts, whether this judgement should be made in relation to the base/vehicle term's metaphorical meaning or to the target/topic term.

The summary effect size of 0.767 [0.605; 0.868] means that according to this set of experiments, there is a very strong relationship between familiarity and comprehensibility ratings. As for the prediction interval, it is [-0.217; 0.978] – that is, it allows everything from a small reverse to a large effect for a future experiment. The heterogeneity analysis reinforces our impression that the experiments do not share a common true effect. Namely, the total amount of the observed between-study variance is very high: $Q = 344.861$. This value is significantly different from its expected value: df(Q) = 11. The I^2 value is 96.81; from this we can conclude that practically the whole amount of the observed variance is real and cannot be ascribed to random error. The standard deviation of the true effect sizes is $T = 0.485$.

These findings clearly motivate a subgroup analysis. A possibility is shown in Table 20.

group	below average effect size	average effect size	above average effect size
experiments	Cardillo2010pred Cardillo2010nom Sandford2010 Bambini2014withcontext Bambini2014with-outcontext	Marschark1983/1 Katz et al.1988 Cardillo2017	Marschark1983/2 McKay2004 Lai2009 CampbellRaney2016
summary effect size	0.392 [0.193; 0.561]	0.814 [0.694; 0.890]	0.929 [0.894; 0.953]
within group variance	9.524 ($p = 0.049$)	0.118 ($p = 0.943$)	28.582 ($p < 0.001$)
between groups variance	306.637		

Table 20. Three groups of experiments on comprehensibility ratings with familiarity as a decisive factor

Only the above average group indicates heterogeneity; this results from the very high precision of the estimates of the true (underlying) effect size by these experiments. The three groups are distinct from each other, as the confidence intervals and the between groups variance indicate.

Duval and Tweedie's trim and fill model does not indicate missing experiments, nor any sign of publication bias; Egger's test is non-significant, too ($p = 0.32$). Figure 24 presents the funnel plot, whose asymmetry might result from the heterogeneity we detected.

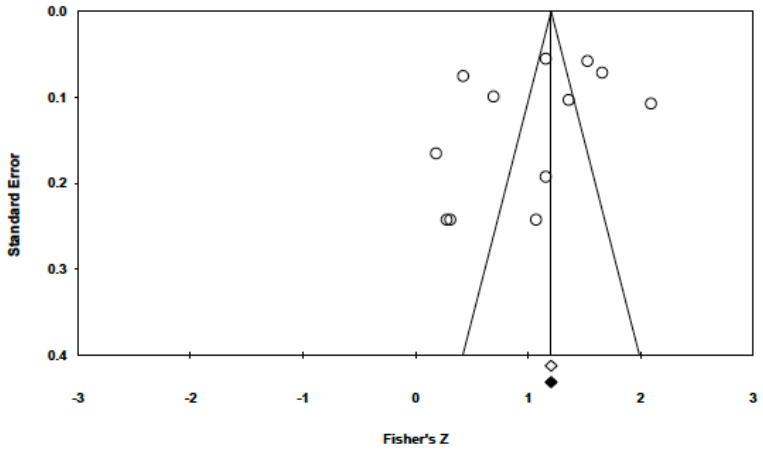

Figure 24. Funnel plot for comprehensibility ratings with familiarity as a decisive factor

B) Aptness

Table 21 in Appendix 1 includes the relevant experimental data on the basis of which the effect of aptness on comprehensibility ratings can be determined.

Figure 25 presents the results of the random-effects analysis.

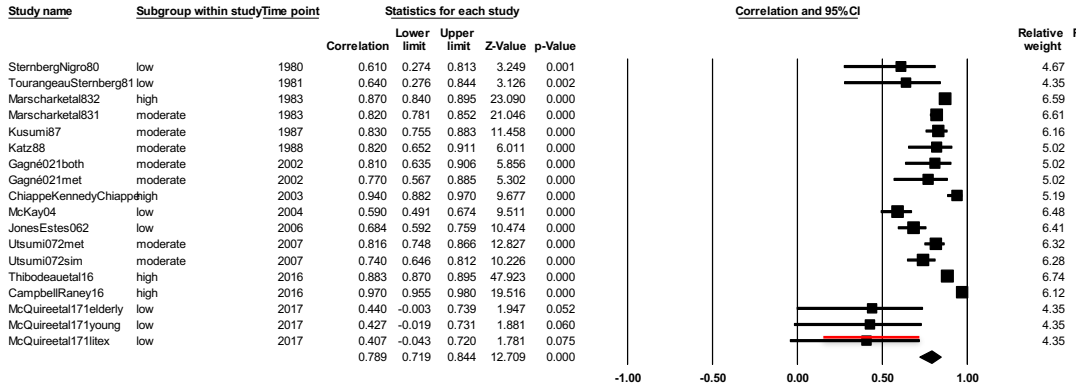

Figure 25. Random-effects model of comprehensibility ratings with aptness as a decisive factor

The most striking feature of this set of experiments is that it contains several experiments which produce a very precise estimation of the effect size, although there are also some experiments which have a quite wide confidence interval. In sum, these experiments provide a very high and very precise summary effect size of 0.789 with a 95% confidence interval as narrow as [0.719; 0.844]. The prediction interval is [0.347; 0.944]. From this we can conclude that a future experiment will yield a moderate to large effect. As for the heterogeneity analysis, the total amount of the observed between-study variance is very high in this case, too: $Q = 244.131$. This value is significantly different from its expected value df(Q) = 17. The I^2 value is 93.037, signalling that the observed variance is not due to random error but is real, i.e. the experiments do not share a common true effect size. The standard deviation of the true effect sizes is $T = 0.312$. These findings clearly motivate a subgroup analysis. We might try to divide up the experiments in such a way that the three experiments by McQuire et al. belong to one group, because they produced an effect size below 0.5, and the other experiments belong to the second group. This grouping is, however, not satisfactory because the second group shows a high amount of heterogeneity. A second attempt might be to classify the experiments into 3 groups. This grouping fares better, yielding three significantly different groups. See Table 22 for the details.

group	below average effect size	average effect size	above average effect size
experiments	SternbergNigro1980 TourengeauSternberg1981 McKay2004 JonesEstes2006/2 McQuire2016/1young McQuire2016/1litexp McQuire2016/1elderly	Marschark1983/1 Kusumi1987 Katz1988 Gagné2002/1both Gagné2002/1met Utsumi2007/2met Utsumi2007/2sim	Marschark1983/2 ChiappeKennedy& Chiappe2003 CampbellRaney2016 Thibodeau2016
summary effect size	0.580 [0.449; 0.687]	0.803 [0.742; 0.852]	0.920 [0.886; 0.944]
within group variance	6.132 ($p = 0.409$)	4.568 ($p = 0.600$)	45.832 ($p < 0.001$)
between groups variance	56.533 ($p < 0.001$)		

Table 22. Three groups of experiments on comprehensibility ratings with aptness as a decisive factor

Finally, we can look for publication bias. Duval and Tweedie's trim and fill model indicates 5 missing studies. See Figure 26.

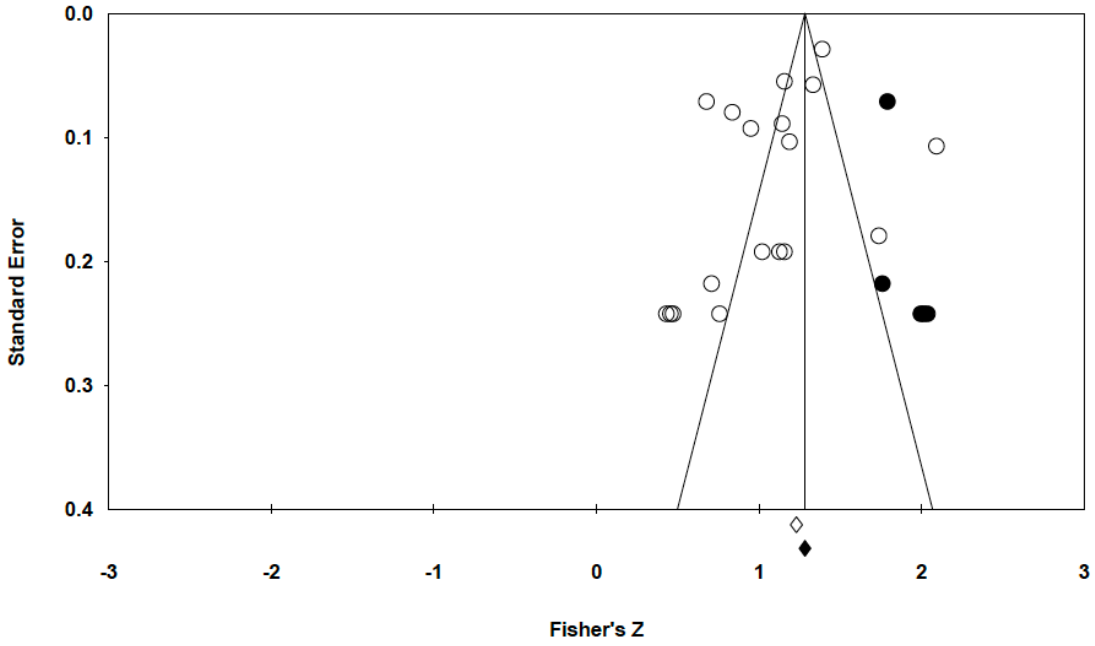

Figure 26. Funnel plot for comprehensibility ratings with aptness as a decisive factor

As in all previous cases, Egger's test is not significant, $p = 0.09$. A cumulative analysis does not provide support for our suspicion that there is publication bias, either, because there is no clear tendency among the cumulative effect sizes. See Figure 27.

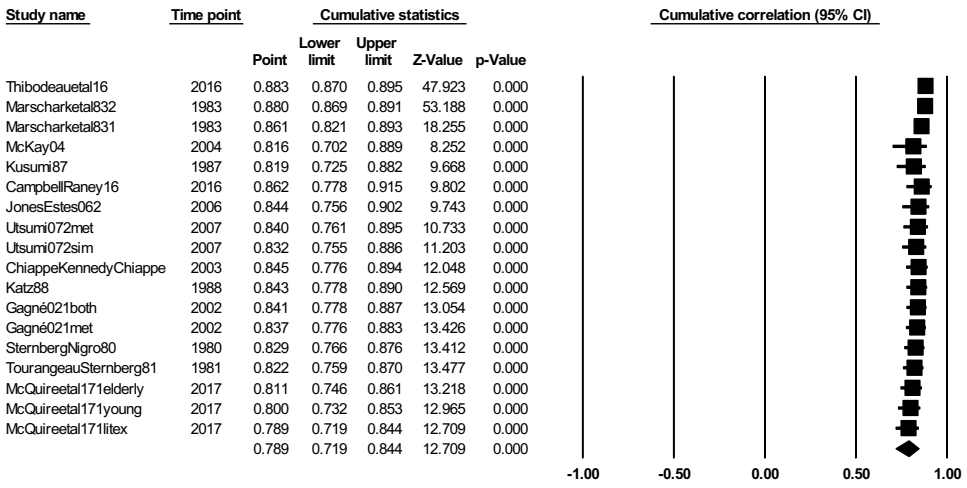

Figure 27. Cumulative meta-analysis for comprehensibility ratings with aptness as a decisive factor

The ambiguity in the tests might result from the circumstance that the heterogeneity was high, which restrains the evaluation of the case.

C) Conventionality

Table 23 in Appendix 1 presents data from experiments investigating the role of conventionality on comprehensibility.

Similarly to our decision in Section 11.3.3A in relation to familiarity, Thibodeau et al. (2016, 2018) was excluded from the analyses. As Figure 28 shows, the results of the experiments are in harmony.

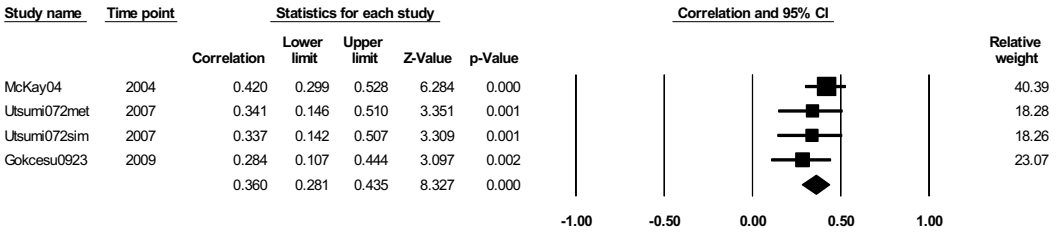

Figure 28. Random-effects model of comprehensibility ratings with conventionality as a decisive factor

The summary effect size of 0.36 with a 95% confidence interval of [0.281; 0.435] indicates a moderately strong relationship between base/vehicle conventionality and comprehensibility ratings with a prediction interval of [0.179; 0.517]. The total amount of the observed between-study variance is very low in this case, $Q = 1.905$, $p = 0.592$, $I^2 = 0$, which means that it is completely due to random error.

11.3.4. Comprehensive analyses

In the interim summaries, we compared the effect of the factors of conventionality, familiarity and aptness on the performance of participants in three distinct rounds, that is, by applying different task types. This perspective can be widened in three directions. First, we can try to generalise these results by co-analysing the outcome of the three rounds conducted in Sections

11.3.1-3 and asking whether a general pattern emerges in the relationship of the three factors with each other.

	conv.	fam.	aptness	average effect size	average I^2	average T
grammatical form preference	0.273	0.393	0.551	0.406	77.338	0.228
comprehension latencies	-0.184	-0.314	-0.269	-0.256	58.213	0.150
comprehensibility ratings	0.36	0.767	0.789	0.639	63.282	0.266
average of the absolute values of the effect sizes	0.272	0.491	0.536			
average I^2	51.986	62.361	84.486			
average T	0.142	0.246	0.256			

Table 24. Comparison of the effect sizes, I^2 and T values related to aptness/conventionality/familiarity in the three groups of experiments

As a check of the columns in Table 24 indicates, the results of the three types of experiments are in harmony and thus *provide converging evidence for the hypothesis that all three factors influence metaphor processing, while conventionality has a weaker effect than aptness and familiarity*. This evidence is considerably stronger than any evidence gained from individual experiments. As the last two rows show, conventionality produced the most consistent results, because the amount and impact of the dispersion in the true effects is the smallest. In contrast, aptness showed a high proportion and large amount of real variance.

The second possible line of analysis is to investigate and compare the data related to the three tasks, i.e., a horizontal analysis of Table 24. Experiments dealing with comprehension times clearly produced the smallest summary effect sizes (in absolute value) and the most consistent results (the smallest amount of real variance and smallest standard deviation of the true effects).

A third promising route could be a deeper and more comprehensive analysis of Tables 9, 11, 13, 16, 18, 20, and 22, that is, separate comparisons of the three factors' behaviour in the three experiment types. For example, we can check whether there is a connection between the researchers and the effect size of the experiments in relation to one particular factor. This was not possible for the analyses we conducted in the subsections 11.3.1-3, because there was not enough data at our disposal. In the tables mentioned above, on the basis of our meta-analyses, we could group the experiments related to an experiment type-relevant factor pair into two or three groups on the basis of their effect sizes. If we assign 1 to experiments belonging to the below average effect size groups, 3 to the experiments in the average effect size groups, as well as to cases when there was no heterogeneity among the experiments, 5 to the above average effect sizes, and 2 to the below average and 4 to the above average experiments in cases in which there were only two groups, and add up the values obtained in the three experiment types for each factor separately, we get the results in Table 25 (see Appendix 1).

As these analyses reveal, there are only a few researchers who have conducted experiments in relation to all three factors:

- Chiappe and Roncero both belong to the below average effect size group with conventionality, and to the above average group with aptness and familiarity.
- Dulcinati is in the average group with conventionality and aptness, and in the below average group with familiarity. This is the most balanced performance.
- McKay was average with conventionality, below average with aptness, and above average with familiarity. This is the most unbalanced performance.
- Utsumi was below average with conventionality and aptness, and above average with familiarity.

As for the two rival theories' point of view regarding the crucial contrast of conventionality vs. aptness, there are three types:

- conventionality in a higher group than aptness: Bowdle & Gentner, McKay;
- aptness in a higher group than conventionality: Chiappe, Roncero;
- conventionality and aptness in the same/similar group: Dulcinati, Jones & Estes, Utsumi.

11.4.4. Interim summary

To sum up, our main results can be summarised as follows:

- With the help of random-effects models, we combined the results of a series of experiments conducted over the past few decades pertaining to the impact of conventionality, familiarity and aptness on metaphor processing in three types of experiments. These analyses yielded considerably more reliable and accurate estimates of the impact of the factors mentioned than single experiments do, because the calculation of the summary effect sizes synthesised the whole range of the available information. Additional analyses also provided information about the precision of these estimates (confidence intervals) and their dispersion (prediction intervals).
  Caveats: The summary effect size is an estimate of the true effect size on the basis of a relatively small set of experiments; in some cases, the number of experiments was very low. Adding further experiments (to be conducted in the future or conducted in the past but unavailable to me and thus not included in these analyses) could modify the results. A further concern is the high amount of heterogeneity we found in the majority of the cases. These problems decrease the reliability of the outcome of our meta-analyses, because they reduce their ability to counter-balance the shortcomings of the individual experiments.
- By performing heterogeneity analyses, we were able to decide whether the results of a set of experiments are consistent or there might be subgroups among them. Our attempts at identifying factors which could distinguish these subgroups, however, were unsuccessful. Therefore, we chose another route: we divided the experiments on the basis of their effect sizes into 2 or 3 subgroups, and then checked with the help of subgroup analysis whether these groups are in fact different from each other.
  Caveats: The small number of experiments as well as the lack of exact replications of the experiments (and via this, the low reliability of the individual experiments) make the groupings questionable, because it is not clear whether the subgroups we identified are stable constructs.

- A simple comparison of the subgroups in relation to the experimental design, the formulation of the task in the control experiments, or the range of metaphors in the stimulus materials pointed to the conclusion that a subgroup analysis based on these factors is pointless. Therefore, it seems that they did not influence the outcome of the experiments.
  Caveats: The concerns we mentioned in relation to the previous point emerge in this case, too.
- In some cases, we found that there is no heterogeneity among the experiments at issue. This makes it possible to resolve the alleged conflict between significant and insignificant results, too. Namely, the relative closeness of the effect sizes and the overlap of their confidence intervals motivate the re-interpretation of the outcome of the experiments as an instance of converging evidence for the summary effect size.
  Caveats: The concerns we mentioned in relation to the previous two points emerge in this case, too.
- Further and deeper theoretical and empirical research should be done in relation to the predictions which can be drawn from theories, as well as the definition and operationalization of the concepts ‘conventionality’, ‘familiarity’ and ‘aptness’. The stimulus materials used in experiments should be revised, too, alongside the experimental designs in order to rule out, for example, boredom effects arising from the huge number of very similar tasks to be performed by participants, or to prevent participants’ naïve theories of metaphor or other conscious considerations from influencing their performance and distorting the results. Our investigations underline several researchers’ concerns that while testing one of the factors, the other two (or even further ones) should be carefully controlled for. The grouping of the experiments we presented in Section 11.3.4 offers an especially promising starting point for a thorough comparative analysis of the experiments conducted so far.
- Our findings prompt the suggestion that statistical meta-analysis should be part of a thorough and radical revision of the methodology in this research field.

12. Conclusions: Inconsistency resolution with the help of cyclic re-evaluation and statistical meta-analysis and possible resolutions of (PPSE)

In Section 8, we raised the Paradox of Problem-Solving Efficacy (PPSE). Now, we are in a position to propose two different resolutions to it.

In Section 10, we applied a meta-scientific model around the concept of 'experimental complex'. From this, we get the first resolution to (PPSE):

Resolution to (PPSE) based on cyclic re-evaluation
Non-exact replications and methodological variants

(a) are *effective tools of problem-solving* if
- they are progressive,
- a limit of the experimental complex can be reached (temporarily, and relative to the informational state), and
- the experimental complex has only one limit, or conflicts with other limits can be resolved;
- conflicts with methodological variants which are limits of other experimental complexes can be resolved;

(b) are *ineffective tools of problem-solving* if
- they are not progressive, or
- the chain of non-exact replications is not capable of reaching a limit of the experimental complex, or
- conflicts among different limits of the same experimental complex cannot be resolved, or
- conflicts with methodological variants which are limits of other experimental complexes cannot be resolved.

(c) The cyclic process of re-evaluation *provides us valuable starting points for the elaboration of new, more refined non-exact replications* which might lead to a limit of the experimental complex (that is, it is an open process); and

(d) is guided by *problem-solving strategies* (Contrastive Strategy, Combinative Strategy).

That is, while progressivity is a *local characteristic* of non-exact replications, effectiveness is a *global feature*. This means that progressivity is relative to an experiment and its non-exact replication, while effectiveness can be judged only relative to an experimental complex. During the re-evaluation process, we are usually "underway" in the sense that we take progressive steps (solve a problem) but we are not in a position to decide about the effectiveness of the re-evaluation process yet. That is, *in most cases we cannot claim that we have reached a limit nor state that no limit can be reached*, etc. but are in the middle of the problem-solving process, where several possible further steps may present themselves.

Nonetheless, there are still two caveats. First, new pieces of information can overrule earlier decisions. Thus, a non-exact replication can turn out to be problematic and lose its limit-status. From this it follows that effectiveness can be judged only in the long run, and decisions are not final but only provisional. Second, a limit of an experimental complex may be incon-

sistent with a limit of another experimental complex. Therefore, *besides intra-complex relations, inter-complex relations have to be reconstructed and evaluated, too*.

The metascientific model we applied supposes that experiments and experimental complexes alike are *open processes* in the sense that, in possession of new pieces of information, they may be continued, modified, or even discarded. Therefore, there are no experiments whose results were unquestionable (both practically and theoretically), nor immune to any improvement, refinement, or criticism. A second key feature of our model is that experimental complexes are supposed to be *not linear but cyclic*. This means that a given step of the re-evaluation process does not necessarily lead to better results.[88] It may turn back to earlier stages and continue the revisions with an experiment for which a non-exact replication has already been conducted. Thirdly, conflicts among experiments cannot be resolved in a simple way, for example, by a mechanical comparison of the plausibility of their results. Instead, strategies of inconsistency resolution as described in Section 10.2 have to be applied.

In Section 11, we took another route and made use of statistical meta-analysis as a tool of conflict resolution. Our considerations pave the way for another possible resolution of the Paradox of Problem-Solving Efficacy (PPSE):

Resolution of (PPSE) based on statistical meta-analysis
Non-exact replications and methodological variants

(a) are *effective tools of problem-solving* if
  – the number of exact and non-exact replications as well as methodological variants is high enough, so that data points processed by meta-analysis are available in a high number, resulting in more stable and well-founded estimates (that is, more plausible data);
  – heterogeneity analyses either indicate consistency among the outcome of the experiments, or there is heterogeneity but the factors which lead to methodological or other kinds of differences, or subgroups among participants which behave differently can be identified;

(b) are *ineffective tools of problem-solving* if
  – the number of exact and non-exact replications as well as and methodological variants is low, or
  – heterogeneity analyses and the resulting subgroupings do not produce useable results but indicate the presence of systematic errors which cannot be identified and ruled out on the basis of the information available. In such cases we have a set of wildly varying results without any plausible explanation for the causes of the divergencies.

(c) The application of statistical meta-analysis
motivates new directions of research insofar as

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

- the use of effect sizes instead of significance testing is more informative about the relationship of the variables at issue. Therefore, it requires more sophisticated explanations, and as a consequence, a refinement of theories;
- the calculation of the summary effect size provides a more reliable and more precise estimation of the strength of the relation between the variables investigated. This necessitates the revision of the empirical basis of theories, and a rethinking of the strength of empirical support for these theories;
- the heterogeneity analyses and the resulting subgroupings might initiate a search for further factors which might influence the results.[89]

(d) *corresponds to the use of the Combinative Strategy* insofar as it synthetises the results of all available experiments.

As with the model based on cyclic re-evaluation, the application of statistical meta-analysis is an open-process, too. Therefore, adding new experiments may lead to different results, and earlier decisions about the effectivity might be in need of revision, too.

[89] Cf. "It would be interesting and vital for further research to examine which of the properties for topic-vehicle pair – interpretive diversity, similarity, aptness, conventionality, relationality – dominates metaphor-simile distinction and how these properties interact with each other." (Utsumi & Kuwabara 2005: 6)

III. THE EVALUATION OF THEORIES WITH RESPECT TO EXPERIMENTAL RESULTS IN COGNITIVE LINGUISTICS

13. Introduction: The paradox of error tolerance (PET) in respect to experiments in cognitive linguistics

When determining the strength of the support a piece of experimental datum provides for a hypothesis/theory, two factors have to be taken into consideration: the plausibility of the datum and the strength of the link between the datum and the hypothesis/theory. This basic idea can be extended in two directions. First, it should be possible to *compare the strength of support* which a datum provides to rival hypotheses/theories. Second, one might want to calculate *the magnitude of the support an experimental complex, or a series of exact and non-exact replications as well as methodological variants* is capable of providing to a hypothesis/theory. We elaborated/used two methods for combining the results of sufficiently similar experiments: a model based on cyclic re-evaluation and statistical meta-analysis. At this point, however, we have to face the *Paradox of Error Tolerance*:

(PET) When determining the strength of support provided by an experimental complex to a hypothesis/theory,
  (a) *the elimination of errors is top priority*, because it is the detection and elimination of problems which makes experiments more reliable data sources;
  (b) *the elimination of errors is not top priority*, because comprehensibility, that is, the involvement of all relevant experiments and the accumulation of all available pieces of information should be ranked higher.

The metascientific model of experimental complexes as presented in Sections 6 and 10 is based on the idea that non-exact replications of experiments are parts of a problem-solving process whose effectiveness depends on the amount and weight of the problems eliminated. In contrast, as mentioned in Section 11.2.1, statistical meta-analysis relies on the assumption that if one collects a large enough number of experiments (which, of course, have to meet certain standards), then shortcomings of individual experiments can be counterbalanced. Therefore, smaller errors can be tolerated. Clearly, the resolution of this conflict is a prerequisite of applying both methods in parallel to the same sets of experiments.

Against this background, we will proceed in Part III of the book as follows. In Section 14, we will present a metascientific tool with the help of which the relationship between single experiments and hypotheses/theories can be modelled and quantified. To this end, we will introduce three different concepts of ‘evidence’. In Section 15, we will show how predictions drawn from theories can be confronted with summary effect sizes obtained with the help of statistical meta-analysis, and how the success of rival theories can be compared. Section 16 aims at elaborating a combined method by integrating the metascientific models based on re-evaluation and problem-solving and on statistical meta-analysis, respectively, and as a result, Section 17 will offer a possible resolution of (PET).

14. The relationship between single experiments and hypotheses/theories: types of evidence

For decades, linguistic data have been viewed as relatively unproblematic entities which provide rich information about linguistic behaviour directly. Evidence was interpreted as a special subset of data, whose certainty is guaranteed by experience ("observation") and intersubjective testability. Thus, evidence was supposed to provide a firm base for the testing of theories and deciding among rival theories. However, this view is untenable for several reasons. First, it has become generally acknowledged in the philosophy of science that experience cannot guarantee the truth of a statement. Second, intersubjective testability does not eliminate the subjectivity immanent in individual experience as a source of knowledge completely. If *n* persons evaluate a phenomenon in the same way, it does not follow that the $n+1^{th}$ person will also agree. That is, the criterion of intersubjectivity is built on induction (that is, on a type of plausible inference), since it infers from a finite number of cases to an infinite number of cases. Third, experiences have to be interpreted: the object of one's experience has to be described with the help of a category system, that is, a theory. Consequently, it is not directly the experience to which one compares the hypotheses of the theory, but its processed and interpreted version.

Therefore, we need a concept of 'data' and 'evidence' which takes into consideration their uncertainty and complexity.

14.1. The p-model's concept of 'data'

The p-model by Kertész & Rákosi does not identify data with "linguistic examples" such as the following Hungarian sentences:

Ennek a tyúknak már megint agymenése van.
['This chick appears to have suffered a brainstorm again.']
Ez a bikapiac sem tart örökké!
['This bull market won't last forever, either!']

Rather, their structure consists of two components: a statement with an information content and a plausibility value supported by a direct source (see Section 4.2.2):

(D) A datum is a statement with a positive plausibility value originating from some direct source.

This means that data are not inferred from other statements but constitute the starting points of a theory (plausible argumentation process). The above "examples" can be transformed into data as follows:

$0 < |$The Hungarian sentences *Ennek a tyúknak már megint agymenése volt.* and *Ez a bikapiac sem tart örökké!* contain conventional metaphors.$|_S < 1$

where *S* is a compound of the linguist's linguistic intuition and linguistic knowledge (i.e., some theory of metaphors).

Accordingly, data are in most cases not claimed to be true with certainty, but they are usually more or less plausible "truth candidates". Their plausibility is usually supported by the sources to some extent, but the sources are not able to make them certainly true. Nevertheless, a datum must possess a certain degree of initial plausibility, that is, it has to receive a plausibility value from some reliable source. Statements which are of neutral plausibility or implausible according to any sources in the p-context, are not data in this sense.

Experimental data can be regarded as data in the sense of (D). Although it is possible to provide a partial reconstruction of the argumentation process leading to the creation of the experimental data, the amount of information which cannot be found in the experimental report and additional materials is too large. Therefore, while the plausibility of experimental data originates from an indirect source from the point of view of the person(s) who conducted the experiment, it is more appropriate to treat them as 'data' in the sense of (D) from the point of view of the reader of the experimental report. Nonetheless, the reliability of their source is strongly influenced by pieces of information pertaining to the components of the experimental process.

14.2. The p-model's concept of 'evidence'

Within the p-model, it is possible to define three types of evidence in order to grasp the relationship between data and other hypotheses of the theory.

*Weak evidence for a hypothesis **H*** simply means that we can build inference(s) on the given datum that make(s) *h* plausible (in the extreme case true with certainty). *Weak evidence against a hypothesis **H*** means a datum on which we can build inference(s) that make(s) *h* implausible (in the extreme case false with certainty):

(EW) (a) A datum *D* is **weak evidence for hypothesis *H***, if the p-context contains statements that extend *D* into an indirect source on the basis of which a positive plausibility value can be assigned to *H*.

(b) A datum *D* is **weak evidence against hypothesis *H***, if the p-context contains statements that extend *D* into an indirect source on the basis of which a positive plausibility value can be assigned to ~*H*.

From this definition it follows that a datum can be weak evidence for a statement and for its rival simultaneously, although it may support them to different extents. The strength of the support it is capable of providing is determined by the peculiarities of the plausible inference connecting the datum and the hypothesis at issue: the plausibility of the datum and the plausibility of the necessary latent background assumptions.[90] That is, the more plausible the datum is and the stronger the link between the datum and the hypothesis at issue is, the stronger is the

[90] On latent background assumptions, see Sections 3.2.7 and 4.2.2.

support this piece of evidence provides to the hypothesis. The strength between the datum and the hypothesis is influenced by the directness of their relationship and the plausibility values of the latent background assumptions. Thus, a great distance between the datum and the hypothesis tested means that a higher amount of latent background assumptions are needed, whose plausibility might be dubious or at least, hard or even impossible to be checked.

Relative evidence for a hypothesis *H* also requires that the datum provides stronger support to *H* than to its rivals:

(ER) (a) A datum *D* is **relative evidence for hypothesis *H***, if
- (i) *D* is weak evidence for hypothesis *H*;
- (ii) the inference(s) connecting the premises and *H* provide(s) *H* with a higher plausibility value than the plausibility values of *H*'s rivals assigned to them by the inferences also using *D* as a premise.

(b) A datum *D* is **relative evidence against hypothesis *H***, if
- (i) *D* is weak evidence against hypothesis *H*;
- (ii) the plausible inference(s) connecting the premises and ~*H* provide(s) ~*H* with a higher plausibility value than the plausibility value of *H* assigned to it by the inferences also using *D* as a premise.

The third type is *strong evidence* which means that the datum makes only hypothesis *H* plausible and does not provide any support to its rivals:

(ES) (a) A datum *D* is **strong evidence for hypothesis *H***, if
- (i) *D* is weak evidence for hypothesis *H*;
- (ii) *D* is not weak evidence for any of *H*'s rivals.

(b) A datum *D* is **strong evidence against hypothesis *H***, if
- (i) *D* is weak evidence against hypothesis *H*;
- (ii) *D* is not weak evidence against any of *H*'s rivals.

Evidence is interpreted by the p-model not as a special subset of data but as a datum with a special function *relative to some hypothesis of the theory*. From this it follows that evidence is not objective, immediately given, theory-independent and completely reliable but source- and theory-dependent and reliable only to a certain extent. Further, data which meet the criteria laid down in (EW)-(ES) in most cases do not perfectly support or refute the given hypothesis. The connection between the datum and the hypothesis is established by plausible inferences relying on plausible premises. A third important corollary of these definitions is that the function of evidence is not restricted to the testing of hypotheses, that is, to the justification of theories, but data and evidence play a role in every stage of the process of linguistic theorising.

15. Summary effect sizes as evidence

While the application of the concepts of 'weak/relative/strong evidence' seems to be quite straightforward with single experiments, it is less clear how to confront the results of statistical meta-analyses with predictions drawn from theories. This section will be devoted to this issue and the question will be answered with the help of the third part of Case Study 5.

15.1. Case study 5, Part 3: Comparing predictions with summary effect sizes

15.1.1. Two models of metaphor processing and their predictions

According to Glucksberg's *Interactive Property Attribution Model* (IPAM), all metaphors are processed in the same way: the target/topic concept is interpreted as belonging to an abstract, unnamed (in certain cases *ad hoc*) metaphoric category prototypically represented by the base/vehicle term. In this way, the base/vehicle term will have dual reference: a literal reference and an abstract one. The base/vehicle and the target/topic concepts do not become connected to each other directly. Rather, they play different but interacting roles. First, from the literal meaning of the base/vehicle term, abstract metaphorical categories are created. Then, an attempt will be made to apply these abstract metaphorical categories to the (literal) target/topic concept. They provide salient properties which characterize both the base/vehicle concept and all other concepts falling within its metaphorical category and which might be attributable to the target/topic. Simultaneously, the target/topic suggests "dimensions", that is, "provide[s] information about what types of properties they can meaningfully inherit and therefore about what types of categories they can meaningfully belong to" (Bowdle & Gentner 2005: 195). As a result, the interplay between the base/vehicle and target/topic determines which of the property-candidates provided by the base/vehicle are relevant and do, in fact, become assigned to the target/topic. This also means that metaphor processing is asymmetrical from the outset.

Table 26 provides a schematic overview of the main steps of IPAM.[91]

	Interactive Property Attribution Model	
Steps 1-2	suggesting (and if needed, elaborating) one or more superordinate abstract metaphorical categories exemplified by the literal base/vehicle concept	these two steps run in parallel
	suggesting "dimensions" of the target/topic concept	
Step 3	determining the salient properties for each metaphorical category candidate	these properties characterize both the literal base/vehicle concept and the metaphorical category
Step 4	projecting the candidate salient properties from the metaphorical categories to the target/topic concept	
Step 5	selecting the most successful projection	
Step 6	attributing the relevant properties to the literal target/topic	

Table 26. Main steps of metaphor processing according to Glucksberg's IPAM

[91] For further details, see Glucksberg et al. (1997).

According to Jones and Estes, this model yields the following general prediction:

> "It follows from this view of metaphor as the interaction of topic and vehicle […] that conventionality should not be a primary predictor of metaphor comprehension, because conventionality refers to only the vehicle concept […]. Instead, aptness should predict comprehension, because aptness reflects both the salience of the vehicle property and its relevance to the topic." (Jones & Estes 2006: 20)

In contrast, *Gentner's Career of Metaphor Hypothesis* (CMH) states that metaphors are analogies. More precisely, metaphor processing is modelled as a two-stage process. First, the two, partially isomorphic concepts of the base/vehicle and target/topic are systematically aligned in such a way that as many connections as possible are established between elements of the two representations, and also their relations; then, these local connections are systematised to one or a few global interpretations. As a second step, if a structurally consistent alignment is achieved, candidate inferences are drawn. That is, further elements are projected from the base/vehicle to the target/topic. Thus, the early stages of metaphor processing are symmetrical; only the projection of candidate inferences is an asymmetrical process. From this it follows that "metaphoric categories are derived from the common relational structure of the target and base concepts and not from the base concept alone" (Bowdle & Gentner 2005: 198). Nonetheless, frequently occurring base/vehicle terms may "become polysemous and […] automatically elicit a metaphoric category" (Bowdle & Gentner 2005: 198). This means that in the case of novel metaphors, there is a *horizontal alignment* between the literal senses of the base/vehicle and the target/topic terms. With conventional metaphors, in contrast, there is a shortcut: besides a horizontal alignment, a *vertical alignment* will also be created between the literal meaning of the target/topic term and the abstract, secondary meaning of the base/vehicle term. In such cases, the metaphorical meaning does not have to be elaborated *ad hoc* because it has become more salient and easily available during repeated encounters. To sum up, novel metaphors are processed as comparisons, while conventional metaphors as categorizations.[92] For later references, Table 27 summarizes the main steps of the metaphor comprehension process as modelled by Gentner's CMH.

[92] For more details, see Gentner & Bowdle (2008).

	Career of Metaphor Hypothesis	
Step 1	matching the identical elements of the literal base/vehicle and target/topic concepts locally	systematic structural alignment of the partially isomorphic concepts of base/vehicle and target/topic (symmetrical phase)
Step 2	combining the matches to structurally consistent connected clusters	
Step 3	combining the clusters to one or a few maximally consistent global system(s), i.e. interpretation(s); structural evaluation of the systems	
Step 4	drawing candidate inferences from the literal base/vehicle to the target/topic concept (projection of further elements)	deriving metaphorical categories (asymmetrical phase)
Step 5	creating a horizontal alignment between the literal senses of base/vehicle and target/topic (mapping between two representations on the same level of abstraction), and, with conventional bases/vehicles, also a vertical alignment between the secondary, abstract meaning of the base/vehicle term and the target/topic (mapping between representations at different levels of abstraction)	
Step 6	the quickest and most systematic alignment wins	

Table 27. Main steps of metaphor processing according to Gentner's CMH

On the basis of Gentner's CMH, the following set of predictions has been set forth:

> "[…] only the simile form directly invites comparison. The metaphor form initially invites an inappropriate comprehension strategy – it invites searching for a category that does not exist. Novel metaphors must therefore be reinterpreted, which should add to the processing time." […] "The metaphor form invites categorization and will therefore promote a relatively simple alignment between the target and the abstract metaphoric category named by the base." (Bowdle & Gentner 2005: 202)
> "[…] if conventionalization increases the likelihood of categorization processing, then (averaging across grammatical forms) conventional figuratives should be easier to interpret than novel figuratives." (Bowdle & Gentner 2005: 202)

As we have seen in Section 11.1.1, several researchers have proposed the use of the factor 'familiarity' instead of 'conventionality', because the former takes into consideration both members of the metaphor. The relationship between the two rival theories and the 'familiarity' factor is, however, not clarified in the literature.

Jones & Estes (2006) tried to produce an overview of relevant research and confront the predictions drawn from the two rival theories with the experimental evidence available. The predictions the experiments made use of pertain to the following tasks: grammatical form preference, comprehension latencies, comprehensibility ratings, and metaphorical categorisation. They found that there are several experiments which indicate that aptness is a decisive factor, while other experiments provided evidence that it is conventionality which is relevant. Besides the highly inconclusive results, Jones and Estes see another reason to urge a new beginning. Namely, they raise the objection that in the great majority of these experiments, either aptness or conventionality was controlled for, but the other factor was not. From this they concluded that it is possible that their impact was confused.

15.1.2. Re-evaluation of the predictions of Gentner's CMH and Glucksberg's IPAM

In Section 11.1.3, we re-evaluated the explication and operationalization of the concepts of 'conventionality', 'familiarity' and 'aptness'. This necessitates a revision of the predictions related to them, too. This subsection will be devoted to this task. We will continuously refer back to Section 15.1.1 and most importantly, to Tables 26 and 27.

A) Grammatical form preference

If we accept the hypothesis that "form reflects function in figurative language" (Bowdle & Gentner 2005: 200), then the metaphor form should be preferred if the figurative statement is processed as a kind of categorization, and the simile form should be chosen whenever the expression is interpreted as a comparison.

For CMH this means that conventionality should be a decisive factor. The pivotal point is at Steps 4 and 5 in Table 27. Namely, with high base/vehicle conventionality, the projection of candidate inferences and the vertical alignment process should be quick and effortless due to an easy access to the secondary figurative meaning of the base/vehicle, resulting in the preference of the metaphor form. In contrast, low conventionality should – in the absence of a secondary abstract meaning – mostly lead to the default horizontal alignment of the literal base/vehicle and target/topic concepts, that is to say, to a comparison. Nevertheless, low conventionality coupled with high familiarity and/or high aptness might result in a vertical alignment process, too. Therefore, the impact of conventionality might be somewhat weaker than would be the case if there were a linear relationship between conventionality and the preference of the vertical processing mode.

In Section 11.1.3, we defined aptness as the proportion of salient features of the base/vehicle which are regarded as important and relevant to the characterisation of the target/topic and those which are deemed to be inapplicable. A high ratio of applicable salient properties could facilitate Steps 4 and 5 in certain cases. The reason for this is that if the majority of the salient features are applicable to the target/topic concept, then the metaphorical category arising does not necessarily arise from a horizontal alignment, i.e. from a lengthy interplay between the literal base/vehicle and target/topic concepts in which the irrelevant features of the two concepts are gradually filtered out and their commonalities are revealed. Instead, the relevant elements originate predominantly from the base/vehicle concept. The decisive point is whether and to what extent the salient features of the base/vehicle are prototypical, higher-level features, which are applicable to a series of other concepts, as well. Thus, although a secondary figurative meaning of the base/vehicle is not given (as with conventional bases/vehicles), it can be emergent, i.e. newly elaborated – initiating a vertical alignment process. Nonetheless, high aptness also aids and speeds up the horizontal alignment of the base/vehicle and target/topic concepts. Thus, the impact of aptness might be palpable but clearly weaker than that of conventionality or familiarity if metaphors are processed as described by CMH.

Familiarity should be highly influential (provided other factors are counterbalanced) in the case of both rival models. Repeated encounters with the metaphor form might have primed the use of this form and make Steps 1-6 of both models run smoothly. Thus, a switch to the simile form in the case of high familiar metaphors seems to be unlikely.

According to IPAM, all metaphors are processed in a similar manner, i.e. as categorizations. Thus, whether a figurative statement can/will be interpreted as categorization, depends

on whether and how easily the 6 steps of metaphor processing in Table 26 can be executed. Aptness should be pivotal, mostly because high aptness strongly boosts the crucial Steps 4-6. Moreover, if the salient properties of the base/vehicle which are relevant to the target/topic are prototypical features (which is a strong possibility with high apt metaphors), then Steps 1-3 should be fluent, too. High conventionality of the base/vehicle might favour the choice of the metaphor form, too. Namely, if a base/vehicle possesses a well-known abstract secondary meaning, a super-ordinate metaphorical category and its salient properties should be more easily identified (Steps 1 and 3). Nevertheless, conventionality should be less effective than aptness or familiarity, because the target/topic is not necessarily rich in dimensions even if the base/vehicle is highly conventional. Thus, high base/vehicle conventionality leaves Steps 2 and 4 intact. As a consequence, the successful accomplishment of Step 6 (and via this, the choice of the metaphor form) also depends on a factor which is unrelated to base/vehicle conventionality.

To sum up, these considerations yield the predictions in Table 28.

	conventionality	**familiarity**	**aptness**
CMH	large effect	large effect	small effect
IPAM	small effect	large effect	large effect

Table 28. Predictions of the two rival theories in relation to grammatical form preference

B) Comprehension latencies

High base/vehicle conventionality triggers an effortless execution of Step 4 of CMH as well as the vertical alignment of base/vehicle and target/topic in Step 5, because identifying the elements to be projected to the target/topic and drawing candidate inferences should be an easy task. In addition, since the vertical alignment wins over the horizontal alignment with conventional bases/vehicles, there can be a harmony between grammatical form and processing mode with conventional metaphors, but not with novel metaphors. Due to these two factors, conventional metaphors should be processed markedly more quickly than novel ones. Familiarity should substantially shorten the processing times, too. To be more precise, high familiarity increases the effect of high base/vehicle conventionality, because it also speeds up Steps 1-3. Similarly, high familiarity combined with low base/vehicle conventionality should be quicker than low familiarity plus low conventionality, because acquaintance with both members of a metaphorical expression could accelerate Steps 1-3, as well as the horizontal alignment process in Step 5. This means that familiarity might be an even more effective factor than base/vehicle conventionality, because it aids all stages of metaphor processing. Aptness should be moderately beneficial to the processing times. Namely, as we have seen in A), high aptness may, in certain cases, initiate a vertical alignment between base/vehicle and target/topic, creating harmony between processing mode and linguistic form. Nonetheless, high base/vehicle conventionality is more successful in ensuring that the vertical alignment mode can be applied and will overrun a horizontal mode than aptness. Since a large proportion of applicable relevant features aids the production and projection of candidate inferences (Steps 4-5), aptness should have some impact on the speed of the horizontal alignment process, as well.

In the case of IPAM, high aptness should markedly accelerate the identification and projection of the relevant salient properties in Steps 4-6. Steps 1-3 might also be boosted, supposing that the larger number of the salient properties of the base/vehicle which are relevant to the target/topic are prototypical features in relation to the category at issue. As for the impact of base/vehicle conventionality on the comprehension times, high conventionality speeds up Steps 1 and 3 but does not influence Step 2, and, as a consequence, its effect on Step 4 is equivocal. Therefore, its impact is clearly weaker than that of aptness. Familiarity could be the most effective factor in this case, too, because it facilitates each step of metaphor processing.

Table 29 summarises the predictions of CMH and IPAM pertaining to comprehension latencies.

	conventionality	**familiarity**	**aptness**
CMH	moderate effect	large effect	small effect
IPAM	small effect	large effect	moderate effect

Table 29. Predictions of the two rival theories in relation to comprehension latencies

C) Comprehensibility ratings

Comprehensibility ratings mirror processing more indirectly than comprehension times, since they show the *felt* effortlessness of the interpretation of metaphorical expressions. That is, comprehensibility ratings might also be influenced by conscious decisions, and there is room for participants' naïve theories in relation to metaphors. This means that although comprehensibility ratings should produce similar effects to comprehension latencies, they might be confounded by subjective factors. For instance, participants might judge high familiar, conventional and high apt metaphors to be considerably easier to interpret than unfamiliar, novel or low apt metaphors on the basis of conscious considerations. Therefore, the related effect sizes might be systematically overestimated. See Table 30.

	conventionality	**familiarity**	**aptness**
CMH	large effect	large effect	moderate effect
IPAM	moderate effect	large effect	large effect

Table 30. Predictions of the two rival theories in relation to comprehensibility ratings

With correlation coefficients, Cohen proposes the following limits for the social sciences: 0.1 (small effect), 0.3 (moderate) and 0.5 (large), respectively.

15.1.3. Comparison of the accuracy of the predictions

A) Grammatical form preference

A comparison of the results of meta-analysis and the predictions of the two rival theories yields the following picture. See Table 31.

	conventionality	familiarity	aptness
CMH	large effect	large effect	small effect
IPAM	**small** effect	large effect	**large** effect
meta-analysis	0.273 [0.127; 0.408]	0.393 [0.215; 0.546]	0.551 [0.424; 0.658]

Table 31. Comparison of the results of meta-analysis and the predictions – grammatical form preference

As the bold emphases in Table 31 show, the predictions of Glucksberg's IPAM are considerably closer to the results of the meta-analysis we conducted than Gentner's CMH, because conventionality produced the smallest effect and aptness the largest. This means that IPAM also predicted the relative magnitude of these effects correctly, while CMH's prediction turned out to be wrong because it stated that conventionality should influence grammatical form preference more than aptness. Nonetheless, the great amount of heterogeneity found in each set of experiments is concerning.

B) Comprehension latencies

As Table 32 shows, the difference between Glucksberg's IPAM and Gentner's CMH is trivial. The effects were considerably weaker than predicted by both models, and, most importantly, both rivals seem to have overestimated the impact of the factor which should be crucial according to their theory.

	conventionality	familiarity	aptness
CMH	moderate effect	large effect	**small** effect
IPAM	**small** effect	large effect	moderate effect
meta-analysis	-0.184 [-0.345; -0.013]	-0.314 [-0.388; -0.237]	-0.269 [-0.408; -0.117]

Table 32. Comparison of the results of meta-analysis and the predictions – comprehension latencies

C) Comprehensibility ratings

The predictions drawn from Glucksberg's IPAM are in harmony with the outcome of the meta-analysis. In contrast, Gentner's CMH only prognosticated the effect of the familiarity factor correctly. See Table 33.

	conventionality	familiarity	aptness
CMH	large effect	**large** effect	moderate effect
IPAM	**moderate** effect	**large** effect	**large** effect
meta-analysis	0.36 [0.281; 0.435]	0.767 [0.605; 0.866]	0.789 [0.719; 0.844]

Table 33. Comparison of the results of meta-analysis and the predictions – comprehensibility ratings

If we try to summarise our results, we have to face the problem that the outcomes of the three meta-analyses are not in accord.

15.1.4. Interim summary

First, we revised the predictions which can be drawn from Gentner's Career of Metaphor Hypothesis and from Glucksberg's Interactive Property Attribution Model. To this end, we re-

interpreted the three concepts ‘conventionality’, ‘familiarity’ and ‘aptness’, and revised their operationalization, as well as their relationship with the two models of metaphor processing at issue. In a second step, we confronted these predictions with the results of our meta-analyses. Since the latter provide a substantially more reliable and accurate estimation of the true effect size than the individual experiments, a decision between the two rival theories reached by the involvement of meta-analytical tools should be more well-founded, too.

Nonetheless, there are some caveats, as well as problematic points which need clarification. First, drawing predictions from the two theories is a complicated task and our re-evaluations presented in Section 15.1.2 are also still open to discussion. Second, as already mentioned in the particular analyses, the large amount and proportion of the real variance in the effect sizes is a serious concern. Third, as we have seen in Section 15.1.3, there was no harmony between the outcome of the three experiment types: while grammatical form preferences and comprehensibility ratings produced similar results, comprehension latencies produced the smallest difference between the effect of the three factors of conventionality, familiarity and aptness, partially contradicting the predictions of both rival theories. This finding requires caution because experiments investigating comprehension latencies showed the smallest amount and proportion of variance, and it is also the experiment type which is capable of eliminating participants’ conscious decisions to the greatest extent. Therefore, it seems to be advisable to deal with the three experiment types separately. Fourth, it is not clear how to compare the predictions and the summary effect sizes. One might come to the idea that if a theory yields the prediction saying that there should be an effect but there is a reverse effect, then the statement about the summary effect size provides evidence *against* this theory; in contrast, in all other cases – that is, if according to the meta-analysis, there is no effect at all or there is some greater or smaller effect –, the summary effect size provides evidence *for* the theory. The situation, however, is more complicated if we give up black-and-white thinking, and realize that the distance between a weak effect and a weak reverse effect might be the same as between a small effect and a large effect. Further difficulties arise from situations in which multiple predictions drawn from a theory have to be evaluated, or the predictions drawn from two or more rival theories have to be compared. Another idea might be to apply a 6-point scale of ‘reverse large – reverse moderate – reverse small – no effect – small – moderate – large effect’. In this case, the most clear-cut solution could be to require that the prediction and the outcome of the statistical meta-analysis have to fall into the same category. Nonetheless, a summary effect size of -0.269 should be interpreted according to Cohen’s proposal as a weak reverse effect, although it is quite close to the lower limit of reverse moderate effects, that is, to -0.3. Thus, if the prediction was made that there should be a reverse moderate effect, then this result should be interpreted as weak evidence *against* the theory. This would be clearly counter-intuitive. In the next section, we will try to elaborate a more acceptable solution.

15.2. Deciding between theories on the basis of summary effect sizes

As we have seen in Sections 11 and 15.1, statistical meta-analysis yields more fine-grained results than the customary practice of hypothesis testing. While the latter provides a dichotomy of significant vs. non-significant results, the former produces effect sizes. Therefore, in order

to confront and compare predictions drawn from rival theories with results of statistical meta-analyses, more refined predictions are needed, too. The application of the concepts of 'weak/ relative/strong evidence' has to be re-thought, too. More precisely, only the concept of weak evidence has to be adjusted; relative and strong evidence can be applied as laid down in (ER) and (ES).

Let's start with the concept of 'weak evidence' as defined in (EW) in Section 13.2. Suppose that the meta-analysis we conducted is a reliable data source. This yields a plausible inference with the following structure:

(EWM) $0 < |$If theory T correctly describes metaphor processing, then the effect of factor F on the processing times of metaphors should be $x.|_R < 1$
$0 < |$The effect of factor F on the processing times of metaphors is $y.|_{meta\text{-}analysis} < 1$
$0 < |x$ and y are in harmony.$|_R < 1$

$0 < |$Theory T correctly describes metaphor processing.$|_I < 1$

The first premise of (EWM) presents a prediction drawn from theory T by researcher R. As we have seen in Section 15.1, this step involves many uncertainties; therefore, this statement is not true with certainty but only plausible. Similarly, as our analyses in Section 11.3 exemplify, the second premise presenting a datum, namely, the summary effect size of a statistical meta-analysis can be only a plausible statement, too, since the synthesis of even a huge number of experiments is not capable of balancing out all possible systematic errors which may burden the single experiments. As the third premise captures, the conclusion can be regarded as plausible on the basis of this inference as an indirect source if the prediction and the result of the meta-analysis are in harmony. y is a real number between -1 and 1. Predictions, however, are in most cases considerably less informative than the summary effect size and often stipulate solely the presence of an effect without characterising its size. This yields a 3-point scale of 'there is an effect – there is no effect – there is a reverse effect'. In other cases, x is a category from the 6-point scale 'reverse large – reverse moderate – reverse small – no effect – small – moderate – large effect'. Thus, the critical point of the evaluation of such inferences is the comparison of a rather rudimentary prediction and a precise summary effect size. Accordingly, the following *rule of thumb* presents itself as a first possible explanation of the "harmony" required by the third premise of (EWM):

(RTR) In cases in which the prediction P drawn from a theory T cannot be refined and it stipulates only the presence and direction of the effect, a statistical meta-analysis provides *weak evidence for P* if the summary effect size indicates the presence and the direction of the effect (irrespective of its magnitude) as P did.
In all other cases, the summary effect size is a datum which has to be interpreted as *weak evidence against P*.

(RTR) yields only a quite rudimentary evaluation. For example, summary effect sizes of 0.2 and 0.9 may both fulfil the requirements and count as weak evidence for the prediction that there should be an effect. In the case of the experiments in Case Study 5, we obtain that the

statistical meta-analyses we conducted provide *weak evidence for the predictions of both theories* if we apply (RTR) and compare the predictions and the summary effect sizes separately. This interpretation of the results is, however, not satisfactory, since it does not take into consideration the relative strength of the effects, that is, whether the predictions the two rival theories yield also stipulate which factor should be stronger.

If we are in possession of predictions which specify an effect size at least on the scale 'reverse large – reverse moderate – reverse small – no effect – small – moderate – large effect', then the following rule of thumb might enable us to apply the definition (EW) to summary effect size data:

(RTS) A statistical meta-analysis provides *weak evidence for* the predictions of theory T if the predictions and the related summary effect sizes indicate similar effect sizes, that is, both show a small/moderate/large (reverse) effect or no effect.
In all other cases, the summary effect size is a datum which has to be interpreted as *weak evidence against* the predictions of the theory at issue.

With the help of (RTS), the results presented in the previous section can be evaluated as follows. The statistical meta-analyses carried out yield in the case of comprehensibility ratings *strong evidence for the predictions from IPAM, and via this, for the model of metaphor processing delineated in Table 26*; in contrast, with grammatical form preference and comprehensibility latencies the summary effect sizes provide *weak evidence against both rival theories*, since none of them was capable of predicting the effect of all three factors correctly. This evaluation of the results is, however, strongly counter-intuitive, since IPAM fared, as Table 31 shows, considerably better with grammatical form preference ratings than did CMH.

Indeed, a less strict version of (RTS) is also possible, which formulates looser stipulations of weak counter-evidence, and establishes a "neutral zone" between evidence and counter-evidence:

(RTL) A statistical meta-analysis provides *weak evidence for* the predictions of a theory if the predictions and the related summary effect sizes indicate similar effect sizes, that is, if both show a (reverse) small/moderate/large effect. In contrast, statistical meta-analysis provides *weak evidence against* the predictions of a theory if the difference between the predictions and the summary effect sizes is at least two levels on the 'large reverse effect – moderate reverse effect – small reverse effect – no effect – small effect – moderate effect – large effect' scale. If the difference is only one level, then the prediction has *neutral plausibility* on the basis of the summary effect size as evidence, i.e., the datum at issue is indecisive.

These looser guidelines yield that the summary effect size data in the case of the comprehension latencies and comprehensibility ratings provide weak evidence for both theories, since the distance between their predictions and the summary effect sizes never transgresses the 1-level mark. Since the predictions of IPAM are more precise (that is, they often correctly predicted the size of the effect), the summary effect size data provide *relative evidence* for the predictions of Glucksberg's theory. As for grammatical form preferences, the summary effect sizes provide

strong evidence for the predictions of IPAM, and via them, to the model presented in Table 26, since they were in two cases correct and in one case there was only 1-level difference; in a sharp contrast, CMH was in a 2-level error with two factors and in a 1-level error in one case.

There is indeed a fourth possibility: we may focus on the relationship of the factors of 'conventionality' and 'aptness' and interpret the predictions of the rivals not in separation but in such a way that the key predictions have to be fulfilled, that is, either conventionality or aptness has to be clearly stronger, while familiarity has to be similar to aptness or slightly weaker or stronger. At a more general level this yields the following rule of thumb for complex (more-factor) predictions:

(RTR) A statistical meta-analysis provides *weak evidence for* the predictions of a theory if the predictions estimate the relative strength of the relevant factors correctly.

If we build our evaluation on (RTR), then we obtain that grammatical form preference and comprehensibility ratings data provide *strong evidence for (IPAM)*, while comprehensibility ratings provide *weak evidence against both rivals*, since according to them, there should be a big difference between conventionality and aptness, while there was no substantial difference.

Of course, it is also possible to elaborate quantitative methods for the comparison of predictions and summary effect sizes. This is, however, a quite complicated task.

To sum up, the best way to interpret the upshot of our analyses is that the statistical meta-analyses we conducted *make the IPAM predictions (and via this, the related hypotheses of the theory) moderately plausible, while they show the CMH predictions and related hypotheses slightly implausible*. This means that a decision between the two theories based on our results is unequivocal but cannot be final and is fallible. Nonetheless, they suggest more clearly and strongly than any single experiment could that IPAM should be preferred over CMH – at least, on the basis of the totality of experiments conducted so far. New experiments making use of revised experimental designs, however, may overwrite this decision in future.

16. The combined method

Now we are in possession of the metascientific tools with the help of which the relationship between (rival) theories and data can be described. Nonetheless, one question is still open: it is not clear whether the metascientific models based on cyclic re-evaluation and problem-solving on the one hand and on statistical meta-analysis on the other hand, are compatible with each other. In order to elaborate a possible resolution of the Paradox of Error Tolerance (PET) as presented in Section 13, we will analyse an experimental complex with both methods, and generalise the moral of this case study.

16.1. Case study 6, Part 1: Cyclic re-evaluation of a debate on the role of metaphors on thinking

Thibodeau and Boroditsky presented the results of two series of experiments (Thibodeau & Boroditsky 2011, 2013) in favour of the hypothesis that "exposure to even a single metaphor can induce substantial differences in opinion about how to solve social problems" (Thibodeau & Boroditsky 2011: 1). Steen and his fellow researchers, however, "consistently found no effects of metaphorical frames on policy preference" (Steen et al. 2014: 21) when they conducted follow-up experiments in order to replicate Thibodeau and Boroditsky's results. Thibodeau and Boroditsky, in response, re-analysed their own, as well as Steen et al.'s earlier experiments and conducted new ones, too. They reported results that reinforced their earlier findings (Thibodeau & Boroditsky 2015). As they put it, this controversy has led to a highly productive proliferation of non-exact replications and control experiments:

> "[…] this example highlights the importance of thinking about replication not in terms of individual studies, but in terms of lines of investigation. Often the interpretation of the results of any one experiment depends on many other ancillary pieces of data, norming results, and control conditions reported elsewhere in the same paper or in the same line [of – Cs. R.] work more broadly. […] Arriving at a meaningful culture of replication will require going beyond a focus on direct replication of disconnected single studies, and instead shifting to a theoretically-informed consideration of the broader set of dependencies needed for interpreting any given finding." (Thibodeau & Boroditsky 2015: 20f.)

Nevertheless, the debate continued in a further publication (Reijnierse et al. 2015). Both the evaluation of the earlier results and the conclusions drawn from the newer series of experiments diverge to an even greater extent. Thus, it is not clear whether the more elaborated experiments reinforce the results of the earlier set of experiments or must be regarded as overruling them. Therefore, these series of replication attempts seem to be typical examples of the *Paradox of Problem-Solving Efficacy* (cf. Sections 8 and 10.4). That is, on the one hand, each replication is a more refined version of the original experiment and their predecessors, *providing more plausible experimental data*. On the other hand, however, instead of leading to converging results, they *trigger cumulative contradictions* among different replications of the original experiment. The question is how the cumulative contradictions between Thibodeau & Boroditsky (2011, 2013, 2015), Steen et al. (2014) and Reijnierse (2015) can be resolved.

In this section, we will apply the metascientific model presented in Section 6 and reformulate the problem at issue as the question of what the limit of the experimental complex evolving from the set of experiments presented in Thibodeau & Boroditsky (2011) might be. In order to answer this question, we have to check the progressivity and the effectiveness of the non-exact replications of the original experiment in Thibodeau & Boroditsky (2011). It is possible that an experimental complex has not reached a limit so far, or it may also have more than one limit at the same time. If there is no limit, the question arises of whether and how a limit could be reached. If there are two limits, then we might face an unresolvable contradiction (at least, on the basis of the information at our disposal).

In Subsection 16.1.1, the structure of the experimental complex evolving from Thibodeau & Boroditsky (2011) will be reconstructed and the progressivity of the non-exact replications judged. This means that this section will focus on the relationship among the experiments: the analyses will try to reveal whether there is at least one problem which is unsolved by an experiment but is solved by its successor, and choose the most elaborated version (limit-candidate) within experiments belonging to the same paper. Subsection 16.1.2 intends to provide an evaluation of the effectiveness of the problem-solving process and identify remaining unsolved problems which burden the limit-candidate. Subsection 16.1.3 will re-evaluate the whole problem-solving process and reveal future prospects.

16.1.1. Reconstruction of the structure of the experimental complex and the progressivity of the non-exact replications

A) The original experiment

OE (Thibodeau & Boroditsky 2011, Experiment 1): Participants were presented with one version of the following passage:

> "Crime is a **{wild beast preying on/virus infecting}** the city of Addison. The crime rate in the once peaceful city has steadily increased over the past three years. In fact, these days it seems that crime is **{lurking in/plaguing}** every neighborhood. In 2004, 46,177 crimes were reported compared to more than 55,000 reported in 2007. The rise in violent crime is particularly alarming. In 2004, there were 330 murders in the city, in 2007, there were over 500."

Then, they had to answer the open question of what, in their opinion, Addison needs to do to reduce crime. The answers were coded into two categories on the basis of the results of a previous norming study: 1) diagnose/treat/inoculate – that is, they suggested social reforms or revealing the causes of the problems, and 2) capture/enforce/punish – that is, they proposed the use of the police force or the strengthening of the criminal justice system. The researchers found that there was a remarkable difference between the answers of participants who obtained the crime-as-beast metaphorical framing and those who read the crime-as-virus framing: the former preferred enforcement significantly more frequently than the latter group (74% vs. 56%).

B) Non-exact replications of the original experiment and control experiments

The experimental complex evolving from OE involves several non-exact replications (NR) and control experiments (CON). Its basic structure looks like this:

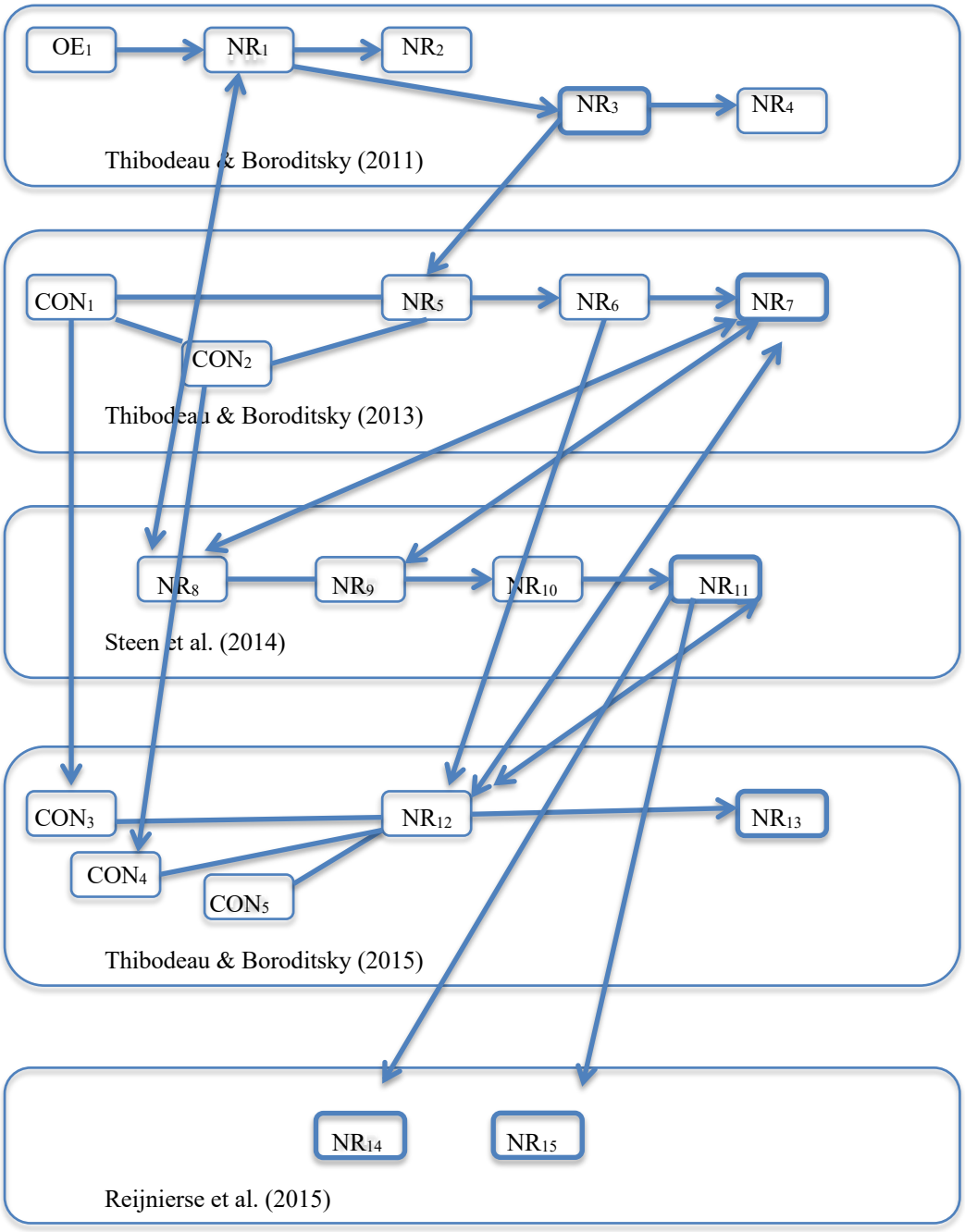

Figure 29. The structure of the experimental complex evolving from Thibodeau & Boroditsky (2011)

In order to provide a common basis for the comparison of the experiments, we will characterise the non-exact replications with the help of 5 parameters:[93]

[93] For a better understanding, Section 16.2.1 can also be consulted which presents a concise description of the experiments.

1) number of stories;
2) metaphorical content;
3) task;
4) coding system;
5) statistical tools applied.

In the case of the OE, this means the following:

OE (Thibodeau & Boroditsky 2011, Experiment 1):
1) Number of stories: 1 story in 2 versions ('virus' frame, 'beast' frame);
2) Metaphorical content: 3 metaphorical expressions belonging to one of the two metaphorical frames;
3) Task: suggesting a measure for solving the crime problem;
4) Coding: binary (social reform vs. enforcement), based on the authors' intuitions;
5) Statistical tools: chi-square test, without controlling for other possibly relevant factors such as age, political views, education, etc.

B1) Thibodeau & Boroditsky (2011)

The first step of our reconstruction is the description of the experiments along the 5 parameters. We will provide a full characterisation only of the original experiment and the limit-candidate; in all other cases, only modifications carried out to the predecessor of the given experiment will be highlighted.

NR_1 (Thibodeau & Boroditsky 2011, Experiment 2), compared to OE:
2) Metaphorical content: *1 metaphor belonging to one of the two metaphorical frames and further ambiguous metaphorical expressions*;
3) Task: suggesting a measure for solving the crime problem + *explaining the role of the police officers*;
4) Coding: binary (social reform vs. enforcement) *with both tasks and averaging the two values*, based on the authors' intuitions.

The first modification is motivated by a case of informational underdetermination insofar as on the basis of the data obtained from OE, one cannot decide whether a metaphorical framing effect can be triggered by many metaphorical expressions belonging to the same frame, or a single metaphor would suffice. The second modification is an improvement of the experimental design aiming at disambiguating the relatively frequent answer "increase the police force". The third modification is a consequence of the second change.

NR_2 (Thibodeau & Boroditsky 2011, Experiment 3), compared to NR_1:
2) Metaphorical content: *0 metaphor*;
3) Task: *providing synonyms for the words 'virus' or 'beast'*, suggesting a measure for crime reduction and explaining the role of police officers.

These changes are motivated by a case of informational underdetermination, too, because OE and NR_1 do not make it possible to rule out the possibility that even a single word might suffice to cause a metaphorical framing effect.

NR_3 (Thibodeau & Boroditsky 2011, Experiment 4), compared to NR_1 – limit-candidate(?):

1) Number of stories: 1 story in 2 versions ('virus' frame, 'beast' frame);
2) Metaphorical content: 1 metaphor belonging to one of the two metaphorical frames, presented at the beginning of the passage and further ambiguous metaphorical expressions;
3) Task: *selecting 1 crime-related issue from a range of 4 for further investigation*;
4) Coding: binary (social reform vs. enforcement), based on the authors' intuitions;
5) Statistical tools: chi-square test, without controlling for other possibly relevant factors such as age, political views, education.

The only change in comparison to NR_1 pertains to the type and focus of the task: instead of the application of an open question about the most important/urgent measure, participants had to choose one issue for further investigation from a 4-member list. This means two things. First, this version may be suitable for reducing informational underdetermination pertaining to the question of whether metaphorical frames can influence people in a similar manner if they have a broader range of possibilities to choose from. Second, asking for possible further investigations may go beyond people's spontaneous decisions and reveal the long term influence of metaphorical frames.

NR_4 (Thibodeau & Boroditsky 2011, Experiment 5), compared to NR_3:

2) Metaphorical content: 1 metaphor belonging to one of the two metaphorical frames, presented *at the end* of the passage, and further ambiguous metaphorical expressions.

Moving the metaphor to the end of the passage to be read might help to find out whether metaphors have an effect in isolation or make their impact by guiding and organising knowledge acquisition.

Summary: Every step of the problem-solving process is progressive in Thibodeau & Boroditsky (2011), because each non-exact replication provides a solution for at least one problem of its predecessor. This means in most cases, the elimination of informational underdetermination. Nonetheless, it is important to realise that while NR_1 is a revised version of OE, which replaces the latter, the relationship between NR_1-NR_4 is rather a complementary one. Jointly, they provide evidence for the hypothesis that even a single metaphor can organise the reception of a text in such a way that it influences both direct and long term decisions, while lexical activation of a metaphorical term cannot fulfil this function. Indeed, it is NR_3 that seems to be viewed by the authors as a limit-candidate within this chain of experiments. For the reasons for this, see the summary of Subsection B2.

B2) Thibodeau & Boroditsky (2013)

CON_1 (Thibodeau & Boroditsky 2013, Experiment 1), control experiment:

1) Number of stories: 1 story in 1 version (without metaphors);
2) Metaphorical content: 1 metaphorical sentence belonging to one of the two metaphorical frames, presented after reading the passage;
3) Task: ordering 1 measure each from a list of 4 to each metaphorical frame;
4) Coding: number of congruent choices (+2, 0, -2);
5) Statistical tools: chi-square test

CON_2 (Thibodeau & Boroditsky 2013, norming study), control experiment:

1) Number of stories: –;
2) Metaphorical content: –;
3) Task: rating the 5 measures on the basis of their reform/enforcement-orientedness;
4) Coding: analysis with the help of a 101-point scale, separately for each measure;
5) Statistical tools: t-test

CON_1 and CON_2 are control experiments. Their function is to check the correctness of the coding system applied in the main experiments.

NR_5 (Thibodeau & Boroditsky 2013, Experiment 2), compared to NR_3:

3) Task: *selecting the most effective crime-reducing measure from a range of 4*;
5) Statistical tools: chi-square test, logistic regression, *also with control for political views*.

The wording of the task was modified substantially in order to touch upon participants' attitude towards crime reducing measures directly. Several potentially relevant factors were taken into consideration during the statistical analyses.

NR_6 (Thibodeau & Boroditsky 2013, Experiment 3), compared to NR_5:

3) Task: selecting the most effective crime-reducing measure *from a range of 5*.

The only change to NR_5 was the extension of the selection of measures with the 'neighbourhood watches' option, whose evaluation was not unanimous, according to CON_1.

NR_7 (Thibodeau & Boroditsky 2013, Experiment 4), compared to NR_6, limit-candidate:

1) Number of stories: 1 story in 2 versions ('virus' frame, 'beast' frame);
2) Metaphorical content: 1 metaphor belonging to one of the two metaphorical frames and further ambiguous metaphorical expressions;
3) Task*: ranking 5 crime-reducing measures according to their effectiveness*;
4) Coding: binary (social reforms vs. enforcement), based on CON_1 and CON_2;
5) Statistical tools: chi-square test, logistic regression, also with control for political views.

There was only a slight difference between this experiment and its predecessor: the technique the participants used to rank the 5 measures was modified.

Summary: From the set of experiments NR_1-NR_4, only NR_3 has been continued in Thibodeau & Boroditsky (2013). Earlier experiments with a negative outcome seem to be regarded by the authors as completed, and the only line of research which was followed was one which entices us with positive results. Thus, the scope of the investigations has been narrowed down. An important improvement, however, is that the assignment of the crime-reducing measures to the metaphorical frames is no longer based on the intuition of the authors but has been checked with the help of two control experiments. The role of potentially relevant further factors was investigated, and the task given to participants was varied, too – more precisely, the formulation of the task was closer to the versions used in OE-NR_2. In this case, the relationship between the members of the chain of experiments NR_5-NR_7 is rather a linear one: each non-exact replication seems to be an improved version of its predecessor. Therefore, this is a progressive series of non-exact replications, too, with NR_7 as its limit-candidate.

B3) Steen et al. (2014)

NR_8 (Steen et al. 2014, Experiment 1, compared to NR_7):

1) Number of stories: *1 story in 3 versions (no-metaphor/'beast'/'virus' frame) in Dutch*;
2) Metaphorical content: 1 metaphor belonging to one of the two metaphorical frames and further ambiguous metaphorical expressions *vs. 1 metaphor belonging to one of the two metaphorical frames without metaphorical support*;
3) Task: ranking 5 crime-reducing measures according to their effectiveness *before and after reading the passage about crime*;
4) Coding: *+2 (two enforcement-oriented choices in the first two places) / +1 (one enforcement-oriented and one social reform oriented choice / 0 (two social reform-oriented choices), based on the authors' intuitions and/or Thibodeau & Boroditsky (2011, 2013)*;
5) Statistical tools: *ANOVA, logistic regression*, also with control for political views, age, etc.

The authors tried to improve on the earlier versions along all 5 dimensions. They added
– a no-metaphor version, in order to provide a neutral point of reference,
– a version without further metaphorical expressions (a 'without support' version), and
– the task of providing a ranking before reading the stimulus material, too.

They modified the coding system, and the method of the control for further possibly relevant factors, as well as the applied statistical tools. For instance, they took into consideration the first two choices instead of only the first one, and coded them in such a way that they obtained a 3-point scale instead of a purely binary classification.

NR_9 (Steen et al. 2014, Experiment 2, compared to NR_8):

1) Number of stories: 1 story in 3 versions (no-metaphor/'beast'/'virus' frame) *in English*;

Only the language was changed to NR_8. This kind of replication provides at least as strong a check of the reliability of the results as an exact replication would do.

NR_{10}-NR_{11} (Steen et al. 2014, Experiments 3-4, compared to NR_9), limit-candidate:

1) Number of stories: 1 story in 3 versions (no-metaphor/'beast'/'virus' frame);

2) Metaphorical content: 1 metaphor belonging to one of the two metaphorical frames and further ambiguous metaphorical expressions vs. 1 metaphor belonging to one of the two metaphorical frames without metaphorical support;
3) Task: ranking 5 crime-reducing measures according to their effectiveness *only after reading the passage about crime*;
4) Coding: +2 (two enforcement-oriented choices in the first two places) / +1 (one enforcement-oriented and one social reform oriented choice / 0 (two social reform-oriented choices), based on the authors' intuitions and/or Thibodeau & Boroditsky (2011, 2013);
5) Statistical tools: ANOVA, logistic regression, also with control for political views, age, etc.

One of the modifications of NR_8-NR_9, namely, pre-reading evaluation of the measures, was rejected. The only difference between NR_{10} and NR_{11} was the number of participants: NR_{11} applied a higher number of participants so as to have the power to detect small effects, as well.

Summary: Each non-exact replication is a clearly progressive step in Steen et al. (2014). Interestingly, NR_{10} and NR_{11} resolve problems which emerged in the previous members of this chain of experiments. Thus, they provide a kind of self-correction, and can be regarded as the limit-candidates within this chain of non-exact replications. Contrasting a 'without metaphorical support' with a 'with metaphorical support' condition also means a return to NR_1, although with a contradictory result.

B4) Thibodeau & Boroditsky (2015)

CON_3 (Thibodeau & Boroditsky 2015, norming task 1), control experiment, compared to CON_1):

1) Number of stories: 1 story in 1 version (without metaphors);
2) Metaphorical content: 1 metaphorical sentence belonging to one of the two metaphorical frames, presented after reading the passage;
3) Task: choosing 1 measure each *from a list of 5* that is most consistent with the given frame;
4) Coding: *analysis separately for each measure*;
5) Statistical tools: *logistic regression*

CON_4 (Thibodeau & Boroditsky 2015, norming task 2), control experiment, compared to CON_2:

1) Number of stories: –;
2) Metaphorical content: –;
3) Task: rating the 5 measures on the basis of their reform/enforcement-orientedness;
4) Coding: analysis with the help of a 101-point scale, separately for each measure;
5) Statistical tools: t-test

CON_5 (Thibodeau & Boroditsky 2015, norming task 3), control experiment:

1) Number of stories: 1 story in 4 versions ('beast', 'virus', 'problem', 'horrific problem');
2) Metaphorical content: 1 metaphorical sentence belonging to one of the two metaphorical frames and two non-metaphorical counterparts;

3) Task: ranking the 4 story versions according their severity, metaphoricity, and conventionality on a 101-point scale, and choosing the best one.
4) Coding: analysis separately for each measure;
5) Statistical tools: t-test

The three control experiments contribute to the inter-subjectivity of the results of NR_7 and NR_8 to a considerable extent.

NR_{12} (Thibodeau & Boroditsky 2015, Experiment 1), compared to NR_7, limit-candidate(?):

1) Number of stories: 1 story in 2 versions ('virus' frame, 'beast' frame);
2) Metaphorical content: 1 metaphor belonging to one of the two metaphorical frames and further ambiguous metaphorical expressions;
3) Task: ranking 5 crime-reducing measures according to their effectiveness;
4) Coding: binary (social reforms vs. enforcement), *based on CON_1 and CON_2, respectively*, and *also separate analyses for each measure*;
5) Statistical tools: chi-square test, logistic regression, also with control for political views *and other possibly relevant factors such as age, education, etc.*

Due to the two modifications and the application of the three control experiments, this non-exact replication is progressive. Both the separate statistical analysis of the full distribution of the first ranked choices and the deeper analysis of the role of several possibly relevant factors are seminal innovations.

NR_{13} (Thibodeau & Boroditsky 2015, Experiment 2), compared to NR_{12}:

3) Task: *choosing between 2 crime-reducing measures*.

The novelty of this member of the experimental complex is that it reduces the impact of the binary coding of the five measures in such a way that only the two most prototypical choices are offered for participants to decide between.

Summary: NR_{12} and NR_{13} add new elements to the experimental designs and rely on carefully elaborated and improved control experiments. At the same time, however, they do not react directly with counter-experiments on the modifications initiated by NR_8-NR_{11}.

B5) Reijnierse et al. (2015)

NR_{14} (Reijnierse et al. 2015, Experiment 1, compared to NR_{11}):

1) Number of stories: *1 story in 2 versions (no-metaphor/'virus' frame)*;
2) Metaphorical content: *0-1-2-3-4 metaphorical expressions*;
3) Task: *evaluating 4+4 crime-reducing measures according to their effectiveness on a 7-point Likert-scale*;
4) Coding: *average of the enforcement-oriented vs. reform-oriented values*;
5) Statistical tools: *one- and two-way ANOVA*, both with and without control for political affiliation, etc.

NR_{15} (Reijnierse et al. 2015, Experiment 2, compared to NR_{14}):
1) Number of stories: *1 story in 2 versions (no-metaphor/'beast' frame)*

NR_{14} and NR_{15} could be combined to make one experiment. The experimental design was improved at several points. Both the application of different numbers of metaphorical expressions and the modification of the task are innovative steps. The use of a Likert-scale is a more sensitive and informative tool than ranking the options and the binary coding of the first choice or the first two choices.

Summary: This pair of experiments is highly progressive, not only in comparison to its immediate predecessors but also because it might be suitable for reducing the informational underdetermination mentioned in relation to NR_1-NR_4.

16.1.2. Evaluation of the effectiveness of the problem-solving process

A) Thibodeau & Boroditsky (2011)

As we have seen in Subsection 16.1.1.B1, all members of the chain of experiments in Thibodeau & Boroditsky (2011) are progressive non-exact replications, because they provide a solution for at least one problem of their predecessors. Despite this, each of them remains multiply problematic, that is, they are burdened with problems which are associated with all parameters:

1) Number of stories: On the basis of solely one pair of metaphors, it is unfounded to generalise the research hypothesis to all metaphors. Moreover, it might, for example, be the case that it is not the metaphors themselves that make people prefer certain measures, but the fact that newspapers, Internet sources, politicians, etc. could have used a metaphor and associate it with a certain style of argumentation or policies. Such bias can be ruled out only with the help of corpus linguistics control research and, more importantly, with the involvement of several different topics and metaphors in the experiments.
2) Metaphorical content: As Steel et al. (2014: 4) also remark, the difference between the two versions of the stimulus material used in NE_1, NE_3 and NE_4 does not only lie in the word 'beast'/'virus', because the text contains further idiomatic expressions that can be interpreted differently in the two metaphorical frames. It is also debatable whether the phrases "was in good shape" or "the city's defence systems have weakened" are equally easily and naturally paired with both metaphors.
3) Task: One measure had to be named, one issue had to be chosen, etc. by participants. Therefore, the analysis of their behaviour is reduced to the choice of one measure. A second concern is that the task of selecting a crime-related issue for further investigation in NR_4 and NR_5 approaches peoples' opinion about the efficacy of the possible measures in a considerably more indirect way than earlier and later formulations of this task, leaving room for other interpretations by the participants.
4) Coding: The binary coding (social reforms vs. enforcement) is considerably less sensitive and informative than coding all possible answers separately, and it is based on a categorization which originates solely in the authors' intuitions.

5) Statistical tools: The first concern is that several possibly relevant factors such as age, political views, and education were taken into consideration only in subsequent statistical analyses. Secondly, and more importantly, the effect size, as both Cramér's V and the odds ratio values in Table 34 show, was small.

experiment	**OE**		**NR1**		**NR3**	
condition	**enforce**	**social**	**enforce**	**social**	**enforce**	**social**
beast	1.59 (0.1)	*-2.17 (0.03)*	1.22 (0.22)	-1.55 (0.12)	1.56 (0.12)	-1.03 (0.3)
virus	-1.61 (0.1)	*2.20 (0.028)*	-1.1 (0.27)	1.4 (0.16)	-1.46 (0.15)	0.96 (0.34)
Cramér's V	0.18 (p = 0.00013)		0.171 (p = 0.009)		0.192 (p = 0.014)	
odds ratio	2.15		2.05		2.32	
rate of congruent choices	59%		57%		60%	

Table 34. Standard residuals (and significance), effect sizes, rate of congruent choices in Thibodeau & Boroditsky (2011)

Effect size should be viewed as at least as important as significance in the interpretation of the results. Therefore, it is highly questionable whether it is justifiable to maintain the (universal) hypothesis that metaphors influence people's opinion if this influence is very limited in its magnitude and/or extent. Thirdly, if we break down the significant chi-square tests with standardized residuals, then we have to confront a further issue. Namely, the standardised residuals in the congruent cells (beast and enforce, virus and social) should be positive and significant, indicating that these cells contribute significantly to the chi-square value (and complementary, the incongruent cells should have significant minus values). Except for the social type answers in the original experiment, the values reveal that the response frequencies do not differ significantly from their expected values in the individual cells. This finding suggests that the differences are in the right direction, but that they are not strong enough. Moreover, since it is only OE that produced a result which is, at least in the case of one condition, in perfect harmony with the predictions, the authors' decision to continue solely with NR_3 in their later publications can be questioned. More specifically, the deeper statistical analysis of the perceptual data indicates that raising open questions as a task should not be abandoned, and the application of several metaphorical expressions belonging to the given frame should be investigated again.

A further interesting point is, as Table 35 shows, that there were changes in the proportions of the answers of the types 'enforce' and 'social'.

experiment	**OE**	**NR1**	**NR2**	**NR3**	**NR4**
enforce	65%	62%	64%	*30%*	*33%*
social	35%	38%	36%	*70%*	*67%*

Table 35. Count proportions in Thibodeau & Boroditsky (2011)

According to the authors' explanation, this shift is due to the application of a closed list of possibilities instead of open questions.[94] On the basis of later developments (see Subsection B), however, this explanation seems to be insufficient.

From these considerations it follows that none of the experiments in Thibodeau & Boroditsky (2011) can be regarded as the limit of this experimental complex, because they are not free of problems.

B) Thibodeau & Boroditsky (2013)

1) Number of stories: No improvement was made in comparison to Thibodeau & Boroditsky (2011).

2) Metaphorical content: No improvement was made in comparison to Thibodeau & Boroditsky (2011).

3) Task: The progressivity of this chain of experiments is to a considerable extent due to the more refined formulation of the tasks.

4) Coding: CON_1, that is, Experiment 1 in Thibodeau & Boroditsky (2013) is a control experiment, intended to test the hypothesis that people "can extract the metaphorical entailments of the two metaphors when they have an opportunity to compare the two frames explicitly" (Thibodeau & Boroditsky 2013: 4). According to the authors, from this "we should expect people to associate enforcement-oriented programs with the beast metaphor and reform-oriented programs with the virus metaphor." This means that this experiment intends to check the correctness of the stimulus material and coding system of Experiments 2-4. It is questionable, however, that this aim has been achieved. The decisive point is the statistical evaluation of the perceptual data. Namely, the authors conducted a chi-square test that showed that significantly more participants gave two congruent responses and significantly fewer participants provided two incongruent responses than expected by chance. If, however, we take into consideration that not all measures must have been assigned to the two metaphorical frames, but that participants had to choose only 1 measure each for both frames, then it seems to be more appropriate to accept only responses with 2 congruent solutions. To put it differently, it seems to be reasonable to collapse the answers into two categories (acceptable, i.e., 2 congruent answers vs. non-acceptable with 1 or 0 congruent answer), and require that at least 66% of participants gave an acceptable answer. This was, however, not the case. A binomial test indicated that the proportion of acceptable answers of 57% was significantly lower than expected, $p = 0.003$ (1-sided).

CON_2 is a control experiment, too. Here, the relatively low number of participants and the high standard deviations can be regarded as weak points. From this point of view, the evaluation of the "neighbourhood watches" option is pivotal, because it was only slightly above the midpoint of the scale. This finding and the large standard deviation indicate that the judgement of this option was rather equivocal. The authors' decision to dichotomize the results and force this option into the enforcement-oriented category exerted a decisive influence on the

[94] Cf. "Laying out four possible approaches to crime shifted the overall likelihood that people wanted to pursue social reform. It seems that explicitly seeing the space of possible responses makes people more likely to attempt reducing crime through reform than enforcement. However, we still found that peoples' responses were influenced by the frame that they read." (Thibodeau & Boroditsky 2011: 8)

interpretation of the experimental data obtained in CON_1 and NR_5-NR_7, too. Moreover, the "neighbourhood watches" option was not included in CON_1; thus, its assignment to the 'enforcement' category is even more questionable.

To sum up, a detailed re-analysis of the data for each option separately with both metaphors in CON_1 could be highly beneficial (see CON_3 on this). A further possibility could be the application of the numerical values obtained in CON_2 instead of the binary coding in the statistical evaluation of the results of CON_1 and the further experiments.

5) Statistical tools: The extension of the statistical analyses to the investigation of the impact of the political affiliation of participants in the main analyses is an important step. The problems mentioned in relation to OE-NR_4 in A), however, remain unsolved. What is more, NR_6 produces only marginally significant results ($\chi^2 = 3.761$, $p = 0.058$). See Tables 36 and 37.

experiment	**NR_6**		**NR_7**	
condition	**enforce**	**social**	**enforce**	**social**
beast	0.72 (0.47)	-1.2 (0.23)	1.1 (0.27)	-0.9 (0.37)
virus	-0.69 (0.49)	1.15 (0.25)	-1.17 (0.24)	0.93 (0.35)
Cramér's V	0.148 (p= 0.058)		0.111 (p = 0.049)	
odds ratio	1.99		1.58	
rate of congruent choices	56%		55%	

Table 36. Standard residuals (and significance) and effect sizes in Thibodeau & Boroditsky (2013)

experiment	**NR_5**	**NR_6**	**NR_7**
enforce	19%	76%	39%
social	81%	24%	61%

Table 37. Count proportions in Thibodeau & Boroditsky (2013)

If we compare the data in Table 37 with those in Table 35, it becomes clear that the authors' explanation for the finding that the rate of enforcement-oriented and social reform-oriented answers changes drastically among experiments cannot be sustained. Thibodeau & Boroditsky (2013: 5f.) identified two possible causes: the number of the measures from which participants could chose (2+2 vs. 3+2), and their political affiliation. These factors, however, do not seem to provide a satisfactory answer, for example, for the differences between NR_6 and NR_7.

A further issue needing a closer look is the choice of the statistical tools. First, the authors used logistic regression in their analyses. Since all data are categorical in NR_5-NR_7, chi-square test and loglinear analysis could be better choices, or, at least, it seems to be reasonable to use them as control analyses. Second, there are further alternatives which seem to be worth investigating. They are based on the abandonment of the questionable binary coding of the measures into reform- and enforcement options. This, as we have already mentioned, could happen in two ways, pointing in opposite directions.

a) *Analysing the relationship between metaphorical frames and the five response options directly*. With NR_6, a chi-square test indicated no effect of the frames: χ^2 (4) = 6.94, p = 0.141. Similarly, a chi-square test indicated no effect of the frames in the case of NR_7, either: χ^2 (4) = 5.876, p = 0.21. Tables 38 and 39 make it possible to reveal the enormous differences between the percentages and standardized residuals of the measures in NR_6 and NR_7, respectively:

	measure				
	economy	**education**	**patrols**	**prison**	**watch**
beast	2.4% 0.3	17.1% -1.2	46.3% 1.1	8.5% -0.8	25.6% 0.4
virus	3.4% 0.2	29.2% 1.1	31.5% -1.1	14.6% 0.8	21.3% -0.4
total	2.9%	23.4%	38.6%	11.7%	23.4%

Table 38. Count proportions in Thibodeau & Boroditsky (2013, Experiment 3)

	measure				
	economy	**education**	**patrols**	**prison**	**watch**
beast	42.1% -0.9	14.2% -0.3	14.2% 0.9	17.5% 1.1	12% -0.1
virus	51.2% 0.9	15.9% 0.3	9.4% -0.9	11.2% -1.1	12.4% 0.1
total	46.5%	15%	11.9%	14.4%	12.2%

Table 39. Count proportions in Thibodeau & Boroditsky (2013, Experiment 4)

b) *Analysing the relationship between metaphorical frames and enforcement-orientedness with the help of the experimental data obtained in CON_2*. Instead of dichotomising the responses, we might try to apply a finer scale with different values for each response. That is, the application of the ratings collected in CON_2 might represent the enforcement- vs. reform-orientedness nature of the measures in a better way. The analyses show that there is an effect of the frames – although the results are more convincing with NR_7. In the case of NR_6, a Mann-Whitney U test showed that the beast frame was significantly more enforcement-oriented (mean rank = 93.54) than the virus frame (mean rank = 79.06), U = 3031, p = 0.046 (two-sided). The mean enforcement value was 66.68 for the beast frame and 59.06 for the virus frame. A Kruskal-Wallis test reinforced the result that the enforcement-orientedness was significantly affected by the choice of the metaphorical frame; H(1) = 3.989, p = 0.046 (two-sided). As for NR_7, a Mann-Whitney U test showed that the beast frame was significantly more enforcement-oriented (mean rank = 187.83) than the virus frame (mean rank = 165.34), U = 1357.5, p = 0.028 (two-sided). The mean enforcement value was 44.5 for the beast frame and 37.05 for the virus frame. A Kruskal-Wallis test produced a similar result; H(1) = 4.813, p = 0.028 (two-sided).

These analyses should have produced similar results in the sense that they should be in harmony (that is, both should be either significant or non-significant). On the basis of the above

considerations, none of these non-exact replications can be regarded as a limit of this experimental complex, either.

C) Steen et al. (2014)

1) Number of stories: The most problematic point of OE-NR_7, namely, the use of only one pair of metaphors in the stimulus materials, questions the generality of the results of NR_8-NR_{11}, too. On the basis of only one pair of metaphors, one can draw neither positive nor negative conclusions about the research hypothesis.

3) Task: NR_8 and NR_9 cannot be regarded as data sources providing plausible experimental data, because raising the same questions before and after the presentation of the stimulus material could have influenced participants' decisions insofar that they might have stuck with their first decision. This could have diminished or masked the influence of the stimuli.

4) Coding: The assignment of the 5 measures to the two metaphors was not controlled for. Thus, the coding system is less reliable than it was in Thibodeau & Boroditsky (2013), because it is based either on the researchers' intuitions or was simply taken from earlier experiments.

5) Statistical tools: The authors applied ANOVA to Likert-type items, which is controversial. Thus, it seems to be advisable to repeat the statistical analyses with the help of tests allowing the dependent variable to be ordinal. Such tests are, for instance, Ordinal Logistic Regression or Optimal Scaling (Categorial Regression). Nevertheless, these tests reinforce the results of the authors: no metaphorical support can be identified. The same result was found with analyses narrowed down to the first chosen options.

We might also try the alternative analyses conducted with NR_6 and NR_7 in the previous subsection in this case, too.

a) *Analysing the relationship between metaphorical frames and the five response options directly*. With NR_{11}, a three-way loglinear analysis resulted in a model with a likelihood ratio of χ^2 (0) = 0. It indicated no three-way interaction between response, metaphorical frame and metaphorical support: χ^2 (8) = 8.228, p = 0.412, and no two-way interactions were found, either: χ^2 (14) = 15.072, p = 0.373. As Table 40 shows, the data produce a different pattern from the data obtained in earlier experiments; moreover, in several cases, their direction (sign) and/or their value is in sharp conflict with the predictions:

		measure				
		economy	**education**	**patrols**	**prison**	**watch**
neutral	**no support**	22.8% 0.1	21% -0.7	18% -0.5	4.8% 0.3	33.5% 0.8
	support	23.9% -0.3	20.9% 0.3	22.1% 0.3	1.2% -1.2	31.9% 0.2
beast	**no support**	*25.6%* *0.9*	20.6% -0.8	*21.1%* *0.4*	*3.9%* *-0.3*	28.9% -0.3
	support	*29.8%* *1.2*	18.5% -0.4	*21.3%* *0.1*	5.1% 1.9	*25.3%* *-1.4*
virus	**no support**	*18.6%* *-1.0*	29.3% 1.5	*19.8%* *0.0*	4.2% -0.1	28.1% -0.5
	support	*22%* *-0.9*	*20.2%* *0.1*	*19.7%* *-0.4*	1.7% -0.8	*36.4%* *1.2*

Table 40. Count proportions in Steen et al. (2014, Experiment 4)

Nonetheless, if we reduce our analyses to the 'with metaphorical support' version and focus solely on the comparison of the 'beast' and 'virus' frames, the results are marginally significant: χ^2 (4) = 8.684, p = 0.069. It is questionable, however, whether this result provides any support to the research hypothesis, because there should be differences between the 'virus' frame and the 'neutral' condition, as well as between the 'beast' frame and the 'neutral' condition, and these differences should point in opposite directions. This was, however, not the case.

b) *Analysing the relationship between metaphorical frames and enforcement-orientedness with the help of the experimental data obtained in CON₄.* When the first two choices were taken into consideration, a multiple regression found no effect of the frames or the presence of metaphorical support on enforcement-orientedness, F(2) = 0.525, p = 0.592, R^2 = 0.01. On a second attempt, only the first choice of participants was investigated. This analysis led to the same results, F(2) = 0.13, p = 0.988, R^2 = 0.00002. Similarly, negative results were produced by an analysis which used a non-parametric test, omitted the variable 'metaphorical support', and took into consideration only the data of participants who received the text with metaphorical support.

Summing up our analyses, we may conclude that no member of this chain of experiments can be regarded as the limit of the experimental complex, because each of them remained multiply problematic.

D) Thibodeau & Boroditsky (2015)

1) Number of stories: The same pair of metaphors was used in one story. Thus, there is no progress in this case, either.

2) Metaphorical content: Since no no-metaphor version was used and the number of metaphorical expressions was not varied, in this respect, this experiment rather counts as a relapse.

5) Statistical tools: Since there are no significant differences between the two conditions in respect to participants' age, political affiliation and gender in the two experiments, it is possible

to check the relationship between frames and responses directly. A chi-square test showed no significant effect of the frames in NR_{12}, χ^2 (1) = 1.432, p = 0.241. NR_{13} produced marginally significant results: χ^2 (1) = 3.322, p = 0.075. Table 41 helps us to compare the data with the outcomes of OE, NR_1, NR_3, NR_6 and NR_7:

experiment	**NR_{12}**		**NR_{13}**	
condition	**enforce**	**social**	**patrols**	**education**
beast	0.7	-0.5	0.8	-0.9
virus	-0.6	0.5	-0.9	1.0
Cramér's V	0.052 (p = 0.241)		0.080 (p = 0.068)	
odds ratio	1.24		1.38	
rate of congruent choices	53.4%		54.7%	

Table 41. Standard residuals and effect sizes in Thibodeau & Boroditsky (2015)

Alternative analyses:
a) *Analysing the relationship between metaphorical frames and the five response options directly*. With NR_{12}, a chi-square test indicated a significant effect of frames on the choice of the measures: χ^2 (4) = 13.748, p = 0.008. As Table 42 shows, however, the only category with significant differences was the response option 'watch'.

	measure				
	economy	**education**	**patrols**	**prison**	**watch**
beast	19.9% 0.9	24% 0.4	33.3% 1.2	6.1% -0.7	*16.7%* *-2.1*
virus	15.2% -0.9	21.6% -0.4	25.9% -1.1	8.5% 0.7	*28.7%* *2.0*
total	17.4%	22.7%	29.4%	7.4%	23.1%

Table 42. Count proportions in Thibodeau & Boroditsky (2015, Experiment 1)

b) *Analysing the relationship between metaphorical frames and enforcement-orientedness with the help of the experimental data obtained in CON_4*. An analysis making use of the ratings collected in CON_4 showed no effect of the frames. According to a Mann-Whitney U test, there is no significant difference between the beast frame (mean rank = 259.79) and the virus frame (mean rank = 268.61), U = 35844.5, p = 0.496 (two-sided). The mean enforcement value was 47.05 for the beast frame and 46.07 for the virus frame. A Kruskal-Wallis test produced the same results; H(1) = 0.464, p = 0.496 (two-sided).

As for NR_{13}, a chi-square test showed only a marginally significant effect of the metaphorical frame: χ^2 (1) = 3.322, p = 0.075. A loglinear analysis indicated a clearly significant interaction between political affiliation and response: χ^2 (2) = 13.203, p = 0.001, a marginal interaction between response and frame: χ^2 (1) = 3.235, p = 0.072, and no three-way interaction among these factors: χ^2 (2) = 0.24, p = 0.887.

This means that there is no unproblematic non-exact replication in Thibodeau & Boroditsky (2015), either.

E) Reijnierse et al. (2015)

1) **Number of stories**: Similarly to NR_5-NR_{13}, there was only one story, although it was presented in two slightly different versions (crime described as a long-term problem vs. a short-term problem) in NR_{14} and NR_{15}, respectively. Thus, only two sets of metaphors were used again.
3) **Task**: Participants had to evaluate 8 crime-reducing measures according to their effectiveness on a 7-point Likert-scale. This step could produce more sensitive measures and lead to more valuable experimental data than was the case in the previous experiments. The authors, however, presented the measures not in a random order for each participant but showed the frame-consistent 4 measures first and the other 4 measures second. This might lead to a bias which seriously calls into question the validity of the results, because the skewing effect of the presentation order could not be eliminated.
4) **Coding**: Besides the basically binary coding (average of the enforcement-oriented vs. reform-oriented values), a comparison of the values separately for each measure could also be informative.
5) **Statistical tools**: Similarly to NR_{10}-NR_{11} in Steen et al. (2014), the application of ANOVA to Likert-scale items is debatable.

Despite the innovative character of the experimental design in Reijnierse et al. (2015), both experiments remained problematic.

16.1.3. Re-evaluation of the problem-solving process and revealing future prospects

Our analyses showed that the later experiments conducted by the same authors do not provide converging evidence for the authors' standpoint. That is, we cannot interpret the situation in such a way that the plausibility values of the experimental data would add up to continuously higher values. Instead, the relationship among these experiments is determined by the operation of *recurrent re-evaluation*: the newer, revised versions replace the earlier ones; nonetheless, this process cannot be regarded as a steady improvement due to the emergence of new problems.

On the basis of the re-evaluation of the non-exact replications in Subsection 16.1.2, we have to conclude that the experimental complex evolving from the set of experiments presented in Thibodeau & Boroditsky (2011) *has not reached a limit* in Thibodeau & Boroditsky (2011, 2013, 2015), Steen et al. (2014) or Reijnierse et al. (2015). Since no non-exact replication was superior in all respects to its rivals, the application of the Combinative Strategy seems to be more appropriate. The process of re-evaluation should be continued by collecting and systematising all relevant and workable elements of the experiments conducted within this experimental complex, and new limit-candidates should be elaborated. The high quality of the experiments analysed does not allow an easy and clear-cut decision but is a rich source of inspiration and guidance for further progress. That is, the experiments conducted by both parties motivate and provide starting points for the continuation of the cyclic re-evaluation process in order to elaborate a limit for this experimental complex along the following lines:

- **Number of stories**: The most important change could be to increase the number of stories.[95] One cannot draw general conclusions on the basis of only one topic and two metaphorical frames. For example, it might be the case that politicians and the press made use of a metaphorical frame, which, as a result, can be associated with a certain political standpoint. Such influences can be circumvented only if the experiments make use of several different stories and a great variety of metaphors.
- **Metaphorical content**: It remained an open question as to whether only one metaphorical expression might influence participants' decisions or a series of expressions belonging to the same metaphorical frame are needed. A no-metaphor control seems to be warranted. In addition, it could be also examined whether novel and conventional metaphors have the same effect.
- **Task**: There were plenty of more or less different versions of the task. The most sensitive and informative seems to be the application of Likert-scales, but the use of an open question (such as suggesting a measure for solving the crime problem in OE) in a first set of experiments could be beneficial, too, in order to provide a comprehensive and varied set of options to participants.
- **Coding system**: Both the basically binary coding (average of the enforcement-oriented vs. reform-oriented values), and an analysis of the values separately for each measure would be beneficial, providing information from different points of view.
- **Applied statistical tools**: The applied statistical tools should be chosen in such a way that their applicability to diverse variable types is taken into consideration. The impact of possible relevant factors such as political affiliation, age, education, etc. should be controlled for properly. A further important task evolves from the small effect size – provided that the further non-exact replications will also find small effect sizes. Namely, one should attempt to narrow down the investigations to subgroups so that the factors characterising people who are responsive to the influence of metaphors can be identified.

16.2. Case study 6, Part 2: Statistical meta-analysis of a debate on the role of metaphors on thinking

The two series of replications by the two camps Thibodeau & Boroditsky vs. Steen et al. not only repeatedly came to opposite conclusions but they also provided different estimates of the effect of the metaphorical frames. While Thibodeau and Boroditsky concluded, for example, that

> "[w]e find that exposure to even a single metaphor can induce *substantial differences* in opinion about how to solve social problems: differences that are *larger*, for example, than pre-existing differences in opinion between Democrats and Republicans" (Thibodeau & Boroditsky 2011: 1; emphasis added),

the other camp stated that

[95] Thibodeau (2016) made important steps in this direction.

> "We *do not find* a metaphorical framing effect." (Steen et al. 2014: 1; emphasis added)
> "Overall, our data *show limited support* for the hypothesis that extended metaphors influence people's opinions." (Reijnierse et al. 2015: 258; emphasis added)

In order to put forward a more reliable estimation of the true effect size, we have to apply the tools of statistical meta-analysis in a novel way.[96] This means, above all, that we intend to investigate whether and how meta-analytic tools can be applied to conflict resolution in a case in which there are only a few experiments at our disposal.[97] Since meta-analysis will be applied to a limited number of experiments, there will unavoidably be some deviations from the customary practice as stipulated by the standard protocol called the 'PRISMA 2009 checklist'.

Section 16.2.1 explains the procedure of selecting the experiments included in the meta-analysis. Section 16.2.2 describes the methods applied in the choice of the effect size index and the data collection process and shows how effect sizes can be calculated at the level of the individual experiments if we focus on the participants' top choices. Section 16.2.3 deals with the combination of the experiments' effect sizes, that is, the calculation of the summary effect size, the methods used to check their consistency, as well as methods for revealing possible publication bias, and then presents the results. Section 16.2.4 presents alternative analyses: an analysis which takes into consideration the whole range of the measures and an analysis comparing the effect of the metaphorical frames on the measures separately. Section 16.2.5 summarises the main findings, draws conclusions and discusses the limitations of the results.

16.2.1. The selection of experiments included in the meta-analysis

As we have seen in Section 11.2.2, the first task is the selection of the experiments which allow us to estimate the strength of the relationship between two variables. The experimental complex evolving from Thibodeau & Boroditsky (2011) comprises a series of experiments investigating the effect of metaphorical framing on readers' preference for frame-consistent/inconsistent political measures. The following short description of the experiments should be sufficient to show that the majority of them are similar enough and that it is possible to apply the tools of meta-analysis to their results.

---

[96] "[…]'research synthesis' and 'systematic review' are terms used for a review that focuses on integrating research evidence from a number of studies. *Such reviews usually employ the quantitative techniques of meta-analysis to carry out the integration*." (Cumming 2012: 255; emphasis added)
"*A key element in most systematic reviews is the statistical synthesis of the data, or the meta-analysis*. Unlike the narrative review, where reviewers implicitly assign some level of importance to each study, in meta-analysis the weights assigned to each study are based on mathematical criteria that are specified in advance. While the reviewers and readers may still differ on the substantive meaning of the results (as they might for a primary study), *the statistical analysis provides a transparent, objective, and replicable framework for this discussion*." (Borenstein et al., 2009: xxiii; emphasis added)

[97] Since the selection of relevant studies always and unavoidably leaves room for subjective factors, nothing precludes a restricted use of the tools of meta-analysis to a smaller but well-defined set of experiments:
"For systematic reviews, a clear set of rules is used for studies, and then to determine which studies will be included in or excluded from the analysis. Since there is an element of subjectivity in setting these criteria, as well as in the conclusions drawn from the meta-analysis, we cannot say that the systematic review is entirely objective. However, because all of the decisions are specified clearly, *the mechanisms are transparent*." (Borenstein et al., 2009: xxiii; emphasis added)

Thibodeau & Boroditsky (2011), Experiment 1: Participants were presented with one version of the following passage:

> "Crime is a **{wild beast preying on/virus infecting}** the city of Addison. The crime rate in the once peaceful city has steadily increased over the past three years. In fact, these days it seems that crime is **{lurking in/plaguing}** every neighborhood. In 2004, 46,177 crimes were reported compared to more than 55,000 reported in 2007. The rise in violent crime is particularly alarming. In 2004, there were 330 murders in the city, in 2007, there were over 500."

Then, they had to answer the open question of what, in their opinion, Addison needs to do to reduce crime. The answers were coded into two categories on the basis of the results of a previous norming study: 1) diagnose/treat/inoculate (that is, they suggested introducing social reforms or revealing the causes of the problems) and 2) capture/enforce/punish (that is, they proposed the use of the police force or the strengthening of the criminal justice system).

Thibodeau & Boroditsky (2011), Experiment 2: In this experiment, the passage to be read, besides a metaphor belonging to one of the two metaphorical frames, also included further ambiguous metaphorical expressions which could be interpreted in both metaphorical frames. The task was to suggest a measure for solving the crime problem and explain the role of the police officers in order to disambiguate the answers.

Thibodeau & Boroditsky (2011), Experiment 4: The only change in comparison to Experiment 2 pertains to the type and focus of the task: instead of the application of an open question about the most important/urgent measure, participants had to choose one issue for further investigation from a 4-member list:

1. Increase street patrols that look for criminals. (coded as 'street patrols')
2. Increase prison sentences for convicted offenders. ('prison')
3. Reform education practices and create after school programs. ('education')
4. Expand economic welfare programs and create jobs. ('economy')

Thibodeau & Boroditsky (2013), Experiment 2: The wording of the task was modified substantially against Experiment 4 of Thibodeau & Boroditsky (2011) in order to touch upon participants' attitudes towards crime reducing measures directly. Namely, it consisted of selecting the most effective crime-reducing measure from a range of 4.

Thibodeau & Boroditsky (2013), Experiment 3: The only change made to Experiment 2 was the extension of the selection of measures with the 'neighbourhood watches' option ("Develop neighborhood watch programs and do more community outreach.").

Thibodeau & Boroditsky (2013), Experiment 4: There was only a slight difference between this experiment and its predecessor: the technique the participants used to evaluate the 5 measures was modified. That is, their task was to rank 5 crime-reducing measures according

to their effectiveness. Nonetheless, only the top choice was used for the creation of the experimental data by the authors.

Steen et al. (2014), Experiment 1: The authors extended the stimulus material with a no-metaphor version, in order to provide a neutral point of reference, and a version without further metaphorical expressions (a 'without support' version). Here, too, participants had to rank the 5 crime-reducing measures according to their effectiveness before and after reading the passage about crime.

Steen et al. (2014), Experiment 2: Only the language was changed from Experiment 1 (English instead of Dutch).

Steen et al. (2014), Experiments 3-4: The idea of a pre-reading evaluation of the measures was rejected. Thus, the task for the participants consisted of ranking the five crime-reducing measures according to their effectiveness only after reading the passage about crime. The only difference between Experiments 3 and 4 was the number of participants: the latter used a higher number of participants so as to have the power to detect small effects, as well.

Thibodeau & Boroditsky (2015), Experiment 1: The only change to Experiment 3 in Thibodeau & Boroditsky (2013) was the application of three control experiments in order to improve the stimulus material's validity.

Thibodeau & Boroditsky (2015), Experiment 2: The novelty of this member of the experimental complex is that it reduces the impact of the binary coding of the five measures in such a way that only the two most prototypical choices were offered for participants to decide between.

Reijnierse et al. (2015), Experiment 1: This experiment made use of 1 story in 2 versions (no-metaphor/'virus' frame). The metaphorical content was varied so that the passage to be read by participants contained 0, 1, 2, 3, or 4 metaphorical expressions. The task consisted of evaluating 4+4 crime-reducing measures according to their effectiveness on a 7-point Likert-scale. Then, the average of the enforcement-oriented vs. reform-oriented values were compared.

Reijnierse et al. (2015), Experiment 2: Identical to Experiment 1, except that there was a 'beast' frame instead of a 'virus' frame.

Christmann & Göhring (2016): This was an attempt at an exact replication of Thibodeau & Boroditsky (2011), Experiment 1 in German.

In contrast, the following three experiments had to be excluded from the meta-analysis:

Thibodeau & Boroditsky (2011), Experiment 3: The stimulus material did not contain metaphors. Instead, participants had to provide synonyms for the words 'virus' or 'beast', suggest

a measure for crime reduction, and explain the role of police officers. Since there were no metaphors in the passage to be read, this experiment will be excluded from the meta-analysis.

Thibodeau & Boroditsky (2011), Experiment 5: In contrast to Experiment 4, the metaphor belonging to one of the two metaphorical frames was presented at the end of the passage. Presentation of the target metaphor at the end of the passage leads to a situation which is substantially different from the previous experiments.

Thibodeau & Boroditsky (2013), Experiment 1 was a control experiment.

16.2.2. The choice and calculation of the effect size of the experiments

A) The data structure of the experiments

The brief characterization of the experiments in the previous section and a closer look at the data handling techniques of the authors reveal a highly important issue: namely, both the tasks which the participants had to perform and the methods for creating experimental data from the raw (perceptual) data were different in the experiments at issue.

Thibodeau & Boroditsky (2011), Experiments 1, 2 and 4

Experiment 1 utilized an open question task. Participants' answers were first coded separately by the authors into the two categories 'social reform' vs. 'enforcement', and then rendered as either purely social-type (1-0), purely enforcement-type (0-1) or mixed (0.5-0.5). In Experiment 2, this procedure was also applied to the question about the role of the police, and the two answers were averaged. In Experiment 4, participants had to choose one measure. Thibodeau and Boroditsky coded the answers as either social reform-oriented or enforcement-oriented.

Thibodeau & Boroditsky (2013), Experiments 2-4

The data sets pertaining to Experiments 3 and 4 have been made accessible by the authors at https://osf.io/r8mac/. These data sets do not include information about the whole ranking of the measures but only participants' first choices. In the evaluation of the data, the authors also included participants' second choices, and examined their orientedness and coherence with the metaphorical frame.

Steen et al. (2014), Experiments 1-4

The data sets can be downloaded from https://osf.io/ujv2f/ as SPSS data files. Both the post-reading and pre-reading responses of participants were captured, and there was also a 'with metaphorical support' versus 'without support' version. The first two choices were taken into consideration by the researchers. The answers were coded with the help of the following 3-point scale: +2 (two enforcement-oriented choices in the first two places) / +1 (one enforcement-oriented and one social reform oriented choice / 0 (two social reform-oriented choices). The results of participants with a shorter reading time than 5s or longer than 60s, and those under 18 years of age were excluded. Residency different from the Netherlands/US and native language different from Dutch/English were not allowed, either.

Thibodeau & Boroditsky (2015)
The range of the experimental data and the methods of their treatment were almost identical with those used in the case of Experiments 2-4 in Thibodeau & Boroditsky (2013).

Reijnierse et al. (2015), Experiments 1-2
The data have been made public on the following Open Science Framework site: https://osf.io/63ym9/. The authors processed the data in such a way that they examined the effect of the number of metaphorical expressions on the perceived efficiency ratings of the two types of measures with the help of a one-way independent ANOVA, separately with both frames.

Christmann & Göhring (2016)
As was the case with Experiment 1 of Thibodeau & Boroditsky (2011), an open question task was applied. The coding system has, however, been modified. Since the number of answers which could not be assigned to the category 'social reform' or 'enforcement' was relatively high, the authors excluded them from their analyses. Table 1 on page 4 contains the response frequencies. The authors, however, made all the answer sheets available at the Open Science Framework site https://osf.io/m7a5u/. I used this data source, and revised the authors' decisions on some occasions.

B) The choice of the effect size indicator
In order to reduce the impact of the diversity of methods applied by the researchers, the data handling techniques have to be standardized. The most straightforward possibility is to analyse *the impact of the frames (beast vs. virus) on the orientedness (social reform vs. enforcement) of the top choices*. The question is, of course, how this can be achieved.

The simplest way to calculate the effect of the metaphorical frames on the choice of the measures consists of comparing the odds of choosing a social type response against choosing an enforcement type response in the first place in the virus condition and the odds of choosing a social type response against choosing an enforcement type response in the first place in the beast condition – i.e. computing the *odds ratio*:[98]

$$\mathrm{OR} = \frac{\text{odds of choosing a social type response against choosing an enforcement type response in the first place in the virus condition}}{\text{odds of choosing a social type response against choosing an enforcement type response in the first place in the beast condition}} =$$

$$= \frac{\text{number of participants choosing a } \textit{social type response} \text{ in the first place } \textit{in the virus condition} \text{ / number of participants choosing an } \textit{enforcement-type response} \text{ in the first place } \textit{in the virus condition}}{\text{number of participants choosing a } \textit{social type response} \text{ in the first place } \textit{in the beast condition} \text{ / number of participants choosing an } \textit{enforcement-type response} \text{ in the first place } \textit{in the beast condition}}$$

[98] There are several effect size indicators which can be calculated with dichotomous variables. Among these, the odds ratio is the most versatile (but not intuitively interpretable).

In order to illustrate how different OR values can be interpreted, let us experiment with some possible scenarios. See Table 43.

metaphorical frame	beast		virus		OR	conf. int.
response type	social	enforcement	social	enforcement		
Scenario 1	50	50	50	50	1	[0.57; 1.74]
Scenario 2	46	54	52	48	1.27	[0.73; 2.22]
Scenario 3	40	60	65	35	2.79	[1.57; 4.94]
Scenario 4	25	75	80	20	12	[6.16; 23.38]
Scenario 5	60	40	35	65	0.36	[0.2; 0.64]
Scenario 6	65	35	70	30	1.26	[0.69; 2.27]

Table 43. OR value calculations

In Scenario 1, we see a perfect tie between social reform- and enforcement-oriented first choices. This yields an odds ratio of 1. That is, if OR is 1, then we can conclude that the metaphorical frame does not affect the choice of the responses. In Scenario 2, with both frames, the frame-consistent answers were slightly preferred by participants. This yields an OR somewhat greater than 1. In Scenario 3, the frame-consistent choices approach a two-thirds majority – and the OR approaches a value of 3. If more than 75% of participants give a frame-consistent answer, then the OR rises to 12. Scenario 5 shows what happens if participants chose frame-inconsistent responses: the OR is between 0 and 1. Finally, in Scenario 6, in both frames it is the social reform-type choices that are in the majority. Since the proportion of the frame-consistent answers is slightly higher in the virus frame than that of the frame-inconsistent responses in the beast frame, we obtain an OR slightly higher than 1.

It is vital to take into consideration the *precision of these estimates*, too. To this end, we can calculate the 95% confidence intervals of the OR values. This shows a range which – in 95% of cases – encompasses the odds of choosing a social type response against an enforcement type response in the virus condition compared to the beast condition. For example, the confidence interval in Scenario 5 is narrow. This indicates that the precision of the estimate is high. In this case, the confidence interval does not overlap the value 1. Therefore, we can conclude that participants who obtained the crime-as-virus metaphorical framing preferred social reform-type answers *significantly less frequently* than those who read the crime-as-beast framing.[99] In contrast, in Scenarios 3 and 4, participants gave frame-consistent answers significantly more often, since the whole confidence interval is above the value 1 – although the precision of these estimates is lower, as the width of the confidence interval shows. Scenarios 1, 2 and 6, however, did not produce significant results, because their confidence intervals include the value 1.

As a next step, we need data from which the odds ratio can be calculated for each experiment. In some cases, this was an easy task, in other cases, further data had to be collected from

[99] "There is a necessary correspondence between the p-value and the confidence interval, such that the p-value will fall under 0.05 if and only if the 95% confidence interval does not include the null value [with the odds ratio, this is 1]. Therefore, by scanning the confidence intervals we can easily identify the statistically significant studies." (Borenstein et al., 2009: 5)

the authors and/or some work was needed to extract the relevant information from the data sets available.

C) Methods of data collection

With the help of the CMA software, effect sizes can be computed from about 100 options, i.e., more than 100 summary data types, but there are also several online effect size calculators such as this one: https://www.psychometrica.de/effect_size.html. Since the data sheets made available by the researchers on a special Open Science Framework site or via email make it possible to collect information about the events and sample size in each group, it is better (i.e., will result in more precise effect size values) to make use of these data (and apply the formula presented in the previous section) than, for example, the Chi-squared and the total sample size, as published in the research papers. This decision is motivated by the finding that if there are several possibilities, then the method which is closer to the raw data should be preferred. Reliance on the summary data presented in the experimental reports is not a compulsory step of meta-analysis but often a necessity, because we do not usually have access to the data sets.

This means that from this set of experiments, data with the following structure should be extracted:

– the number of participants choosing a social reform type measure in the 'beast' condition;
– the number of participants choosing an enforcement type measure in the 'beast' condition;
– the number of participants choosing a social reform type measure in the 'virus' condition;
– the number of participants choosing an enforcement type measure in the 'virus' condition.

In most cases, these data could not be found in the research report but could be produced from the information in the data sheets.

Thibodeau & Boroditsky (2011), Experiments 1, 2 and 4

With the help of the data sets provided by Paul Thibodeau via email communication, response frequencies of the social reform-type vs. enforcement-type answers in the two conditions (beast frame vs. virus frame) could be computed easily.[100] Nevertheless, in the case of Experiments 1 and 2, it had to be decided whether and how answers to an open question can be transformed into the orientedness of the most preferred measure. Following the authors' proceedings, it was supposed that the number of the answers belonging to one category of orientedness (i.e., social reform type or enforcement-type) may be regarded as mirroring participants' preferences.

Thibodeau & Boroditsky (2013), Experiments 2-4

An extended set of data, together with a data file related to Experiment 2, were provided for me by Paul Thibodeau on email request. I converted the csv files into SPSS data files (sav), and in accordance with the data handling methods applied in Thibodeau & Boroditsky (2015), excluded data from participants quicker than 10s or slower than 300s from the analyses. Furthermore, only the results of participants with US residence and English as a native language were taken into consideration.

[100] Table 1 of Thibodeau & Boroditsky (2011: 5) presents practically the same information content.

Thus, the first choice of participants had to be coded as social reform-oriented or enforcement-oriented and the number of participants choosing them in the two conditions (beast frame, virus frame) calculated, as was the case with Thibodeau & Boroditsky (2011).

Steen et al. (2014), Experiments 1-4
I restricted the analyses to the post-reading responses. The 'with metaphorical support' versus 'without support' distinction, however, was not taken into account, but these data sets were unified. I changed the coding system applied by the authors to the first choice of participants.

Thibodeau & Boroditsky (2015)
See the procedures applied with Thibodeau & Boroditsky (2013), Experiments 2-4.

Reijnierse et al. (2015), Experiments 1-2
At the outset, the 8 measures were recoded with the help of the authors' instructions, so that the four measure categories used in the previous experiments could be applied in this case, too.[101] The highest social-reform oriented and enforcement-oriented values were compared in the following way. If the measure 'economy' or 'education' obtained the highest value from a participant, then the first choice of this participant was coded as social-reform oriented. If he/she gave the highest rating to the measure 'patrols' or 'prison', then his/her top choice was coded as enforcement-oriented. If the highest values in the two categories were identical, then data from this participant were omitted from the analysis.

Christmann & Göhring (2016)
Since an open question task format was applied, I recoded participants' answers either as enforcement-type or social reform-type. Then, in the case of each participant, I regarded the majority of the answers as decisive. If there was a tie, the data of the given participant were omitted from the analysis.

Table 44 in Appendix 2 shows the response frequencies the data extraction process yields.

D) The effect size of the individual experiments
Figure 30 shows the individual effect sizes, their confidence intervals, Z-values, p-values, and weights.

[101] See Table 2 in Reijnierse et al. (2015: 251).

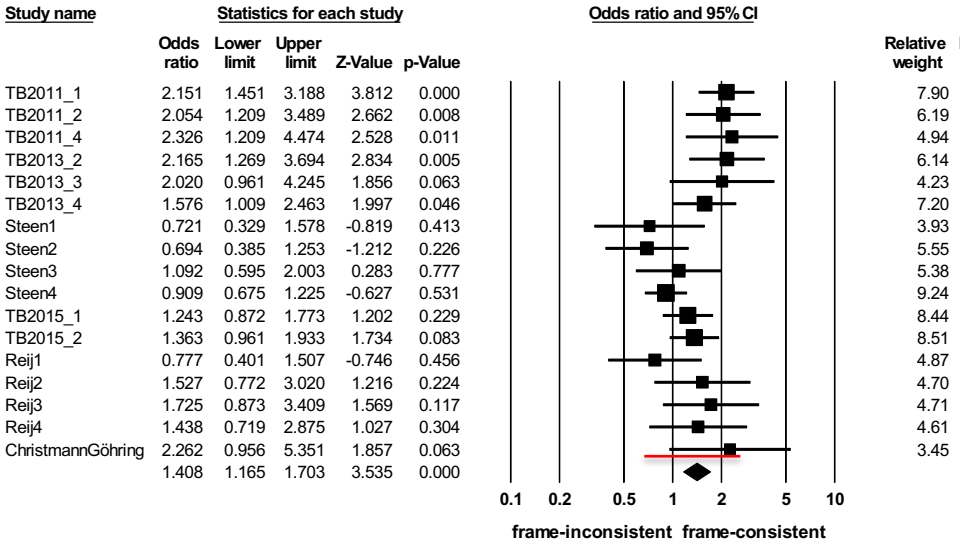

Figure 30. Effect sizes of the experiments and the summary effect size in the first analysis (top choices of participants)

The odds ratios of the individual experiments ranged from 0.694 (Steen et al., 2004, Experiment 2) to 2.326 (Thibodeau & Boroditsky 2011, Experiment 4). An odds ratio greater than 1 means that participants preferred frame-consistent answers, while an odds ratio below 1 means the opposite. In 13 of the 17 cases, the odds ratio was higher than 1. There seem to be subgroups regarding the effect size values. Experiments in Thibodeau & Boroditsky (2011, 2013), except for Experiment 4 of their 2013 series of experiments, as well as the replication by Christmann & Göhring indicate an effect size slightly greater than 2. Thibodeau and Boroditsky's 2015 paper shows effect sizes somewhat above 1. Experiments 1 and 2 in Steen et al. (2014) and the 1-metaphor condition in Reijnierse (2015) produced effect sizes clearly below 1. Experiments 3 and 4 in Steen (2014) are very close to 1, while the remaining experiments conducted by these authors indicate an effect size around the 1.5-mark. This means, in sum, that the experiments show a weak or no effect of the metaphorical frame.

There were only 5 experiments for which the confidence interval did not include the value 1. These all are completely above the 1-mark line and represent a significant result *for* the research hypothesis. In contrast, there was no experiment which would provide a significant result *against* Thibodeau & Boroditsky's research hypothesis. The lowest point of the confidence intervals was 0.329, while the highest was 5.351.

Experiment 4 of Steen et al. (2014) provides the most precise estimation of the effect size with a quite narrow confidence interval of [0.675, 1.225], while the replication by Christmann & Göhring (2016) is the least precise: its confidence interval of [0.956, 5.351] is noticeably wide.

From these results it would be premature to conclude that the majority decides, and the experiments together yield a statistically insignificant, weak support for Thibodeau and Boroditsky's research hypothesis. The aim of meta-analysis is, as we have already said in Section 11.2.1, not to count votes but to calculate an estimate of the effect size *on the basis of all the information inherent in the data from the experiments synthesized*. The next section will show which method should be chosen and how it can be applied in this case.

16.2.3. Synthesis of the effect sizes

A) Calculation of the summary effect size

Having established the effect sizes of the experiments, the next step consists of estimating the summary effect size. This step was carried out with the help of the CMA software. As we have seen in Section 11.2.4, there are basically two methods to combine the effect sizes of individual experiments: the fixed-effect model and the random-effect model. In our case, the application of the random-effect model seems to be unequivocal, since the experiments were conducted at different times by different researchers, the task of participants was modified several times, and the data on which the calculation of the effect sizes is based does not take into consideration any possible relevant factors such as political affiliation, age, education, etc. This means that the mean effect size is calculated as a weighted mean of the experiments' effect sizes in such a way that the weights are the inverse of the sum of the between-studies variance and the within-studies variance. In this way, two components are taken into consideration. The first component consists of the differences between the individual effect sizes, since we cannot suppose that all experiments share a common effect size. The second component is the size of the experiments, since larger experiments will be assigned a greater weight than smaller ones. The last row of Figure 30 shows the summary effect size with its 95% confidence interval.

The summary effect size of 1.408 is significant, $Z = 3.535$, $p = 0.004$. Its confidence interval [1.165, 1.703] does not include the value 1, and overlaps with the majority of the confidence intervals of the individual experiments. This confidence interval is quite narrow, indicating a precise estimation of the summary effect. To put it differently, the mean effect size probably (in 95% of the cases) falls between 1.165 and 1.703. From these data we can conclude that the experiments together provide evidence for Thibodeau and Boroditsky's research hypothesis, although the summary effect size is quite low.

B) The consistency of the effect sizes

Following this, the consistency of the (true) effect sizes needs to be investigated.[102] In this case, the Q-value, i.e. the total amount of the between-experiments variance observed, is 34.486. Its expected value is df(Q) = 16. These two values differ significantly from each other; $p = 0.005$. This means that the total variation is significantly greater than the sum of the within-study variations, indicating that *these experiments do not share a common true effect size*. The second relevant indicator is the estimate for the between-study standard variation of the true effects, denoted as T^2. This is 0.078 in log units with a standard error of 0.055. This yields that the standard deviation of the true effects, i.e. T, is 1.322. Finally, the I^2 value is 53.605, which means that about 54% of the observed variance in effect sizes cannot be attributed to random error but reflects differences in the true effect sizes of the experiments. This indicates *a moderate amount of variation in the true effect sizes*. Therefore, we should try to find subgroups among the studies which constitute more homogenous classes, or perform meta-regression in order to identify possible covariates.

[102] Cf. Section 11.2.6.

C) Subgroup analysis

a) Subgroup analysis – by authors

Our analyses in the previous subsection yielded the result that there is a moderate amount of variation in the true effect sizes. If we want to reveal the cause of this heterogeneity, one possibility is subgroup analysis. Since the great majority of Thibodeau & Boroditsky's experiments produced effect sizes above the 1-mark line, while the opposite is true of Steen et al.'s experiments, it seems to be well-motivated first to classify the experiments into two groups on the basis of their authors. There are several methods of subgroup analysis. In this case, a mixed-effects model with a pooled estimate of τ^2 seemed to be the most appropriate. This means that a random-effects model was used within subgroups and a fixed-effect model was applied to combine the two subgroups.

Figure 31 presents the outcome of the subgroup analysis based on participants' first choices.

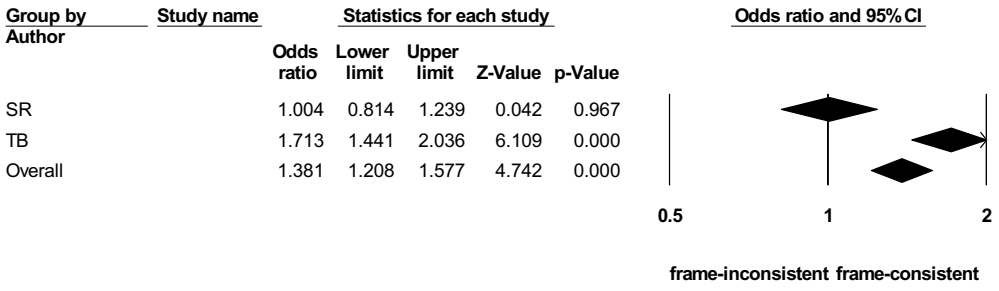

Figure 31. Subgroup analysis by authors

There is a marked contrast between the two groups, since, as Figure 31 shows, there is no overlap between the confidence intervals of the two groups. With the experiments by Steen et al., the summary effect size verges on 1, and the confidence interval of [0.814, 1.239] includes 1. This means that these experiments do not provide support for the research hypothesis that metaphors influence reasoning about crime. This group is quite homogenous in the sense that only about 14% of the observed variance reflects differences in the true effect sizes of the experiments ($I^2 = 14.221$). In contrast, the experiments conducted by Thibodeau and Boroditsky have a summary odds ratio significantly higher than 1 and exhibit a very high degree of consistency (about 9% of the observed variance is real variance in true effect sizes, $I^2 = 8.949$). Accordingly, this group of experiments seems to provide support for the hypothesis that metaphors influence thinking about crime.

A Q-test based on analysis of variance reinforces our impression that the two groups are different. Namely, the difference between the groups is statistically significant: $Q_{betw} = 14.833$, $df = 1$, $p = 0.0001$. A fully random analysis (in which both the experiments within the groups, as well as the two groups themselves are combined with the help of a random-effects model) produces similar results, except that the confidence intervals are, of course, wider. Therefore, we may conclude that *the variation of the true effect sizes pertaining to the first choice of participants might, to a large extent, be due to the different methods applied by the two groups of researchers*.

b) Sub-group analysis – by political affiliation

The variation in the effect sizes might be also due to factors which do not pertain to the peculiarities of the experiments as in the previous case, but to idiosyncrasies of subgroups within the participants in the experiments. Since political affiliation was one of the variables which were found to influence participants' preferences for crime reduction measures in some experiments by the researchers who conducted them, it seems reasonable to check its impact with meta-analysis tools, too.[103] Here again, a mixed-effects model seemed to be appropriate. Figure 32 summarizes the results.

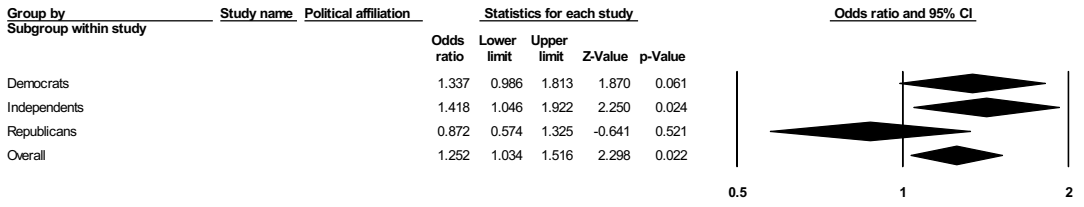

Figure 32. Subgroup analysis with political affiliation as a variable

As Figure 32 shows, there is a considerable overlap among the three confidence intervals. And in fact, the comparison of the three groups yields that the between-studies Q-value is 3.690 with 2 as a degree of freedom and a corresponding p-value of 0.158. This means that there are no substantial differences among the three groups. Furthermore, the within-group variance in effects is significantly greater than the degree of freedom in the case of the Democrats, indicating a great amount of dispersion in the true effect sizes in this subgroup of participants. As the corresponding I^2-statistics indicate, about 44% of the observed variance in effect sizes cannot be attributed to random error but reflects differences in the true effect sizes of the experiments in this subgroup. In contrast to the Democrats and the Republicans, in the case of the Independents, a mean effect size significantly above 1 was obtained. Therefore, the metaphorical frame seemed to influence only this group of participants.[104] To sum up, a subgroup-analysis based on the political affiliation of participants does not seem to be a good fit for the data.

D) Cumulative meta-analysis

Although the subgroup analysis by authors presented in Subsection C)a) indicated that both groups of experiments produced highly consistent results, Figures 33 and 34 reveal an interesting feature of these experiments. Namely, with the experiments conducted by Thibodeau & Boroditsky, the effect sizes *gradually decrease*. A cumulative meta-analysis reinforces this finding: if we calculate the summary effect size of these experiments stepwise in such a way that we always add an experiment and re-calculate the summary effect size, then we can see that it grows smaller over time.[105] See Figure 33.

---

[103] Christmann & Göhring (2016) does not include information about participants' political affiliations, thus this experiment is excluded from this analysis.

[104] We have to add that the results of the Democrats were marginally significant.

[105] Christmann & Göhring's (2016) exact replication attempt is omitted from this analysis.

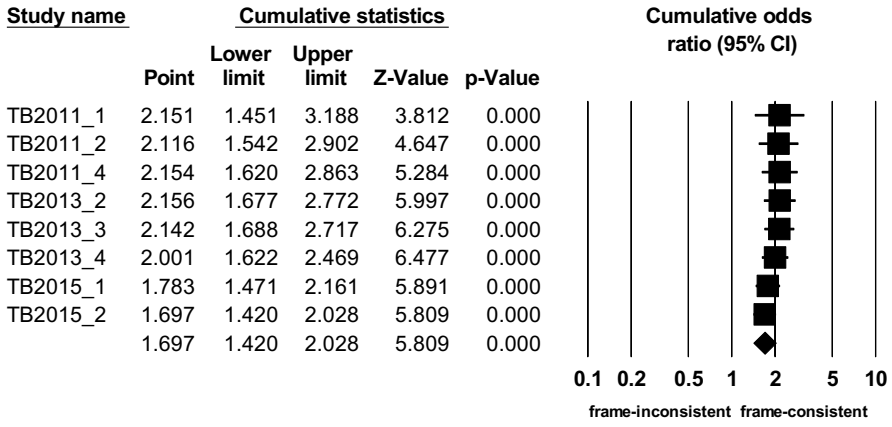

Figure 33. Cumulative meta-analysis – Thibodeau & Boroditsky

In contrast, as Figure 34 indicates, the cumulative summary effect size of the experiments conducted by Steen and his colleagues *increased almost continuously*:

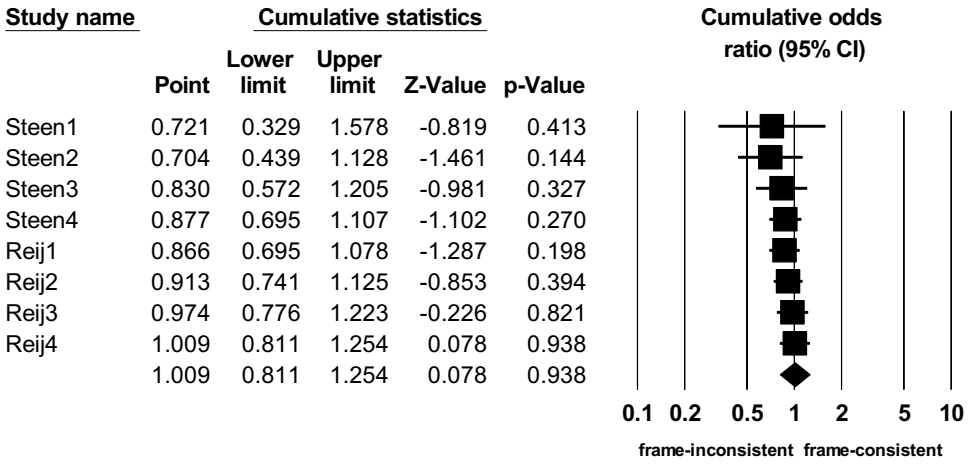

Figure 34. Cumulative meta-analysis – Steen and his colleagues

Nonetheless, it is important to remark that the experiments Reij1-4 did not follow each other in a chronological order nor are they improved versions of each other. Rather, they originate from the same experiments (as the 1-4 metaphor conditions) – that is, they should be regarded as one data point.

To sum up, this might mean that there is a slight tendency to convergence between the results of the two rival camps. If we try to identify the cause of these trends, it is not the temporal relationships among the experiments which seem to be decisive but rather changes in the methodology applied by the researchers. This hypothesis is supported by the finding that the exact replication of Thibodeau & Boroditsky's first experiment by Christmann and Göhring in 2016 yielded a higher effect size value than the original experiment.

E) The prediction interval

The 95% confidence interval of the summary effect size characterizes the precision of its estimate but does not provide information about the *amount of the dispersion of the effect sizes*. As we have discussed in Section 11.2.5, if we ask the question of whether *a new experiment* will have a true effect size falling between certain limits in 95% of the cases, then we have to calculate the *prediction interval*. In this case, the prediction interval is [0.749, 2.648], as indicated by the red line in Figure 30. This means that the true effect size for any similar *experiment* will fall into this range in 95% of the cases, provided that the true effect sizes are normally distributed (while the true *mean* effect size will fall into the confidence interval of [1.165, 1.703] in 95% of the cases). That is, on the basis of the information included in these experiments, one cannot predict whether a similar experiment would indicate any effect of the metaphorical frame – a weak effect or no effect are similarly possible.

F) Publication bias

Duval and Tweedie's trim and fill method[106] indicates no missing study. This means that there is no evidence for publication bias resulting from missing small experiments.

16.2.4. Alternative analyses

The diversity of the data handling techniques applied by the researchers might motivate alternative analyses. The analysis presented in Sections 16.2.2-3 took into consideration only the top choices of participants. Nonetheless, there are other possibilities. In this section, we will discuss two of them.

A) The rankings/ratings analysis

The rankings/ratings analysis takes the *rankings/ratings of the social reform-oriented vs. the enforcement-oriented measures* into consideration. That is, while for the first (top choices) analysis, we needed data about the orientedness (social reform vs. enforcement) of the top choices in the beast and in the virus frames, respectively, for the second analysis data are needed about the *whole range of the measures in the beast and in the virus frames*, respectively.

a) The choice of the effect size indicator

The experiments can be divided into three groups in terms of the information they contain about participants' evaluations of the measures. The data sheets belonging to Thibodeau & Boroditsky (2013, 2015) and Steen et al. (2014) contain data about the *ranking* of the measures; those by Reijnierse et al. (2015) include data about the *rating* of the measures; Christmann & Göhring (2016) applied an open question task, thus their answer sheets make it possible to count the *number of the social reform vs. enforcement-oriented answers* given by each participant. In order to calculate the effect of the metaphorical frames on the evaluation of the measures, we can compare

– the means of the *rankings* of the social reform type/enforcement-oriented measures in the virus vs. beast condition;

[106] See Section 11.2.7.

– the means of the *ratings* of the social reform-type/enforcement-oriented measures in the virus vs. beast condition;
– the means of the *number* of the social reform-type/enforcement-oriented measures in the virus vs. beast condition.

This data type motivates the use of the effect size indicator *standardized mean difference*, i.e. Cohen's *d*. Cohen's *d* is calculated in such a way that the difference of the sample means in the two conditions is divided by the within-group standard deviation pooled across conditions. In contrast to the odds ratio, the null-value (neutral value) is 0. That is to say, $d = 0$ indicates that there is no difference between the rankings/ratings/number of items in the two conditions (metaphorical frames). According to Cohen's recommendation, a *d* value of 0.2 indicates a small effect; 0.5 indicates a medium effect; 0.8 means a large effect. A negative value shows that there is an effect in the opposite direction – i.e. participants rank/evaluate frame-inconsistent measures higher.

b) Methods of data collection

As was the case with the first analysis presented in Sections 16.2.2-3, data had to be extracted and computed from the information in the data sheets.

i) In order to compare the whole *range* of the measures in the beast and in the virus frames, respectively, if a social reform type answer was ranked as 1, then the variable 's1' was assigned the value 5 (or 4, if there were only 4 options), otherwise it received the value 0. Similarly, if a social type answer was chosen in the second place, 's2' was assigned the value 4 (or 3, if there were only 4 options), but if the participant's second choice was an enforcement-type answer, then s2 received the value 0 – and so on. The $5^{th}/4^{th}$-ranked social type measure could have the value 1 or 0, depending on the last ranked answer of the given participant. As a second step, the variables s1-s4/5 were summarised as the variable 'social'. The highest value of this variable is 9 in the case of 5 response options (if the measures 'economy' and 'education' were chosen in the first and second places), while its lowest value is 3 (if these two measures were ranked as options No. 4 and 5).

Since the rankings of the enforcement-oriented are complementary to the social-reform type rankings, it is enough to take the variable 'social' into account for the calculation of the effect size.[107] Thus, in the case of the experiments by Thibodeau & Boroditsky (2013, 2015) and Steen et al. (2014), the effect size was calculated with the help of the mean and standard deviation of the variable 'social'[108] in the beast and virus conditions, respectively.

ii) As for the *ratings* of the measures in the beast and in the virus frames, respectively, the ratings of the social-reform oriented and enforcement-oriented measures are independent from each other in the sense that a participant may prefer one type of measure, or deem all measures equally effective, etc. Therefore, the ratings of the social-reform oriented and enforcement-oriented measures alike had to be taken into consideration by calculating the standardized mean

[107] I.e. in the case of a 4-member list of possible choices, the sum of the rankings of the social reform-type and the enforcement-type measures is 10, while with 5 response options the rankings add up to 15.
[108] See Section 2.2.

difference. Thus, 4 data sets were created in such a way that the mean and standard deviation of the *difference* of the social reform-oriented ratings and the enforcement-oriented ratings in the two metaphorical conditions (beast vs. virus frame) were calculated separately in the one-metaphor, two-metaphor, three-metaphor and four-metaphor conditions. This procedure allows us to calculate the SDM in 4 cases.

iii) With the *open question task* experiment, the number of the social reform vs. enforcement-oriented answers was captured by each participant. Then, the means and standard deviations of their *differences* were calculated for both conditions (metaphorical frames).

Table 45 and Table 46 in Appendix 2 present the mean rankings and the ratings/choices of the measures yielded by the data extraction process, respectively (standard deviations are in parentheses in both cases).

c) The effect size of the individual experiments

Figure 35 shows the individual effect sizes, their confidence intervals, Z-values, p-values, and weights.

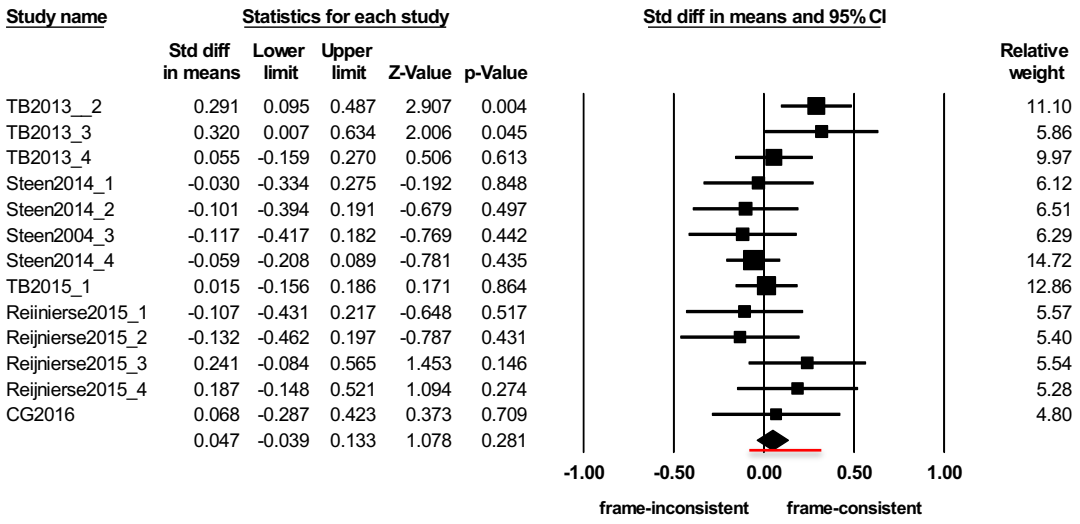

Figure 35. Effect sizes of the experiments and the summary effect size in the complex analysis

The standardized mean difference of the individual experiments ranged from -0.132 (Reijnierse et al., 2015, 2-metaphor condition) to 0.32 (Thibodeau & Boroditsky 2013, Experiment 3). In contrast to the top choices analysis, only 7 experiments out of 13 indicated an effect of the metaphorical frames, i.e., provided a positive SMD. This could suggest the opposite conclusion to the previous case. A decision on the basis of these pieces of information, however, would be unfounded, too. We have also to take into consideration that in the second analysis, there were only 2 experiments for which the confidence interval did not include the value 0. Thus, the majority of the experiments did not provide a significant result, and the confidence intervals ranged from -0.462 to 0.634, which yields a rather wide spectrum.

d) Synthesis of the results

As with the first (top choices) analysis, the application of the random effect model is appropriate in this case, too. As the last row of Figure 35 shows, the summary effect size of 0.047 is not significant; $Z = 1.078$, $p = 0.281$. Its confidence interval [-0.039, 0.133] includes the value 0, and overlaps with the majority of the confidence intervals of the individual experiments. This confidence interval is very narrow, indicating a very precise estimation of the summary effect. From these results we can conclude that the experiments together do not provide evidence for Thibodeau and Boroditsky's research hypothesis in this case.

As for the *consistency of the effect sizes*, the Q-value, i.e. the total amount of the observed between-experiments variance is 17.409. Its expected value is df(Q) = 12. These two values are not significantly different from each other, $p = 0.135$. This means that the total variation is not significantly greater than the sum of the within-study variations, suggesting that these experiments might share a common true effect size. The second relevant indicator is the estimate for the standard variation of the true effects, denoted as T^2. This is 0.007 in log units with a standard error of 0.01. This means that the standard deviation of the true effects, i.e. T, is 1.09. Finally, the I^2 value is 31.068, which means that about 31% of the observed variance in effect sizes cannot be attributed to random error but reflects differences in the true effect sizes of the experiments. This indicates a rather small amount of variation in the true effect sizes in this case.

The *prediction interval* is [-0.163, 0.256]. This means that the true effect size for any similar study will fall into this range in 95% of the cases, provided that the true effect sizes are normally distributed. That is, we may expect a result indicating no or a weak effect.

The first reaction to the discrepancy between the two analyses might be that the reason for the first analysis (top choices) yielding a higher summary effect size could be the fact that it includes 8 experiments by Thibodeau & Boroditsky, while the second analysis (ratings/rankings) only includes 4. Undeniably, this is a factor that has a major influence on the summary effect size. To wit, if we omit Experiments 1, 2, and 4 of Thibodeau & Boroditsky (2011) and Experiment 2 of Thibodeau & Boroditsky (2015) from the first analysis, a random effects model yields 1.260 as the summary effect size with a confidence interval of [1.019, 1.557]. But there is a second factor, too, which seems to be more interesting. Namely, if we compare the effect sizes of the individual experiments by transforming the odds ratios into standardized mean difference, we get the following picture. See Table 47.[109]

[109] In the case of the experiments in Reijnierse et al. (2015), the value of the complex analysis is computed as the average of the effect size of the social reform-type answers and the enforcement-type answers.

experiment	top choices	ratings/rankings
TB2013/2	0.426	0.291
TB2013/3	0.388	0.320
TB2013/4	0.251	0.055
Steen2014/1	-0.181	-0.030
Steen2014/2	-0.201	-0.101
Steen2014/3	0.048	-0.117
Steen2014/4	-0.053	-0.059
TB2015/1	0.120	0.015
Reijnierse/1	-0.139	-0.107
Reijnierse/2	0.233	-0.132
Reijnierse/3	0.301	0.241
Reijnierse/4	0.200	0.187
CG2016	0.450	0.068
Summary effect size	0.127	0.047

Table 47. Comparison of the effect sizes of experiments in the first (top choices) and second (ratings/rankings) analyses

The contrast is startling: the values in the second analysis are in most (although not all) cases considerably lower than in the first analysis. A possible reason might be that metaphors seem to be capable of slightly influencing people's initial reactions, but that when we take into account the whole spectrum of responses the impact of the metaphorical frames is substantially reduced, or even eliminated.

e) Subgroup analyses

Figure 36 summarizes the outcome of a *subgroup analysis by author* as a grouping variable.

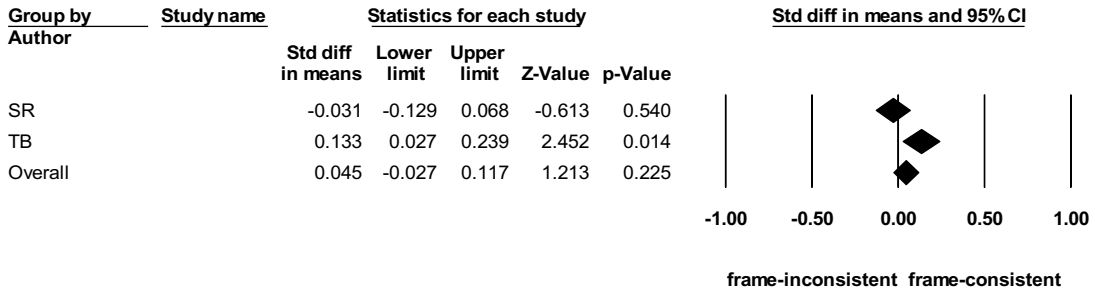

Figure 36. Subgroup analysis by authors – rankings/ratings analysis

As was the case with the first (top choices) analysis, in the second (rankings/ratings) analysis only the experiments conducted by Thibodeau & Boroditsky produce a standardized mean difference significantly higher than 0, and provide support for the research hypothesis. In this case, however, there is an overlap between the two confidence intervals. Nevertheless, the two groups are significantly different ($Q_{betw} = 4.915$, $df = 1$, $p = 0.027$). Furthermore, the within-group variances and the I^2 values indicate that the group of the experiments conducted by Steen et al. is more homogenous than it was in the previous case, while the experiments by Thibodeau

& Boroditsky are less homogenous. Namely, the I^2 value of the Steen et al.'s group is 0, while that of Thibodeau & Boroditsky is 36.707, indicating that about 37% of the observed variance reflects differences in the true effect sizes of the experiments.

In sum, if the whole ranking of the orientedness of the measures is taken into consideration, then the impact of the researchers' methods seems to be considerably weaker than it was in the first (top choices) analysis, but still remarkable. This finding should motivate further investigations. The results point to a search for further possibly relevant moderator variables and, accordingly, to corresponding between-participant subgroup analyses to test them.

A *subgroup analysis by the political affiliation of participants* yielded the following results (see Figure 37).

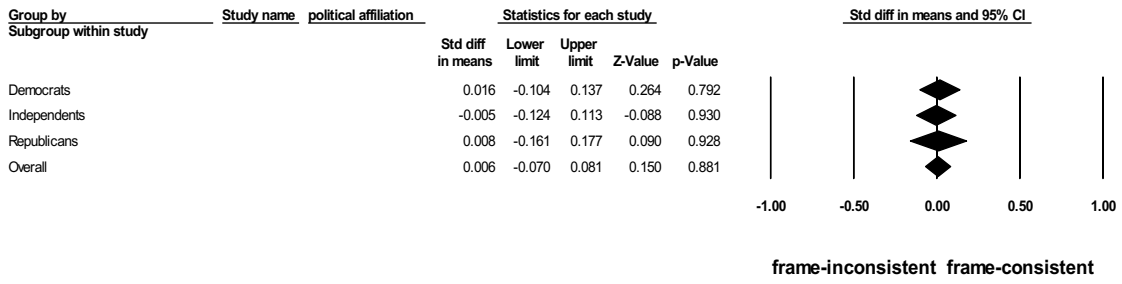

Figure 37. Subgroup analysis with political affiliation as a variable – rankings/ratings analysis

The between-groups Q-value is 0.063 with 2 as a degree of freedom, and a corresponding p-value of 0.969 in the random effects analysis. This means that *there are no substantial differences among the three political affiliations* in this case, either; the overlap among the three confidence intervals is huge.

f) Cumulative meta-analysis

As Figure 38 shows, in the case of the experiments conducted by Thibodeau & Boroditsky, a cumulative meta-analysis produces similar results to those produced in the first (top choices) analysis. Namely, there is a decrease in the effect sizes:

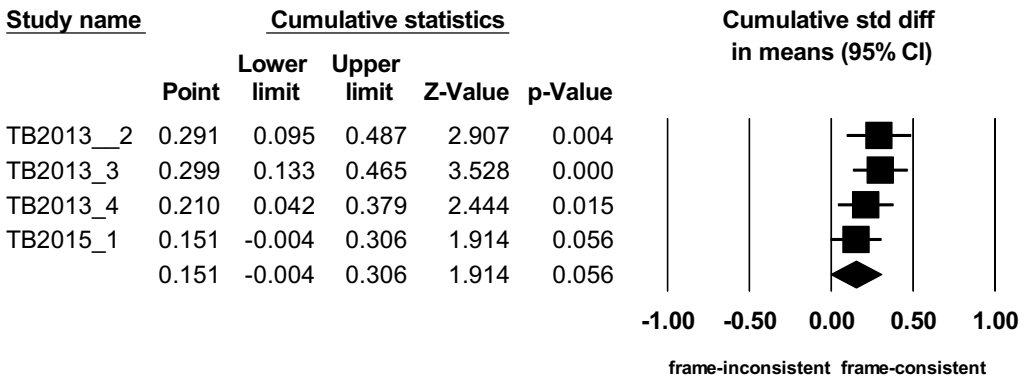

Figure 38. Cumulative meta-analysis – Thibodeau & Boroditsky

In contrast, the experiments by Steen and his colleagues show no clear tendency in the values. See Figure 39.

Study name	Point	Lower limit	Upper limit	Z-Value	p-Value
Steen2014_1	-0.030	-0.334	0.275	-0.192	0.848
Steen2014_2	-0.067	-0.278	0.144	-0.623	0.533
Steen2004_3	-0.084	-0.256	0.089	-0.952	0.341
Steen2014_4	-0.070	-0.182	0.043	-1.213	0.225
Reiinierse2015_1	-0.074	-0.180	0.033	-1.359	0.174
Reijnierse2015_2	-0.079	-0.180	0.022	-1.534	0.125
Reijnierse2015_3	-0.051	-0.147	0.046	-1.033	0.302
Reijnierse2015_4	-0.033	-0.125	0.060	-0.689	0.491
	-0.033	-0.125	0.060	-0.689	0.491

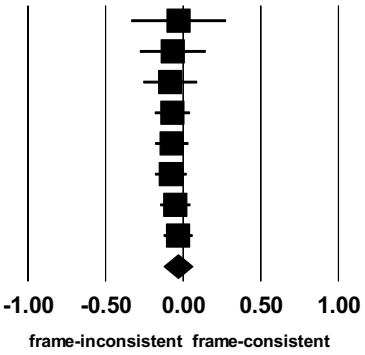

Figure 39. Cumulative meta-analysis – Steen and his colleagues

g) The prediction interval

The *prediction interval* is [-0.163, 0.258]. This means that the true effect size for any similar study will fall into this range in 95% of the cases, provided that the true effect sizes are normally distributed. Thus, the true effect size for any similar experiment will likely indicate either a weak reversed effect of the metaphorical frame, or more likely, a low effect.

h) Publication bias

Similarly to the first (top choices) analysis, Duval and Tweedie's trim and fill model indicates no missing study. To put it differently, the second (rankings/ratings) analysis seems to have estimated the true effect size correctly.

B) The measures analysis

A third possibility is to examine *the impact of the metaphorical frames on the measures separately*. Thus, the rankings/ratings of the five measures are investigated separately.

a) The choice of the effect size indicator

In the case of Thibodeau & Boroditsky (2013), Experiments 2-4, the rankings of the individual measures have to be collected. From this, we get a 2x5 (or 2x4) data matrix:

– mean of the rankings of the measures 'economy' / 'education' / 'patrols' / 'prison' / 'neighbourhood watches' in the beast condition;
– mean of the rankings of the measures 'economy' / 'education' / 'patrols' / 'prison' / 'neighbourhood watches' in the virus condition.

As for Reijnierse et al. (2015), the ratings of the individual measures could be directly averaged and compared in the two conditions.

This data type motivates the use of the effect size indicator *standardized mean difference*, i.e. Cohen's *d* in this case, too.

b) Methods of data collection

Table 48 in Appendix 2 present the mean and standard deviation of the rankings/ratings of the measures.

c) The effect size of the measures in the individual experiments

Table 49 summarises some relevant features related to the SMD of the individual experiments.

	economy	education	patrols	prison	watches
highest SMD	0.211	0.337	0.496	0.453	0.281
lowest SMD	-0.281	-0.267	-0.165	-0.272	-0.170
SMD higher than 0	5	5	7	6	4
number of signifi-cant results	0	1	3	1	0
smallest lower limit	-0.606	-0.566	-0.465	-0.572	-0.475
greatest upper limit	0.427	0.650	0.825	0.781	0.582

Table 49. Characterisation of the SMDs of the individual experiments in the third (measures) analysis

The most interesting finding is that the measure 'street patrols' has the highest value in all comparisons: it had an experiment with the highest SMD, with the least high lowest SMD, with the largest number of SMDs above 0; it had the largest amount of significant SMDs, and its lower and upper limits were the highest, too. Thus, it was the most popular measure. At the other extreme we find the measure 'economy'; its lowest values in almost all comparisons indicate that this was the participants' least popular choice.

d) Synthesis of the results

As Figure 40 shows, there is no substantial difference among the five measures; only the 'street patrols' measure shows a marginally significant effect of the metaphorical frame.

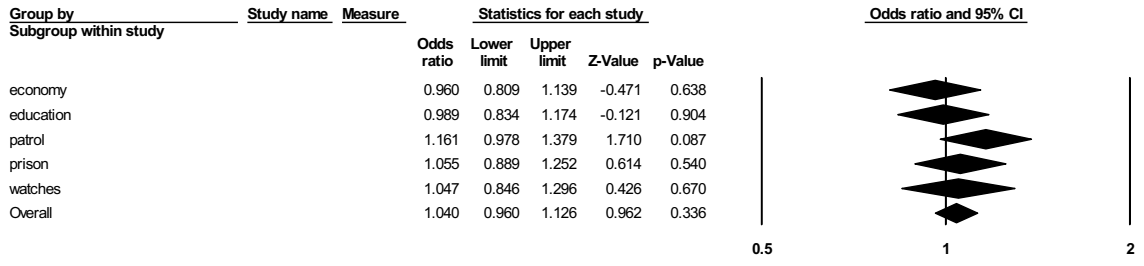

Figure 40. Effect sizes of the measures in the measures analysis

The *Q* statistics reinforce this impression: the difference between the measures is statistically not significant: $Q_{betw} = 2.792$, $df = 4$, $p = 0.593$.

16.2.5. Interim summary

Instead of a mechanical summary and comparison of the outcomes of the experiments belonging to an experimental complex, statistical meta-analysis offers a multifaceted evaluation of the available data:

(a) *In general:* The calculation of effect sizes with their 95% confidence intervals for each experiment makes it possible to compare the magnitude of the effect of one variable on another.
Specifically: The effect sizes of the individual experiments indicate that the impact of the frames (beast vs. virus) on the orientedness (social reform vs. enforcement) of the choices made by participants ranges from no effect to a significant weak effect.

(b) *In general:* With the calculation of the summary effect size, all pieces of information included in the individual experiments can be synthetized so that the shortcomings of individual experiments might be counterbalanced, and the results are more robust. The 95% confidence interval informs us about the precision of this estimate.
Specifically: The first analysis focused on the top choices of participants. It yielded a significant but weak effect of the metaphorical frame very precisely. The second analysis covered the whole ranking/rating of the measures. It yielded a lower summary effect size than the first analysis. As a further contrast, this result was not significant. The third analysis compared the effect of the metaphorical frames on the measures separately but found that they showed a similar pattern. To wit, the measures do not provide support for the research hypothesis.

This means that the results of the meta-analyses seem to take *a middle course* between the researchers' extreme evaluations of their findings. Steen and his colleagues stated that there is no, or only a minimal, effect. This is in accordance with the outcome of the second (rankings/ratings) analysis but in conflict with the first (top choices) analysis. In contrast, Thibodeau and Boroditsky (2011: 10) stated that there is a strong effect. This evaluation contradicts the results of all the meta-analyses we conducted. Finally, Thibodeau & Boroditsky's (2013: 21) more cautious formulation is in harmony with the outcome of the first (top choices) analysis but not with the second (rankings/ratings): "In sum, the results confirm that natural language metaphors can affect the way we reason about complex problems."

(c) *In general:* The prediction interval specifies where the true effect of a new experiment would fall in 95% of the cases. Thus, it informs us about the dispersion of the effect sizes.
Specifically: The prediction interval of the first and second analyses indicates that the true effect size for any similar experiment will indicate either a weak reversed effect of the metaphorical frame, no effect, or most likely, a low effect.

(d) *In general:* Subgroup analyses may reveal whether there are subgroups among the experiments indicating some methodological or other differences, or there are subgroups among participants which behave differently.
Specifically: Both in the first and the second analyses, a moderate amount of heterogeneity was found. Subgroup analyses identified one possible cause of this finding: namely, the variation in the true effect sizes seems to be due to a considerable extent to the different methods applied by the two groups of researchers. Namely, while Thibodeau and

Boroditsky applied open questions or used only the top choices of participants, Steen and his colleagues took either the first two responses into consideration or they applied Likert-type scales.[110] Further, the formulation of the task of participants was modified by the researchers many times. The contrast between the two groups of experiments was considerably sharper in the case of the first analysis, which used experiments with a broader range of data eliciting techniques. Our results suggest that further, finer details of data processing, such as the application of open vs. closed questions, the exact formulation of the task, or the usage of rankings or ratings, etc. might turn out to be relevant factors, too. Conversely, the political affiliation of participants did not influence the results.

(e) *In general:* Performing a cumulative meta-analysis enables us to check whether the effect size is affected by some factor. For this end, first we have to arrange the experiments into a sequence based on this variable. Then, we have to add the experiments one after another, re-calculate the summary effect size again and again, and compare them in order to find out whether there is a tendency in the values.
Specifically: Cumulative meta-analyses showed that if experiments are sorted chronologically, then the effect sizes in 3 of 4 cases converge towards the summary effect size. We raised the hypothesis that this might be due to the changes in the stimulus materials, and the tasks participants had to perform.

(f) *In general*: If researchers conducting the experiments make their data sets public, there is room for more exact, deeper analyses, as well as re-analyses.
Specifically: Raw data included in the data sheets made public by the researchers enabled us to calculate the effect sizes more precisely than on the basis of summary data presented in the experimental reports. Further, we were able to conduct and compare three different analyses (top choices, rankings/ratings, measures), so that the diversity of the methods of data processing adopted could to some extent be controlled for. Nonetheless, the impact and theoretical consequences of the application of diverse data processing methods should motivate further research.

Nonetheless, some *limitations* have to be imposed on our results. First, we made use of statistical meta-analysis in an unorthodox way, because we applied it to a debate between two parties and did not conduct a thorough search for further experiments testing the same research hypothesis in the literature. This necessitates the extension of the set of experiments analysed by further studies. Second, while statistical meta-analysis is an indispensable tool for summarising and synthesizing the results of (sufficiently) similar experiments, its resources for revealing (systematic) errors present in the experiments at issue are limited. Third, with the help of statistical meta-analysis, some inconsistencies among experiments could be resolved. Therefore, it is an effective method of problem-solving. At the same time, however, it also led to the emergence of new problems.

[110] Nonetheless, it is important to mention that Steen et al. (2014: 15ff.) also present an analysis of the top ranked solutions in their *Alternative analyses* section.

16.3. The combination of the two methods

Our analyses in Section 16.1 and Section 16.2 illustrate that the methods based on cyclic re-evaluation and statistical meta-analysis *complement each other*: they reveal problems and prospects which cannot be dealt with the help of the other method. Therefore, their relationship is built not on rivalry but on cooperation:

– While the model based on cyclic re-evaluation deals mostly with problems related to the experimental design, statistical meta-analysis might provide tools to check whether they in fact influence the results.
– The method of cyclic re-evaluation is also vital in judging whether the experiments at issue fulfil some basic requirements and can be regarded as reliable data sources. Statistical meta-analysis must not include experiments which are not capable of producing plausible experimental data.
– Statistical meta-analysis may counterbalance errors present in one subgroup of experiments but cannot identify and eliminate problems burdening all or most experiments. Therefore, it could be fruitfully complemented by analyses aimed at identifying possible error sources in the experiments, such as the reconstruction of the relationship among the experiments and their replications with the help of the concept of the ‘experimental complex’ as presented in Section 16.1.
– Statistical meta-analysis may not only check whether certain factors are relevant or not with the help of sub-group analyses or meta-regression, it is also capable of treating one of the most acute problems of experiments in cognitive linguistics: the low power resulting from the small number of participants and replications.
– As we have seen in Section 16.2, statistical meta-analysis can also contribute to a deeper analysis of the interpretation of the data by investigating different versions of the relationship between the perceptual data and the experimental data.

From this it follows that statistical meta-analysis has to be integrated into a more comprehensive model of the evaluation of the replication of experiments in which its results can motivate new directions of research in order to find novel solutions to problems.

17. Conclusions: Experimental data as evidence for/against theories and a possible resolution of (PET)

From the considerations presented in Section 16.3 we can conclude that the contradiction suggested by the Paradox of Error Tolerance is only apparent. Namely, *the gradual elimination of problems does not mean rigid problem-intolerance, while statistical meta-analysis is not based on an uncritical problem-tolerance*. Therefore, the task is to find the balance between the requirements of comprehensiveness and perfection by uniting the virtues of the two methods.

18. Results

In Section 1, we formulated the main problem this book centres on as follows:

(GP) (a) How can the uncertainty of data be treated in cognitive linguistics?
(b) What are the methods of inconsistency resolution in cognitive linguistics?
(c) Which guidelines should govern the evaluation of theories in cognitive linguistics?

Then, we decided to narrow down the general (GP) to a progressive, prototypical empirical data type within cognitive linguistic research, to *experimental data*. This yielded the special (SP):

(SP) (a) How can the uncertainty of experimental data be treated in cognitive linguistics?
(b) What are the methods of the treatment of inconsistencies emerging from conflicting results of experiments in cognitive linguistics?
(c) Which guidelines should govern the evaluation of theories with respect to experimental results in cognitive linguistics?

We devoted the three parts of the book to the three sub-problems of (SP). On our way to finding a solution to the subproblems, we had to face paradoxes, whose resolution was a prerequisite to providing a solution to (SP). We presented several case studies to illustrate the workability of the proposed metascientific models.

18.1. A solution to (SP)(a)

In Part I, we presented a metascientific model of experiments as well as series of closely related experiments with the help of which the uncertainty of experimental data can be described and a method for its treatment can be put forward. Its application yields the following resolution to (SP)(a):

(RP) (a) (1) Experiments in cognitive linguistics are, like experiments in science, not completely reliable data sources. The uncertainty of experimental data results basically from the *inherent fallibility* of the components of the experimental process.
(2) The uncertainty of experimental data can be explicated as their degree of acceptability/unacceptability on the basis of the peculiarities of the experiment from which they originate, that is, *as plausibility/implausibility*. These plausibility values may range from falsity with certainty through neutral plausibility to truth with certainty, while the two endpoints of the scale representing rather theoretical than real possibilities.
(3) The plausibility of the experimental data is *a function of the plausibility of statements related to* the *components of the experimental process* such as the experimental design, the experimental procedure, the authentication and inter-

pretation of the perceptual data, and the presentation of the results. Although it is not possible to subsequently reconstruct all relevant details of the experimental process and give an accurate estimate of the plausibility of all related statements and their contribution to the plausibility of the experimental datum, revealing the sources from which the plausibility of the perceptual data, theoretical hypotheses, background assumptions etc. in the experimental process and in the experimental report originate is pivotal. Although their impact cannot be determined perfectly, it is decisive to find out whether they increase or decrease (or even seriously question) the plausibility of the resulting experimental data.[111]

(4) Experiments are *cyclic processes*: the plausibility of the statements related to different stages of the experimental process is re-evaluated again and again during the elaboration and conduct of the experiment; conflicts between diverging evaluations are revealed and attempts are made to resolve them.

(5) Experiments are *open processes*: the evaluation of experiments in nothing other than the continuation of the cyclic process of re-evaluation by another researcher(s) by new plausible argumentation cycles, and, if possible, the elaboration of proposals for its continuation by new experimental cycles. This involves the reconstruction of the stages of the experimental process, conduct of thought experiments, the identification of problems, the re-evaluation of the plausibility value of the experimental data, as well as proposals for the possible resolution of the open problems.

(6) We *do not have direct access* to the components of the experimental process but have to reconstruct them from the experimental report and the additional materials made public by the researchers having conducted the experiment. Thus, experiments have a *dual argumentative structure*: the experimental process is organised by a non-public plausible argumentation process that is then transformed into the experimental report, i.e. into a public piece of plausible argumentation.

(7) The re-evaluation of experiments includes the comparison of the reconstructed version of the plausible argumentation process organising the experimental process with the experimental report. An overestimation of the plausibility value of the experimental data in the experimental report is a grave error. *Transparency* is key: the experimental report and the additional materials should contain all information that might be relevant for the evaluation of the steps of the experimental process. In this way, the reader can be made *a virtual participant of the creation, analysis and evaluation of the experimental data*.

[111] Plausible inferences may be enthymematic which means that they may have missing premises. Despite this, the amount of information which cannot be found in the experimental report and additional materials is too large. Therefore, it is more appropriate to treat experimental data as 'data' in the sense of our definition (D) in Section 14.1 than as plausible statements whose plausibility value originates from indirect sources (plausible inferences). See also Section 14.1 on this.

(8) Experiments are mostly not isolated entities but parts of *experimental complexes* consisting of exact and non-exact replications, control and counter-experiments. Non-exact replications may lead to more plausible experimental data. The increasing plausibility results from the successes in the problem-solving process and/or the refinement of the research hypothesis. While non-exact replications can eliminate identified problems, exact replications may secure the reliability of the results. Methodological variants (i.e. experiments belonging to other experimental complexes but investigating the same variables) may increase the plausibility of the experimental data, too. 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. 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.

(9) With the help of the concepts of 'progressivity', 'limit', 'convergence', and 'efficacy', it is possible to *describe the progress of the problem-solving process and evaluate its current state*. Nevertheless, it is important to bear in mind that new information may require a revision of our earlier decisions. Therefore, convergence and efficacy can be judged only temporarily, relative to our current knowledge.

(10) The plausibility of experimental data dynamically changes due to the exact and non-exact replications, control and counter-experiments and methodological variants. Therefore, the uncertainty resulting from the inherent fallibility of the experimental process can only be effectively reduced through the replications and revisions, that is, by improvements and experiments conducted by the research community. This means that an important change of view is necessary: experiments should be regarded as *collective works of a research field* and not private affairs of single minds. This requires, above all, openness, transparency and cooperativeness.

In relation to (SP)(a), we raised the following paradox:

(RPE) *The rhetorical paradox of experiments in cognitive linguistics*:
The reliability of experiments as data sources in cognitive linguistics is both directly *and* inversely proportional to the rhetoricity of the experimental report.

Clarification of the role of argumentation in relation to experiments in cognitive linguistics was a decisive point in the elaboration of our metascientific model since it motivated the involvement of argumentation theoretical tools in the model. As we have seen, argumentation plays a more significant role in experiments in linguistics than in science. On the basis of the metascientific model presented in Part I, in Section 7 we put forward *a resolution of the Rhetorical Paradox of Experiments*, which *couples the reliability of experiments as data sources to the*

effectiveness of the plausible argumentation process, mirroring the successfulness of the problem-solving process.

This resolution of (RPE) provides the précis of the train of thought presented in Part I insofar as it highlights the importance of a radical change of view in relation to experiments in cognitive linguistics. The evaluation of experiments should not be reduced to an isolated, theory-guided and unsystematic judgement of the experimental report but should systematically reveal and analyse the inner structure of the experiments and their outer relations to other experiments. While absolute objectivity is an unrealistic aim, and experimental data cannot be seen as "hard facts", the thoroughness of the re-evaluation of as many details of the experimental process as possible is decisive. Thus, the argumentation presented in the experimental report and the additional materials should assist the deep and methodical analysis of the experimental process, which is nothing other than the continuation of the argumentation process with new cycles by other researchers.

18.2. A solution to (SP)(b)

In Part I, we described experiments and series of closely related experiments in cognitive linguistics as problem-solving processes. As the case studies we presented exemplified, non-exact replications are often capable of ruling out possible systematic errors, and methodological variants (that is, experiments making use of different techniques but investigating the same research hypothesis) may further increase the plausibility value of the experimental data at issue by raising their reliability and validity. Exact and non-exact replications and methodological variants, however, often produce conflicting results or lead to the emergence of new problems. This yielded the Paradox of Problem-Solving Efficacy:

(PPSE) Non-exact explications and methodological variants are
- (a) *effective tools of problem-solving* in cognitive linguistics because by resolving problems they lead to more plausible experimental results; and they are also
- (b) *ineffective tools of problem-solving* because they trigger cumulative contradictions among different replications and methodological variants of an experiment and lead to the emergence of new problems.

In order to resolve (PPSE), in Part II we extended our metascientific model so that it provides us with tools for describing the emergence, function and the treatment of inconsistencies. On the basis of our considerations and case studies, we propose the following resolution to (SP)(b):

(RP) (b) (1) Inconsistencies related to experiments in cognitive linguistics are mostly not fatal failures which would require the immediate rejection of the experiments at issue. *Inconsistencies and, more generally, problems are one of the major driving forces of experimental work*, because they motivate the elaboration of non-exact replications, control and counter-experiments as well as methodological variants, and often prompt researchers to elaborate more refined theories.

(2) Conflicts among experiments can be explicated as *p-inconsistencies*: a hypothesis is made plausible by one (series of) experiment(s) as a source, while another (series of) experiment(s) makes it implausible. P-inconsistencies may emerge within an experimental complex between exact/non-exact replications or experiments and counter-experiments, as well as between results of experiments belonging to different experimental complexes such as methodological variants. P-inconsistencies cannot be resolved with the help of a mechanical comparison of the plausibility values of the conflicting statements but they have to be resolved *in the context of all related experiments, that is, by the re-evaluation of the problem-solving process(es)*.

(3) The re-evaluation of the problem-solving process(es) has to involve *future prospects* as well. That is, it should not be a static snapshot of the current state of the experimental complex(es). Rather, it should be a *dynamic analysis* of the development of the relationship among a series of related experiments, which involves the search for starting points for the elaboration of new, more refined non-exact replications which might be free of problems according to our current knowledge.

(4) The resolution of inconsistencies is guided by *problem-solving strategies*. The first strategy (Contrastive Strategy) involves the separate continuation of the conflicting chains of experiments by conducting further non-exact replications, counter- or control experiments, a systematic confrontation and comparison of the results and it may lead to a decision if a limit (valid and reliable experiment) has been reached by one of the conflicting series of experiments while the other series gets stuck. The second strategy (Combinative Strategy), in contrast, is based on *a refinement of the research hypothesis and experimental design in such a way that all factors found relevant so far are taken into consideration*. In this way, it keeps both members of the conflict in order to integrate them and provide a more comprehensible picture and avoid information loss.

(5) One method of inconsistency resolution consists of *the reconstruction and judgement of the effectiveness of the related cyclic process(es) of problem-solving* put forward in the experimental reports. This means that the first step is the *reconstruction of the structure of the experimental complex*: one has to identify the limit-candidates as well as the chains of non-exact replications, control- and counter-experiments which produced them. The second step consists of *re-evaluating the problem-solving process* within the chains of experiments, and the comparison of them. One has to take *the number, seriousness and resolvability of all problems burdening the experiments belonging to the experimental complex(es)*. We can differentiate between *progressivity* which is a local characteristic of non-exact replications and means that a problem of the predecessor has been solved, and *effectiveness* which is a global feature and means that the problem-solving process reached a limit, that is, an experiment which is valid and reliable on the basis of the information at our disposal. The re-evaluation process mostly does not terminate because we are usually

not in a position to decide about its effectiveness but can only propose new versions of the experiments which have to be conducted and evaluated in future. Thus, if the p-inconsistencies cannot be resolved on the basis of the information at our disposal because no limit has been reached, then the third step should be the *determination of the directions of the continuation of the cyclic process of re-evaluation*.

(6) Another method of inconsistency resolution is based on *statistical meta-analysis*. Statistical meta-analysis provides us with tools for *combining the results of a series of experiments* conducted in the past. The calculation of the summary effect sizes synthesises the whole range of the available information, yielding considerably more reliable and accurate results than single experiments could. Additional analyses may provide information about the precision of these estimates (confidence intervals) and their dispersion (prediction intervals). Heterogeneity analyses, subgroup analyses and meta-regression can be applied in order to find out whether the results of the experiments are consistent or there are subgroups among the experiments indicating some methodological or other differences, or subgroups among participants which behave differently. This also means that experiments with a significant value and ones indicating a non-significant result are not necessarily in conflict, because they may indicate a similar effect size, and their confidence intervals may overlap to a great extent.

As we have shown in Section 12, both metascientific models we proposed as tools of inconsistency resolution yield *a resolution to (PPSE)* as well. The metascientific model based on the reconstruction and re-evaluation of the cyclic process of problem-solving offers us the concepts of ‘progressivity’ and ‘limit’ with the help of which the exact differences between effective and ineffective problem-solving processes can be stipulated. The use of statistical meta-analysis makes it possible, too, to distinguish between efficacy and inefficacy. The problem-solving process is effective if there is a high enough number of non-exact replications and methodological variants at our disposal, and the results of the heterogeneity analyses conducted either indicate consistency among the experiments, or the causes of the heterogeneity can be identified and they result in the refinement of the research hypothesis.

In both cases, it is important to emphasise that effectiveness can be judged only in the long run. Decisions are not final but only provisional: new pieces of information can overrule earlier decisions. Thus, a non-exact replication can turn out to be problematic and lose its limit-status, and the addition of new experiments may modify the summary effect sizes and the results of the heterogeneity analyses, subgroup analyses or meta-regression. This means that both methods interpret experiments and experimental complexes as *open processes* and suppose that there are no experiments whose results were final, and immune to revision or improvement.

18.3. A solution to (SP)(c)

The evaluation of theories with respect to experimental results in cognitive linguistics involves three tasks: one has to draw predictions from the rival theories, summarise the results of a series of experiments, and make a principled decision about the predictions/theories on the basis of the experimental results. As our case studies exemplified, all three endeavours are highly problematic and require the elaboration of clear guidelines. On the basis of our analyses put forward in Part III, we propose the following solution to (SP)(c):

(RP) (c) (1) The relationship between single experiments and theories can be modelled with the help of *three different concepts of 'evidence'*. A common characteristic of the three concepts is that the connection between the data and the theories is established by plausible inferences which make use of the given datum as one of their premises and lead to predictions drawn from the theories or their negation as their conclusions. The three concepts of 'evidence' cover three basic constellations. A datum may provide support for the theory because there is a plausible inference connecting the datum and the prediction(s) of the theory, independently of whether it is possible to build a similar inference between the datum at issue and the rival theory ('weak evidence'). 'Relative evidence', in contrast, is a comparative concept: it describes scenarios in which a datum provides weak evidence for the predictions of both theories but supports one of them significantly more strongly. Finally, 'strong evidence' means that a datum differentiates between the rival theories even to a greater extent because it provides weak evidence for the predictions of one theory and against the other – that is, it makes one of them plausible while its rival becomes implausible. Nonetheless, since experimental data are not true with certainty but only plausible, they cannot prove or falsify (that is, demonstrate the truth or falsity of) theories. Consequently, there is no such a thing as "experimentum crucis" in the sense that no experiment is capable of warranting a final decision among rival theories.

(2) Drawing predictions from cognitive linguistic theories is a highly complicated task. In order to produce a prediction, *thought experiments* have to be carried out so that a strong enough link can be established between hypotheses of the theory, peculiarities of the relevant linguistic phenomena, and linguistic behaviour under certain circumstances. Attempts at performing such thought experiments often reveal that the concepts used in cognitive linguistic theories are not defined properly and they are in need of *explication* so that their meaning becomes clear.

(3) Predictions stipulate a state of affairs which should occur under certain well-specified circumstances. This requires a connection among the high-level theoretical concepts of the theory, lower-level theoretical constructs (phenomena) and perceptible/measurable manifestations of linguistic behaviour. That is, one has to *operationalise* the theoretical concepts. Further, the use of statistical meta-analysis requires more detailed predictions because it not sufficient to

confront summary effect size data with predictions stating solely the presence or absence of an effect; one has to quantify to some extent the strength of the effect a variable should have on some other variable(s).

(4) We presented two methods for *summarising the results of a series of closely related experiments*. The first method is based on the *reconstruction and cyclic re-evaluation of the related experiments as well as the progressivity and effectiveness of the related problem-solving process*. Non-exact replications in most cases are progressive because they solve at least a problem of their predecessor but do not necessarily increase the plausibility value of the experimental data. Further, due to the emergence of new problems and the unsolved problems burdening the original experiment and its successors alike, it is possible that an experimental complex does not produce a limit, that is, an experiment which would be valid and reliable. In such cases, the best choice may be the continuation of the problem-solving process and the elaboration and conduct of more refined versions of the experiments. Nonetheless, if the experimental data resulting from the least problematic member of the experimental complex can be deemed to be plausible, it can be used as evidence for or against the related experiments. Indeed, one has to keep in mind that this decision can be only provisory and is fallible. The other method for summarising the results of similar experiments is *statistical meta-analysis*. Although the calculation of the summary effect size is based on relatively exact and clearly applicable rules, the reliability of the outcome of statistical meta-analyses depends heavily on the number and plausibility of the summarised experimental data.

(5) Confronting experimental results with predictions gives rise to several difficulties, too. First, statistical meta-analysis yields more fine-grained results than the customary practice of hypothesis testing. While the latter provides a dichotomy of significant vs. non-significant results, the former produces effect size values. Therefore, we proposed the application of a 6-point scale of 'reverse large – reverse moderate – reverse small – no effect – small – moderate – large effect' for predictions. At this point, there are still two possibilities: a stricter requirement which stipulates that the summary effect size has to fall into the predicted category, otherwise it counts as weak evidence against the prediction, or a more permissive which tolerates a one-point difference between predictions and summary effect sizes by creating a neutral zone. Second, further difficulties may arise in cases in which we have not sole predictions but a compound of several related predictions.

Nonetheless, the two methods for combining the results of similar enough experiments seemed to be based on contradictory assumptions, yielding the *Paradox of Error Tolerance*:

(PET) When determining the strength of support provided by an experimental complex to a hypothesis/theory,

(a) *the elimination of errors is top priority*, because it is the detection and elimination of problems which makes experiments more reliable data sources;
(b) *the elimination of errors is not top priority*, because comprehensibility, that is, the involvement of all relevant experiments and the accumulation of all available pieces of information should be ranked higher.

As we have seen in Section 17, this paradox can be resolved easily if we realise that the two methods are complementary, and should best be used in parallel. They shed light on the experiments at issue from different angles and provide information which is unavailable by using the other method, making it possible to achieve a comprehensive, well-founded and balanced evaluation of the related experiments. First, the method of cyclic re-evaluation can be used to check whether the experiments at issue produce plausible experimental data so that statistical meta-analysis can be applied to them. Second, while the model based on cyclic re-evaluation is capable of revealing problems related to the experimental design, statistical meta-analysis provides us tools to check whether they in fact influence the results or are only minor problems that can be counterbalanced. Nonetheless, one needs a high number of experiments from a longer time span which were conducted by different researchers for this, and it has to be checked whether the problem at issue is only present in a few experiments but not in all or most of them. To put it otherwise, statistical meta-analysis may counterbalance errors present in one subgroup of experiments but cannot identify and eliminate problems burdening the majority of experiments. Therefore, it should be fruitfully complemented by analyses aimed at identifying possible error sources in the experiments, such as the reconstruction of the relationship among the experiments and their replications with the help of the concept of the 'experimental complex'. Statistical meta-analysis is also capable of treating one of the most acute problems of experiments in cognitive linguistics: the low power resulting from the small number of participants. A severe setback to its application is, however, the low number of exact and non-exact replications and methodological variants. Finally, the model based on cyclic re-evaluation of the problem-solving process is strongly future-oriented insofar as it can effectively contribute not only to a deeper analysis of the current state of experimental work but also it motivates new directions of research. Nonetheless, statistical meta-analysis also has resources which can facilitate the search for novel solutions to older and newer problems in cognitive linguistics. The strictness and thoroughness in the analysis of experiments, the elaboration of control experiments, counter-experiments, methodological variants and non-exact replications 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 as collaborative works carried out with the participation of researchers belonging to different theoretical backgrounds or even to rival approaches.

REFERENCES

Andor, J. (2004): The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics* 1, 93-111.

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

Archangeli, D. (1997): Optimality Theory: An introduction to linguistics in the 1990s. In: Archangeli, D. & Langendoen, T. (eds.): *Optimality Theory. An overview*. Oxford: Blackwell, 1-33.

Aristotle: *Metaphysics*. Book 4/3.
http://ebooks.adelaide.edu.au/a/aristotle/metaphysics/book4.html#book4

Arzouan, Y., Goldstein, A., & Faust, M. (2007): Brainwaves are stethoscopes: ERP correlates of novel metaphor comprehension. *Brain Research* 1160, 69-81.

Bambini, V., Resta, D., & Grimaldi, M. (2014): A dataset of metaphors from the Italian literature: Exploring psycholinguistic variables and the role of context. *Plos One* 9(9), e105634. doi:10.1371/journal.pone.0105634.

Blasko, D.G. & Connine, C.M. (1993): Effects of familiarity and aptness on metaphor processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19(2), 295-308.

Bogen, J. (2002): Experiment and observation. In: Machamer, P. & Silberstein, M. (eds.): *The Blackwell guide to the philosophy of science*. Malden & Oxford: Blackwell, 128-148.

Bogen, J. & Woodward, J. (1988): Saving the phenomena. *The Philosophical Review* 97(3), 303-352.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., Rothsteind, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods* 8, 5–18. doi: 10.1002/jrsm.1230.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., Rothstein, H. R. (2009): *Introduction to meta-analysis*. Chichester: John Wiley & Sons.

Borsley, R.D. (2005): Introduction. *Lingua* 115, 1475-1480.

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.

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

Brisand, F., Frisson, S., & Sandra, D. (2001): Processing unfamiliar metaphors in a self-paced reading task. *Metaphor and Symbol* 16(1-2), 87-108.

Caillies, S. & Declercq, C. (2011): Kill the song – steal the show: What does distinguish predicative metaphors from decomposable idioms? *Journal of Psycholinguistic Research* 40(3), 205-223.

Campbell, S.J. & Raney, G.E. (2016): A 25-year replication of Katz et al.'s (1988) metaphor norms. *Behavior Research Methods* 48, 330-340.

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

Cantor, G. (1989): The rhetoric of experiment. In: Gooding, D., Pinch, T. & Schaffer, S. (eds.): *The Uses of Experiment*. Cambridge: Cambridge University Press, 159-180.

Cardillo, E.R., Schmidt, G.L., Kranjec, A., & Chatterjee, A. (2010): Stimulus design is an obstacle course: 560 matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior Research Methods* 42(3), 651-664.

Cardillo, E.R., Watson, C., & Chatterjee, A. (2017): Stimulus needs are a moving target: 240 additional matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior Research Methods* 49(2), 471-483.

Chang, H. (2011): Beyond case studies: History as philosophy. In: Mauskopf, S. & Schmaltz, T. (eds.): *Integrating history and philosophy of science*. Netherlands: Springer, 109-124.

Chiappe, D.L. & Kennedy, J.M. (1999): Aptness predicts preference for metaphors or similes, as well as recall bias. *Psychonomic Bulletin & Review* 6(4), 668-676.

Chiappe, D.L. & Kennedy, J.M. (2001): Literal bases for metaphor and simile. *Metaphor and Symbol* 16(3-4), 249-276.

Chiappe, D.L., Kennedy, J.M., & Chiappe, P. (2003): Aptness is more important than comprehensibility in preference for metaphors and similes. *Poetics* 31, 51-68.

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

Chomsky, N. (1965): *Aspects of the theory of syntax*. MIT Press, Cambridge.

Chomsky, N. (1969 [1957]): *Syntactic structures*. Mouton, The Hague & Paris.

Chomsky, N. (1969): Language and philosophy. In: Hook, S. (ed.): Language and philosophy: A symposium. New York: New York University Press, 51-94.

Chomsky, N. (1980): On binding. *Linguistic Inquiry* 11, 1-46.

Christmann, U. & Göhring, A.-L. (2016). A German-language replication study analysing the role of figurative speech in reasoning. *Scientific Data* 3, 160098. https://doi.org/10.1038/sdata.2016.98.

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

Consten, M., Loll, A. (2010): Circularity effects in corpus studies – why annotations sometimes go round in circles. *Language Sciences* 34(6), 702-710.3.

Cumming, G. (2012). *Understanding the new statistics. Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Deignan, A. (2008): Corpus linguistics and metaphor. In: Gibbs, R.W. (ed.): *The Cambridge handbook of metaphor and thought*. Cambridge: Cambridge University Press, 280-294.

Dulcinati, G., Mazzarella, D., Pouscoulous, N., & Rodd, J. (2014): Processing metaphor: The role of conventionality, familiarity and dominance. *UCL Working Papers in Linguistics* 26, University College London.

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

Franklin, A. (2009): Experiments. Stanford Encyclopedia of Philosophy. *http://plato.stanford.edu/entries/physics-experiment/*.

Gagné, C.L. (2002): Metaphoric interpretations of comparison-based combinations. *Metaphor and Symbol* 17(3), 161-178.

Galison, P. (1987): *How experiments end*. Chicago: Chicago University Press.

Geeraerts, D. (2006): Methodology in cognitive linguistics. In: Kristiansen, G., Achard, M., Dirven, R. & de Mendoza Ibáñez, F.J.R. (eds.)(2006): *Cognitive Linguistics: Current Applications and Future Perspectives*. Berlin & New York: de Gruyter, 21-49.

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

Gentner, D., Bowdle, B. (2008): Metaphor as structure-mapping. In: Gibbs, R.W. (ed.): *The Cambridge handbook of metaphor and thought*. Cambridge: Cambridge University Press, 109-128.

Gentner, D., Bowdle, B., Wolff, P., Boronat, C. (2001): Metaphor is like analogy. In: Gentner, D., Holyoak, K.J., Kokinov, B.N. (eds.): *The analogical mind: Perspectives from cognitive science*. Cambridge: MIT Press, 199-253.

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., Lonergan, J.E. (2007): Identifying, specifying and processing metaphorical word meanings. In: Rakova, M., Pethő, G., Rákosi, Cs. (Eds.), The cognitive basis of polysemy. New sources of evidence for theories of word meaning. Frankfurt a.M. & New York & Oxford: Peter Lang, 71-90.

Gibbs, R.W. (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.

Giere, R.N. (2011): History and philosophy of science. Thirty-five years later. In: Mauskopf, S. & Schmaltz, T. (eds.): *Integrating history and philosophy of science*. Netherlands: Springer, 59-65.

Giora, R., Gazal, O., & Goldstein, I. (2012): Salience and context: Interpretation of metaphorical and literal language by young adults diagnosed with Asperger's syndrome. *Metaphor and Symbol* 27, 22-54.

Glucksberg, S. (2001): *Understanding Figurative Language: From Metaphors to Idioms*. Oxford: Oxford University Press.

Glucksberg, S. (2003): The psycholinguistics of metaphor. *Trends in Cognitive Science* 7(2), 92-96.

Glucksberg, S., McGlone, M.S. (1999): When love is not a journey: What metaphors mean. *Journal of Pragmatics* 31, 1541-1558.

Glucksberg, S., Keysar, B., McGlone, M.S. (1992): Metaphor Understanding and Accessing Conceptual Schema: Reply to Gibbs (1992). *Psychological Review* 92(3), 578-581.

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

Gokcesu, B.S. (2009): Comparison, categorization, and metaphor comprehension. In: Taatgen, N. & H. van Rijn (eds.): *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*, 567-572.

Gooding, D.C. (2000): Experiment. In: Newton-Smith, W.H. (ed.): *A Companion to the Philosophy of Science*. Malden & Oxford: Blackwell, 117-126.

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

Hacking, I. (1983): *Representing and intervening*. Cambridge University Press, Cambridge.

Hacking, I. (1992): The Self-Vindication of the Laboratory Sciences. In: Pickering, A. (ed.): *Science as Practice and Culture*. Chicago: University of Chicago Press, 29-64.

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*. Amsterdam & Philadelphia: Benjamins, 304-324.

Hjelmslev, L. (1969): *Prolegomena to a theory of language*. Madison: The University of Wisconsin Press.

Jones, L.L. (2004): *Metaphor comprehension: An exemplar of ad hoc category creation*. MSc dissertation. Athens, Georgia.

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

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

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

Katz, A.N., Paivio, A., Marschark, M., & Clark, J.M. (1988): Norms for 204 literary and 260 nonliterary metaphors on 10 psychological dimensions. *Metaphor and Symbolic Activity* 3(4), 191-214.

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.

Kepser, S., Reis, M. (2005): Evidence in linguistics. In: Kepser, S., Reis, M. (eds.): *Linguistic evidence. Empirical, theoretical and computational perspectives*. Berlin & New York: de Gruyter, 1-6.

Kertész, A. (2004): *Philosophie der Linguistik*. Tübingen: Narr.

Kertész, A., Rákosi, Cs. (2008a): Daten und Evidenz in linguistischen Theorien: Ein Forschungsüberblick. In: Kertész, A., Rákosi, Cs. (eds.): *New approaches to linguistic evidence. Pilot Studies / Neue Ansätze zu linguistischer Evidenz. Pilotstudien*. Frankfurt am Main & Bern & Bruxelles & New York & Oxford & Wien: Lang, 21-60.

Kertész, A., Rákosi, Cs. (2008b): Conservatism vs. innovation in the (un)grammaticality debate. In: Kertész, A., Rákosi, Cs. (eds.): *New approaches to linguistic evidence. Pilot studies*

/ Neue Ansätze zu linguistischer Evidenz. Pilotstudien. Frankfurt am Main & Bern & Bruxelles & New York & Oxford & Wien: Lang, 61-84.

Kertész, A., Rákosi, Cs. (2008c): Conservatism vs. innovation in the debate on data in generative grammar. In: Kertész, A., Rákosi, Cs. (eds.): *New approaches to linguistic evidence. Pilot studies / Neue Ansätze zu linguistischer Evidenz. Pilotstudien*. Frankfurt am Main & Bern & Bruxelles & New York & Oxford & Wien: Lang, 85-118.

Kertész, A., Rákosi, Cs. (2009): Cyclic vs. circular argumentation in the Conceptual Metaphor Theory. *Cognitive Linguistics* 20(4), 703-732.

Kertész, A. & Rákosi, Cs. (2012): *Data and evidence in linguistics: A plausible argumentation model*. Cambridge: Cambridge University Press.

Kertész, A. & Rákosi, Cs. (2013): Paraconsistency and plausible argumentation in Generative Grammar: A case study. *Journal of Logic, Language and Information* 22 (2), 195-230.

Kertész, A. & Rákosi, Cs. (2014): The p-model of data and evidence in linguistics. In: Kertész, A. & Rákosi, Cs. (eds.): *The evidential basis of linguistic argumentation.* Amsterdam & Philadelphia: John Benjamins, 15-48.

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

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

Kuhn, Th. (1962, 1970): *The structure of scientific revolutions*. Chicago: The University of Chicago.

Kusumi, T. (1987): Effects of categorical dissimilarity and affective similarity between constituent words on metaphor appreciation. *Journal of Psycholinguistic Research* 16(6), 577-595.

Lai, V.T., Curran, T., & Menn, L. (2009): Comprehending conventional and novel metaphors: An ERP study. *Brain Research* 1284, 145-155.

Lakatos, I. (1978): Falsification and the methodology of scientific research programmes. In: Worrall, J. & Currie, G. (eds.): The methodology of scientific research programmes, Vol. 1. Cambridge & New York: Cambridge University Press, 8-101.

Lakoff, G., 1993. The contemporary theory of metaphor. In: Ortony, A. (ed.): *Metaphor and thought*. Cambridge: Cambridge University Press, 202-252.

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

Laudan, L. (1977): *Progress and Its Problems: Towards a Theory of Scientific Growth*. University of California Press.

Lehmann, C. (2004): Data in linguistics. *The Linguistic Review* 21, 175-210.

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

Marschark, M., Katz, A.N., & Paivio, A. (1983): Dimensions of metaphor. *Journal of Psycholinguistic Research* 12(1), 17-40.

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

McKay, M.T. (2004): The development and comprehension of conventional metaphors. PhD dissertation. University of Florida.

McQuire, M., McCollum, L., & Chatterjee, A. (2017): Aptness and beauty in metaphor. *Language and Cognition* 9(2), 316-331.

Meheus, J. (ed.)(2002): *Inconsistency in science*. Dordrecht: Kluwer.

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

Nagy C., K. (2014): Methods and argumentation in historical linguistics. In: Kertész, A. & Rákosi, Cs. (eds.): *The evidential basis of linguistic argumentation*. Amsterdam & Philadelphia: Benjamins, 71-102.

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

Nickles, Th. (1989): Justification and experiments. In: Gooding, D., Pinch, T., Schaffer, S. (eds.): *The uses of experiment. Studies in the natural sciences*. Cambridge: Cambridge University Press, 299-333.

Nickles, T. (2000): Lakatos. In: Newton-Smith, W.H. (ed.): *A companion to the philosophy of science*. Blackwell, 207-212.

Nickles, T. (2002): From Copernicus to Ptolemy: Inconsistency and method. In: Meheus, J. (ed.): *Inconsistency in science*. Dordrecht/Boston/London: Kluwer, 1-33.

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.

Penke, M., Rosenbach, A. (2004): What counts as evidence in linguistics? *Studies in Language* 28(3), 480-526.

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

Pickering, A. (1989): Living in the material world: On realism and experimental practice. In: Gooding, D., Pinch, T., Schaffer, S. (eds.): *The uses of experiment. Studies in the natural sciences*. Cambridge: Cambridge University Press, 275-297.

Pierce, R.S. & Chiappe, D.L. (2009): The roles of aptness, conventionality, and working memory in the production of metaphors and similes. *Metaphor and Symbol* 24, 1-19.

Pitt, J.C. (2001): The dilemma of case studies. Toward a heraclitian philosophy of science. *Perspectives on Science* 9(4): 373-382.

Popper, K. (1959): *The logic of scientific discovery*. London: Hutchinson.

Pullum, G.K. (2007): Ungrammaticality, rarity, and corpus use. *Corpus Linguistics and Linguistic Theory* 3, 33-47.

Rákosi, Cs. (2011a): Metatheoretical reconstruction of psycholinguistic experiments. Part 1. *Sprachtheorie und germanistische Linguistik* 21 (1), 55-93.

Rákosi, Cs. (2011b): Metatheoretical reconstruction of psycholinguistic experiments. Part 2. *Sprachtheorie und germanistische Linguistik* 21 (2), 159-187.

Rákosi, Cs. (2012): The Fabulous Engine: Strengths and flaws of psycholinguistic experiments. *Language Sciences* 34(6), 682-702.

Rákosi, Cs. (2014): On the rhetoricity of psycholinguistic experiments. *Argumentum* 10, 533-547.
Rákosi, Cs. (2016a): On the evaluation of psycholinguistic experiments on metaphor: Part I: The metatheoretical background. *Argumentum* 12, 278-287.
Rákosi, Cs. (2016b): On the evaluation of psycholinguistic experiments on metaphor: Part II: Case studies. *Argumentum* 12, 288-302.
Rákosi, Cs. (2017a): Replication of Psycholinguistic Experiments and the Resolution of Inconsistencies. *Journal of Psycholinguistic Research* 46(5), 1249-1271.
Rákosi, Cs. (2017b): 'Experimental complexes' in psycholinguistic research on metaphor processing. *Sprachtheorie und germanistische Linguistik* 27(1), 3-32.
Rákosi, Cs. (2018a): Dealing with the Conflicting Results of Psycholinguistic Experiments: How to Resolve Them with the Help of Statistical Meta-analysis. *Journal of Psycholinguistic Research* 47(4), 777-801.
Rákosi, Cs. (2018b): Remarks on the margins of a debate on the role of metaphors on thinking. *Sprachtheorie und germanistische Linguistik* 28(1), 3-35.
Reijnierse, W. G., Burgers, C., Krennmayr, T., & Steen, G. J. (2015). How viruses and beasts affect our opinions (or not): The role of extendedness in metaphorical framing. *Metaphor and the Social World*, 5, 245-263. doi:10.1075/msw.
Rescher, N. (1976): *Plausible reasoning*. Van Gorcum, Assen/Amsterdam.
Rescher, N. (1977): *Methodological Pragmatism*. Blackwell, Oxford.
Rescher, N. (1987): How serious a fallacy is inconsistency? *Argumentation* 1, 303-316.
Roncero, C. (2013): *Understanding figurative language: Studies on the comprehension of metaphors and similes*. PhD Thesis, Concordia University, Montreal, Canada.
Roncero, C., Almeida, R.G., Martin, D.C., & de Caro, M. (2016): Aptness predicts metaphor preference in the lab and on the Internet. *Metaphor and Symbol* 31(1), 31-46.
Sampson, G. (1975): *The form of language*. London: Weidenfeld & Nicholson.
Sampson, G.R. (2007a): Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory* 3, 1-32.
Sampson, G.R. (2007b): Reply. *Corpus Linguistics and Linguistic Theory* 3, 111-129.
Sanford, D. (2010): *Figuration & frequency: A usage-based approach to metaphor.* PhD dissertation. Albuquerque, New Mexico: The University of New Mexico.
Schickore, J. (2011): The significance of re-doing experiments: A contribution to historically informed methodology. *Erkenntnis* 75, 325-347.
Schiffrin, D. (1981): Tense variation in narrative. *Language* 57(1), 45-62.
Schlesewsky, M. (2009): Linguistische Daten aus experimentellen Umgebungen: Eine multiexperimentelle und multimodale Perspektive. *Zeitschrift für Sprachwissenschaft* 28, 169-178.
Scholl, R. & Räz, T. (2016): Towards a methodology for integrated history and philosophy of science. In: Sauer, T. & Scholl, R. (eds.): The philosophy of historical case studies. Boston Studies in the Philosophy and History of Science 319. Switzerland: Springer.
Schütze, C.T. (1996): *The empirical base of linguistics. Grammaticality judgments and linguistic methodology*. Chicago & London: The University of Chicago Press.
Simone, R. (2004): The object, the method, and the ghosts. Remarks on a terra incognita. *The Linguistic Review* 21, 235-256.

Smith, N. (2000): Foreword to Chomsky, N., 2000. *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

Steen, G. J., Reijnierse, W. G., & Burgers, C. (2014). When do natural language metaphors influence reasoning? A follow-up study to Thibodeau and Boroditsky (2013). *PLoS ONE*, 9(12), e113536. DOI: 10.1371/journal.pone.0113536.

Stefanowitsch, A., Gries, S.Th. (eds.)(2007): *Grammar without grammaticality*. Special issue of Corpus Linguistics and Linguistic Theory.

Sternberg, R. & Nigro, G. (1980): Interaction and analogy in the comprehension and appreciation of metaphors. *The Quarterly Journal of Experimental Psychology A (Human Experimental Psychology)* 35(1), 17-38.

Sternefeld, W. (ed.)(2007): Data in generative linguistics. *Theoretical Linguistics* 33(3), 269-410.3.

Thibodeau, P.H. (2016): Extended Metaphors are the Home Runs of Persuasion: Don't Fumble the Phrase. *Metaphor and Symbol* 31(2), 53-72.

Thibodeau, P.H., & Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. *PLoS ONE*, 6(2), e16782. DOI: 10.1371/journal.pone.0016782.

Thibodeau, P.H., & Boroditsky, L. (2013). Natural language metaphors covertly influence reasoning. *PLoS ONE*, 8(1), e52961. DOI: 10.1371/journal.pone.0052961.

Thibodeau, P.H., & Boroditsky, L. (2015). Measuring effects of metaphor in a dynamic opinion landscape. *PLoS ONE*, 10(7), e0133939. doi:10.1371/journal.pone.0133939.

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

Thibodeau, P. & Durgin, F.H. (2011): Metaphor aptness and conventionality: A processing fluency account. *Metaphor and Symbol* 26(3), 206-226.

Thibodeau, P., Sikos, L., & Durgin, F.H. (2016): What do we learn from rating metaphors? In: Papafragou, A., Grodner, D.J., Mirman, D., & J. Trueswell (eds.): *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 1769-1774.

Thibodeau, P., Sikos, L., & Durgin, F.H. (2018): Are subjective ratings of metaphors a red herring? The big two dimensions of metaphoric sentences. *Behavior Research Methods* 50(2), 759-772.

Tourangeau, R. & Sternberg, R.J. (1981): Aptness in metaphor. *Cognitive Psychology* 13, 27-55.

Utsumi, A. (2007): Interpretive diversity explains metaphor–simile distinction. *Metaphor and Symbol* 22(4), 291-312.

Utsumi, A. & Kuwabara, Y. (2005): Interpretive diversity as a source of metaphor-simile distinction. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, 2230-2235.

Utsumi, A. & Sakamoto, M. (2010): Predicative metaphor comprehension as indirect categorization. In: Ohlsson, S. & R. Catrambone (eds.): *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*. Cognitive Science Society, Austin, Texas, USA, 1034-1039.

Utsumi, A. & Sakamoto, M. (2011): Indirect categorization as a process of predicative metaphor comprehension. *Metaphor and Symbol* 26, 299-313.

Veale, T. (2006): Computability as a test on linguistic theories. In: Kristensen, G., Achard, M., Dirven, R. & Mendoza Ibañez, F.J.R. (eds.): *Cognitive linguistics: Current applications and future perspectives.* Berlin, New York: De Gruyter, 461-483.

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

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.

APPENDIX 1

	N	conventional figuratives	novel figuratives
Bowdle & Gentner (1999) = Bowdle & Gentner (2005), Experiment 3	48	6.13 (1.27)	3.52 (1.335)
Chiappe, Kennedy & Smykowsky (2003)	44	correlation: +0.01, $p > 0.9$	
Jones (2004), Experiment 1	51	0.69 (0.21)	0.54 (0.21)
Jones (2004), Experiment 2	48	0.61 (0.24)	0.4 (0.24)
Bowdle & Gentner (2005), Experiment 1	16	4.35 (0.87)	2.81 (0.83)
Bowdle & Gentner (2005), Experiment 2	32	0.33 (0.29)	0.24 (0.26)
Jones & Estes (2005), Experiment 1	51	0.69 (0.32)	0.54 (0.29)
Jones & Estes (2005), Experiment 2	60	3.11 (0.93)	2.47 (0.77)
Jones & Estes (2006), Experiment 1	48	3.33 (0.76)	3.51 (0.83)
Jones & Estes (2006), Experiment 3	31	2.98 (1.06)	3.03 (1.11)
Utsumi (2007), Experiment 1	30	3.14 (0.71)	2.55 (0.59)
Pierce & Chiappe (2009)	275	$t = 1.23, p = 0.22$	
Roncero (2013) = Roncero et al. (2016)	104	$r_s = 0.04, p = 0.72$	
Dulcinati et al. (2014)	82	Pearson's correlation coefficient: $r = 0.26, p < 0.05$.	

Table 8. Experimental data for grammatical form preference with conventionality as a decisive factor

	N	low apt figuratives	high apt figuratives
Chiappe & Kennedy (1999), Experiment 2	46	correlation: 0.75, $p < 0.005$	
Chiappe, Kennedy & Chiappe (2003), metaphors	34	1-0.31 = 0.69 (0.25)	1-0.2 = 0.8 (0.16)
Chiappe, Kennedy & Chiappe (2003), similes	34	0.4 (0.28)	0.58 (0.26)
Chiappe, Kennedy & Smykowsky (2003)	44	correlation: +0.63, $p < 0.001$	
Bowdle & Gentner (2005), Experiments 1-2	48	$r = -0.65$, $p < 0.01$	
Jones & Estes (2005), Experiments 1-2	111	Aptness was strongly correlated with category membership, $r = 0.75$, $p < 0.001$.	
Jones & Estes (2005), Experiment 3	60	2.29 (0.77)	4.23 (1.16)
Jones & Estes (2006), Experiment 1	48	3.27 (0.9)	3.57 (0.83)
Jones & Estes (2006), Experiment 3	31	2.38 (1.06)	3.63 (1.11)
Utsumi (2007), Experiment 1	30	2.47 (0.64)	3.34 (0.71)
Roncero (2013) = Roncero et al. (2016)	104	$r_s = 0.61$, $p < 0.001$	
Dulcinati et al. (2014)	82	Pearson's correlation coefficient: $r = 0.58$, $p < 0.01$.	

Table 10. Experimental data for grammatical form preference with aptness as a decisive factor

	N	low familiar figuratives	high familiar figuratives
Chiappe & Kennedy (2001), Experiment 3	16	correlation: +0.57, $p < 0.001$	
Utsumi & Kuwabara (2005), Experiments 1-2	30	$r = 0.47$, $p < 0.01$	
Utsumi (2007), Experiment 1	30	2.45 (0.61)	3.15 (0.77)
Utsumi (2007), Experiment 1	30	correlation: 0.46 ($p < 0.01$)	
Roncero (2013) = Roncero et al. (2016)	104	$r = 0.47$, $p < 0.001$	
Dulcinati et al. (2014)	82	Pearson's correlation coefficient: $r = 0.1$	

Table 12. Experimental data for grammatical form preference with familiarity as a decisive factor

	N	high-familiar metaphors	low-familiar metaphors
Blasko & Connine (1993), Experiment 1	81	887 (173)	983 (208)
Blasko & Connine (1993), Experiment 2	36	795 (129)	926 (210)
Arzouan et al. (2007), Experiment 1	31	875 (264)	1115 (318)
Arzouan et al. (2007), Experiment 3	15	987 (194)	1560 (449)
Lai et al. (2009)	29	$F(1, 69) = 23.437$, $p < 0.0005$	
Sanford (2010)	26	5371.2 (850.33)	5865.2 (814.66)
Thibodeau & Durgin (2011)	72	$r = -0.249$, $t(126) = 2.89$, $p < 0.01$.	
Caillies & Declercq (2011), Experiment 1	20	845 (131)	952 (160)
Caillies & Declercq (2011), Experiment 1	20	748 (132)	802 (170)
Caillies & Declercq (2011), Experiment 1	20	765 (89)	820 (143)
Caillies & Declercq (2011), Experiment 1	18	702 (133)	728 (128)
Giora, Gazal & Goldstein (2012), Experiment 1	28	2374 (823)	3198 (1618)
Giora, Gazal & Goldstein (2012), Experiment 2	28	2109 (860)	2550 (1181)
Cardillo et al. (2017)	20	correlation: -0.15	

Table 14. Experimental data for comprehension latencies with familiarity as a decisive factor

	N	low apt	high apt
Blasko & Connine (1993), Experiment 3	39	889 (145)	816 (117)
Brisand et al. (2001), Experiment 1	60	650 (305)	631 (327)
Brisand et al. (2001), Experiment 2	60	498 (80)	525 (129)
Gagné (2002), Experiment 1, both forms	30	$r = -0.46$, $p < 0.009$	
Gagné (2002), Experiment 1, metaphors	30	$r = -0.5$, $p < 0.009$	
Chiappe, Kennedy & Chiappe (2003)	34	correlation: -0.55, $p < 0.01$	
Jones & Estes (2006), Experiment 2	60	4121 (782)	3302 (682)
Utsumi & Sakamoto (2010)	38	837.4 (218.5)	810.3 (223.2)
Utsumi & Sakamoto (2011), Experiment 2	38	814.8 (234.5)	831.7 (206.7)

Table 15. Experimental data for comprehension latencies with aptness as a decisive factor

	N	conventional figuratives	novel figuratives
Bowdle & Gentner (2005), Experiment 2, figuratives	32	2160 (834)	3058 (1327)
Bowdle & Gentner (2005), Experiment 2, metaphors	32	2063 (873.5)	3245 (1672.5)
Jones & Estes (2006), Experiment 2	60	3697 (1464)	3590 (1309)
Utsumi & Sakamoto (2010)	38	831.7 (206.7)	816 (235)
Utsumi & Sakamoto (2011), Experiment 2	38	809.1 (222.7)	837.4 (218.5)

Table 17. Experimental data for comprehension latencies with conventionality as a decisive factor

	N	high-familiar metaphors	low-familiar metaphors
Marschark, Katz, Paivio (1983), Experiment 1	334	Correlation between comprehensibility and familiarity: 0.82, $p < 0.01$	
Marschark, Katz, Paivio (1983), Experiment 2	303	Correlation between comprehensibility and familiarity: 0.91, $p < 0.01$	
Katz et al. (1988)	30	Correlation between comprehensibility and familiarity, non-literal metaphors: $r = 0.82$	
McKay (2004)	200	Pearson correlation between comprehensibility and familiarity: 0.93, $p = 0.01$	
Lai et al. (2009)	29	Familiar vs. novel: $F(1,69)=185.692$, $p<.0005$	
Cardillo et al. (2010), predicative metaphors	20	Correlation coefficient between familiarity and interpretability: 0.30, $p < 0.01$	
Cardillo et al. (2010), nominal metaphors	20	Correlation coefficient between familiarity and interpretability: 0.27, $p < 0.01$.	
Sanford (2010)	18	3.072 (0.64)	2.85 (0.56)
Bambini, Resta, Grimaldi (2014), without context	105	An inverse robust correlation between difficulty and familiarity: $r_s (113) = -0.60$, $p < 0.01$	
Bambini, Resta, Grimaldi (2014), with context	180	Difficulty correlated inversely with familiarity ($r_s (63) = 0.40$, $p_s < 0.01$)	
Campbell & Raney (2016)	90	Correlation between comprehensibility and familiarity: 0.97, $p < 0.001$	
Cardillo et al. (2017)	20	Correlation coefficient between familiarity and ease of interpretation: 0.79, $p < 0.01$	

Table 19. Experimental data for comprehensibility ratings with familiarity as a decisive factor

	N	low apt	high apt
Sternberg & Nigro (1980)	24	The correlation between ratings of comprehensibility and of aptness was 0.61, $p < .0001$, across the five forms.	
Tourangeau & Sternberg (1981), Experiment 1	20	Correlation between aptness and comprehensibility: 0.64 ($p < 0.01$)	
Marschark, Katz, Paivio (1983), Experiment 1	334	Correlation between comprehensibility and aptness: 0.82, $p < 0.01$	
Marschark, Katz, Paivio (1983), Experiment 2	303	Correlation between comprehensibility and aptness: 0.87, $p < 0.01$	
Kusumi (1987)	96	Correlation between comprehensibility and aptness: $r = 0.83$, $p < 0.01$	
Katz et al. (1988)	30	Correlation between comprehensibility and aptness, non-literal metaphors: $r = 0.82$	
Gagné (2002), Experiment 1, both forms	30	The higher the aptness of a comparison, the higher the comprehensibility rating was for that combination: $r = 0.81$, $p < .0001$	
Gagné (2002), Experiment 1, only metaphors	30	The higher the aptness of a comparison, the higher the comprehensibility rating was for that combination: $r = 0.77$, $p < .0001$	
Chiappe, Kennedy & Chiappe (2003)	34	We found a correlation of 0.94 between the comprehensibility judgments and the aptness judgments, $p < 0.001$	
McKay (2004)	200	Pearson correlation between comprehensibility and aptness: 0.59, $p = 0.01$	
Jones & Estes (2006), Experiment 2	60	4.91 (0.46)	5.71 (0.39)
Utsumi (2007), Experiment 2, metaphors	42	2.86 (1.05)	5.54 (0.84)
Utsumi (2007), Experiment 2, similes	42	3.31 (1.26)	5.65 (0.82)
Thibodeau et al. (2016=2018)	1193	Correlation between comprehensibility and aptness: 0.883, $p < 0.001$	
McQuire et al. (2017), Experiment 1, young adults	20	Aptness correlated positively with […] interpretability, Pearson $r = 0.427$, $p < 0.0005$	
McQuire et al. (2017), Experiment 1, literary experts	20	Aptness correlated positively with […] interpretability, Pearson $r = 0.407$, $p < 0.0005$	
McQuire et al. (2017), Experiment 1, elderly adults	20	Aptness correlated positively with […] interpretability, Pearson $r = 0.44$, $p < 0.0001$	
Campbell & Raney (2016)	90	Correlation between comprehensibility and metaphor goodness (aptness): 0.97, $p < 0.001$	

Table 21. Experimental data for comprehensibility ratings with aptness as a decisive factor

	N	conventional figuratives	novel figuratives
McKay (2004)	200	Pearson correlation between comprehensibility and conventionality: 0.42, p = 0.01	
Utsumi (2007), Experiment 2, metaphors	42	4.67 (0.92)	3.97 (1.01)
Utsumi (2007), Experiment 2, similes	42	4.8 (0.83)	4.15 (0.98)
Gokcesu (2009), Experiment 3		Conventional metaphors (M = 0.863) were rated as more sensible than novel metaphors (M = 0.697), $t(53) = 3.077$, $p < 0.05$	

Table 23. Experimental data for comprehensibility ratings with conventionality as a decisive factor

factor	researcher	calculation	average points
conventionality	Bowdle & Gentner	1 x 3 + 2 x 5 + 2 x 4	4.2
	Chiappe	2 x 1	1
	Dulcinati	1 x 3	3
	Gokcesu	1 x 3	3
	Jones & Estes	2 x 1 + 4 x 3 + 1 x 2	2.3
	McKay	1 x 3	3
	Roncero	1 x 1	1
	Utsumi	3 x 3 + 2 x 2	2.6
aptness	Blasko & Connine	1 x 3	3
	Bowdle & Gentner	1 x 1	1
	Brisand	2 x 1	1
	Campbell	1 x 5	5
	Chiappe	2 x 1 + 1 x 3 + 3 x 5	3.3
	Dulcinati	1 x 3	3
	Gagné	2 x 3 + 2 x 5	4
	Jones & Estes	2 x 1 + 1 x 3 + 3 x 5	3.3
	Kusumi	1 x 3	3
	Marschark[112]	2 x 3 + 1 x 5	3.67
	McKay	1 x 1	1
	McQuire	3 x 1	1
	Roncero	1 x 3	3
	Sternberg	2 x 1	1
	Thibodeau	1 x 5	5
	Utsumi	2 x 1 + 3 x 3	2.2
familiarity	Arzouan	2 x 3	3
	Bambini	2 x 1	1
	Blasko & Connine	2 x 3	3
	Caillies & Declercq	4 x 3	3
	Campbell	1 x 5	5
	Cardillo	2 x 1 + 2 x 3	2
	Chiappe	1 x 4	4
	Dulcinati	1 x 2	2
	Giora	2 x 3	3
	Lai	1 x 5 + 1 x 3	4
	Marschark	2 x 3 + 1 x 5	3.67
	McKay	1 x 5	5
	Roncero	1 x 4	4
	Sanford	1 x 1 + 1 x 3	2
	Thibodeau & Durgin	1 x 3	3
	Utsumi	2 x 4	4

Table 25. Researchers and effect size groups – overview

[112] Katz et al. (1988) is referred to as Marschark (1988) here.

APPENDIX 2

metaphorical frame	beast		virus	
response type	social	enforcement	social	enforcement
Thibodeau & Boroditsky (2011), Experiment 1	61	170	98	127
Thibodeau & Boroditsky (2011), Experiment 2	33	80	61	72
Thibodeau & Boroditsky (2011), Experiment 4	50	33	74	21
Thibodeau & Boroditsky (2013), Experiment 2	136	44	174	26
Thibodeau & Boroditsky (2013), Experiment 3	14	62	26	57
Thibodeau & Boroditsky (2013), Experiment 4	97	73	111	53
Steen et al. (2014), Experiment 1	17	63	14	72
Steen et al. (2014), Experiment 2	45	46	36	53
Steen et al. (2014), Experiment 3	35	52	36	49
Steen et al. (2014), Experiment 4	169	189	152	187
Thibodeau & Boroditsky (2015), Experiment 1	148	97	184	97
Thibodeau & Boroditsky (2015), Experiment 2	112	175	109	125
Reijnierse et al. (2015), 1-metaphor condition	35	34	32	40
Reijnierse et al. (2015), 2-metaphor condition	28	36	38	32
Reijnierse et al. (2015), 3-metaphor condition	33	37	40	26
Reijnierse et al. (2015), 4-metaphor condition	35	34	37	25
Christmann & Göhring (2016)	25	21	35	13

Table 44. Response frequencies in the first (top choices) analysis

metaphorical frame	beast			virus		
response type	N	social	enforcement	N	social	enforcement
Thibodeau & Boroditsky (2013), Experiment 2	180	5.61 (1.404)	4.39 (1.404)	200	5.99 (1.156)	4.01 (1.156)
Thibodeau & Boroditsky (2013), Experiment 3	76	5.2 (1.789)	9.8 (1.789)	83	5.77 (1.769)	9.23 (1.789)
Thibodeau & Boroditsky (2013), Experiment 4	170	6.89 (1.504)	8.11 (1.504)	164	6.97 (1.381)	8.03 (1.381)
Steen et al. (2014), Experiment 1	80	5.3 (2.009)	9.7 (2.009)	86	5.24 (2.011)	9.76 (2.011)
Steen et al. (2014), Experiment 2	91	6.46 (2.024)	8.54 (2.024)	89	6.25 (2.123)	8.75 (2.123)
Steen et al. (2014), Experiment 3	87	6.43 (2.061)	8.57 (2.061)	85	6.18 (2.199)	8.72 (2.199)
Steen et al. (2014), Experiment 4	358	6.54 (2.037)	8.46 (2.037)	339	6.42 (2.018)	8.58 (2.018)
Thibodeau & Boroditsky (2015), Experiment 1	245	9.64 (2.047)	5.36 (2.047)	281	9.67 (1.969)	5.33 (1.969)

Table 45. Mean rankings in the second (rankings/ratings) analysis

metaphorical frame	beast			virus		
response type	**N**	**social**	**enforcement**	**N**	**social**	**enforcement**
Reijnierse et al. (2015), 1-metaphor condition	72	5.3924 (1.12887)	4.9479 (1.36282)	75	5.18 (0.85578)	4.94 (1.48235)
Reijnierse et al. (2015), 2-metaphor condition	67	5.4179 (1.07166)	4.8321 (1.24954)	75	5.32 (1.01472)	4.9867 (1.41415)
Reijnierse et al. (2015), 3-metaphor condition	76	5.5 (0.99415)	5.1349 (1.16078)	71	5.2782 (1.19713)	4.4401 (1.59489)
Reijnierse et al. (2015), 4-metaphor condition	70	5.4964 (1.06449)	5.1321 (1.22602)	68	5.5184 (0.92973)	4.8051 (1.50549)
Christmann & Göhring (2016)	61	1.23 (1.055)	1.02 (0.904)	61	1.18 (0.806)	0.87 (1.056)

Table 46. Ratings/choices of the measures in the second (ratings/rankings) analysis

experiment	meta-phorical frame	N	economy	education	patrols	prison	neighbour-hood watches
TB2013/2	**beast**	180	3.02 (1.167)	2.58 (1.051)	2.38 (1.01)	2.01 (1.003)	–
	virus	200	3.23 (1.05)	2.76 (1.009)	2.18 (0.927)	1.84 (0.945)	–
TB2013/3	**beast**	76	2.59 (0.912)	2.61 (1.461)	3.61 (1.575)	2.96 (1.248)	3.24 (1.539)
	virus	83	2.66 (0.928)	3.11 (1.506)	2.99 (1.721)	3.04 (1.401)	3.2 (1.377)
TB2013/4	**beast**	170	3.76 (1.394)	3.13 (1.219)	2.81 (1.328)	2.72 (1.468)	2.57 (1.345)
	virus	164	4.04 (1.248)	2.93 (1.325)	2.65 (1.212)	2.76 (1.371)	2.63 (1.411)
Steen2014/1	**beast**	80	2.6 (1.472)	2.7 (1.084)	3.36 (1.407)	2.79 (1.49)	3.55 (1.359)
	virus	86	2.5 (1.281)	2.74 (1.238)	3.45 (1.428)	2.53 (1.436)	3.77 (1.224)
Steen2014/2	**beast**	91	3.21 (1.465)	3.25 (1.287)	3.12 (1.452)	1.86 (1.216)	3.56 (0.98)
	virus	89	3.34 (1.453)	2.91 (1.258)	3.19 (1.372)	2.12 (1.388)	3.44 (1.388)
Steen2014/3	**beast**	87	3.33 (1.436)	3.09 (1.197)	3 (1.276)	1.69 (1.113)	3.89 (1.05)
	virus	85	2.94 (1.499)	3.24 (1.342)	3.22 (1.383)	2.01 (1.239)	3.59 (1.083)
Steen2014/4	**beast**	358	3.27 (1.432)	3.27 (1.249)	3.05 (1.305)	1.76 (1.164)	3.65 (1.117)
	virus	339	3.04 (1.447)	3.38 (1.24)	3.11 (1.274)	1.72 (1.091)	3.75 (1.103)
TB2015/1	**beast**	245	2.73 (1.478)	3.48 (1.179)	3.44 (1.356)	1.92 (1.268)	3.43 (1.075)
	virus	281	2.67 (1.437)	3.42 (1.15)	3.42 (1.228)	1.91 (1.344)	3.58 (1.156)
Reij2015/1	**beast**	72	5.271 (1.25)	5.514 (1.236)	5.903 (1.128)	4.63 (1.608)	–
	virus	75	4.953 (1.004)	5.407 (1.012)	5.667 (1.446)	4.698 (1.645)	–
Reij2015/2	**beast**	67	5.209 (1.178)	5.627 (1.191)	5.791 (1.238)	4.512 (1.501)	–
	virus	75	5.147 (1.224)	5.493 (1.095)	5.84 (0.987)	4.702 (1.697)	–
Reij2015/3	**beast**	76	5.263 (1.121)	5.737 (1.005)	5.974 (0.993)	4.855 (1.368)	–
	virus	71	5.106 (1.284)	5.451 (1.355)	5.338 (1.53)	4.141 (1.774)	–
Reij2015/4	**beast**	70	5.364 (1.161)	5.629 (1.099)	5.929 (1.159)	4.867 (1.458)	–
	virus	68	5.419 (1.053)	5.618 (1.163)	5.765 (1.34)	4.485 (1.708)	–

Table 48. Mean and standard deviation of the rankings/ratings of the measures in the third (measures) analysis