Author's version. Published in: Language Sciences 34(6), 682-701, 2012

The Fabulous Engine:
Strengths and flaws of psycholinguistic experiments

Csilla Rákosi
University of Debrecen, Research Group for Theoretical Linguistics, Pf. 47, H-4010 Debrecen, Hungary
E-mail address: rakosics@gmail.com

> *"People once believed a fabulous engine called the Scientific Method harvests empirical evidence through observation and experimentation, discards subjective, error ridden chaff, and delivers objective, veridical residues from which to spin threads of knowledge. Unfortunately, that engine is literally fabulous."* (Bogen, 2002, p. 128)

Abstract
In methodological debates in linguistics, the scientificness of linguistics is often felt to be unsatisfactory in comparison to the standards of natural sciences. This inferiority complex is mostly articulated as the requirement to turn linguistics into a mature empirical theory with the help of, for example, the elaboration and conduct of psycholinguistic experiments. This paper argues that the proposals which have been put forward to fulfil this requirement are based not on workable and generally applied norms of natural sciences but on outmoded and untenable tenets of the standard view of the analytical philosophy of science. Therefore, a two-step strategy is suggested: 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 natural sciences on the other. The feasibility of this strategy is supported by a case study on psycholinguistic experiments in metaphor research.
Keywords: psycholinguistic experiments, linguistic data, linguistic evidence, methodology of linguistics

1. Introduction

While many researchers are uninterested in methodological issues and seem to be of the opinion that linguistics can be practised without making explicit the methodological background of linguistic theorising, there is a deep feeling of unease about methodological 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, p. 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, p. 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 borrow methodological tools and norms from other disciplines.

The presence of the two strivings can be traced back to the same cause: the scientificness 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 turn linguistics into a mature empirical theory.[3] The following *general requirements* have been imposed and found wide acceptance:

(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) 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'.

(d) Data are immediately given and primary to the theory.

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, among others, in Geeraerts (2006), Lehmann (2004) and Simone (2004). 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 experiments, use of corpora, surveys, fieldwork. Evidence consists of observation statements capturing different perceptible manifestations of linguistic behaviour.[5]

---

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 far from applying the methodology of 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, p. 127)

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, p. 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, p. 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, p. 23; emphasis as in the original)

(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]

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

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.[9] 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.[10]

Second, consequently, in opposition to (SR)(b), neither corpus data, data gained by experiments or 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 types have to be assumed to be problematic, and they are inevitably highly theory- and problem-dependent. Moreover, although linguistic data cannot be treated as irrevisable facts, the everyday practice of linguistic research and the metascientific reflection of a considerably 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.[11]

---

6 "[Something] may nevertheless function as a datum in some research that assigns it the role of unquestionable evidence in the argumentation." (Lehmann, 2004, p. 181)

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 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, p.185-186; emphasis added)
"[…] linguistics should primarily develop an independent *observational language* that the different *theoretical languages* of linguistics can be mapped onto […]." (Geeraerts, 2006, p. 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, p. 24; emphasis as in the original)

9 "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, p. 14)
"If linguistics is indeed based on intuition, then it is not a science […] Science relies exclusively on the empirical." (Sampson, 1975, p. 60)

10 Cf. Kertész and Rákosi (2008a,b,c, 2012); Schütze (1996); Lehmann (2004); Penke and Rosenbach (2004); Kepser and Reis (2005); Borsley (2005); Stefanowitsch and Gries (eds.)(2007); Sternefeld (ed.)(2007); Consten and Loll (this issue).

11 For an overview, see Kertész and Rákosi (2008a, b, c, 2012).

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 seems to be principled reasons for the failure of this requirement.

These problems cast doubt on (GR) as well. The uncertainty, problem- and theory-dependence of linguistic data is irreconcilable with (GR)(a), (c) and (d). 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.

(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.[12] Although these two strategies have the same goal, and share the same metascientific commitments, they sharply criticise and reject each others' 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).[13]

(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. The latter, however, have never been

---

[12] 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, p. 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], p. 13-16; Andor, 2004, p. 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 to formulate *general laws* of linguistic competence (Chomsky, 1969 [1957], p. 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], p. 49-60).

[13] 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)

accepted methodological principles of natural sciences but remained 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, p. 12)

This discrepancy between the "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.[14] 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 of obsolete elements that have already been eliminated from among the tools of the philosophers of science (cf. Kertész, 2004; Kertész and Rákosi, 2005b, 2008a,b, 2012).

(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.[15] As opposed to these insights, (SR) still tries to derive methodological norms of linguistics from the alleged principles of scientific theorising in general.

---

[14] "In the late 1950s, philosophers too began to pay more attention to actual episodes in science, and began to use actual historical and contemporary case studies as data for their philosophizing. Often, they used these cases to point to flaws in the idealized positivistic models. These models, they said, did not capture the real nature of science, in its ever-changing complexity. The observation language, they argued, could not be meaningfully independent of the theoretical language since the terms of the observation language were taken from the scientific theory they were used to test. All observation was theory-laden. Yet, again, trying to model all scientific theories as axiomatic systems was not a worthwhile goal. Obviously, scientific theories, even in physics, did their job of explaining long before these axiomatizations existed. In fact, classical mechanics was not axiomatized until 1949, but surely it was a viable theory for centuries before that. Further, it was not clear that explanation relied on deduction, or even on statistical inductive inferences. […] All the major theses of positivism came under critical attack. But the story was always the same – science was much more complex than the sketches drawn by the positivists, and so the concepts of science – explanation, confirmation, discovery – were equally complex and needed to be rethought in ways that did justice to real science, both historical and contemporary. Philosophers of science began to borrow much from, or to practice themselves, the history of science in order to gain an understanding of science and to try to show the different forms of explanation that occurred in different time periods and in different disciplines. Debates began to spring up about the theory ladeness of observation, about the continuity of scientific change, about shifts in meaning of key scientific concepts, and about the changing nature of scientific method. These were both fed by and fed into philosophically new areas of interest, areas that had existed before but which had been little attended to by philosophers." (Machamer, 2002, p. 6-7)

[15] "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,

At this point, of course, the question emerges of what linguists should do. The 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. Despite this, it appears to be highly *risky*. First, the silent majority of linguists do not reflect on methodological issues. 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 possible.

This is the problem around which this paper centres on. We formulate it as (P1):

(P1) Is metascientific reflection in linguistics possible, and can it be a fruitful enterprise?

(P1) cannot be answered within the limits of the present paper. Therefore, it has to be narrowed down to a more specific problem. Since one of the most important developments in linguistics is the acceptance and application of a wide range of data types, and experiments count as frequently used and extremely valuable data sources, (P2) presents itself as an interesting and instructive special case of (P1):

(P2) Is metascientific reflection on the nature and limits of experiments as data sources in linguistics possible and can it be a fruitful enterprise?

We will argue for the following solution to (P2):

(H2) 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 on the other:

- (a) Results of methodological reflection on experiments carried out by philosophers of physics, psychology, social sciences, biology etc. have to be analysed to determine whether there is *analogy* between them and the situation in linguistics.
- (b) *Research practice* as well as the *self-reflection of linguists* has to be taken into consideration.
- (c) Methodological guidelines or principles have to be in accord with *a general account of linguistic theorising* that covers not only specific issues related to the treatment of experimental data but comprises the whole process of theory formation.

In order to make (H2) plausible, we will proceed as follows. In Section 2, we will provide a concise overview of recent views on the nature and limits of experiments in physics and of the relationship between experimental data and theories. In Section 3, we will present a case

---

"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, p. 9-11)

study in order to illustrate to what extent the findings of Section 2 can be applied to linguistics. In Section 4, we will summarise our results and draw some tentative general conclusions.

2. Experiments in physics

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, p. 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)[16] 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, 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,[17] and it has to be spelled

[16] For example: the atomic mass of silicone, neutron currents, recessive epistasis, Broca's aphasia.
[17] Cf. Bogen (2002, p. 135).

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, not only the results of the experiment have to be put forward but 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 for deciding by the scientific community 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, p. 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 experiments cannot be reduced to the examination of the end products of the experimentation process but 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 and Woodward, 1988, p. 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 results of a series of measurements (a set of perceptual data) and the expected phenomenon.[18] 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, p. 230)

Individual measurements are always influenced by measurement errors. While *random errors* are unpredictable but with statistical methods controllable, *systematic errors* systematically distort the results. It is very difficult to reveal their presence because they bias every single measurement in the same way, into 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.[19] 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

---

[18] "[…] 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 and Woodward, 1988, p. 308-309; emphasis as in the original)

[19] 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.

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, p. 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 the 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.

(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, p. 276-278; 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, p. 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 methodological norms determine to a considerable extent what in "normal science" happens: 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*.[20] To reduce this danger, Franklin (2002, p. 3-6, 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;[21]
- 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, p. 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 hypotheses of 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.

---

[20] 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, p. 236)

[21] 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.

> 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, p. 159)

Thus, there is a danger that the researcher eliminates relevant information from the published report and important decisions remain without 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, p. 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 lost 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.[22] 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

---

[22] "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, p. 165-166; emphasis as in the original)

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-ladeness 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:

> "[…] 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, p. 307)

For the role of plausible inferences in science, see also Rescher (1976, 1987) and Kertész and Rákosi (2012).

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, p. 122-123)

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, p. 3; emphasis as in the original)

This paper's 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 the motto 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:[23]

> "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, p. 56)

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 the there is circularity or not. This question, as Kertész and Rákosi (2009) has shown, can only be decided *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, p. 119; emphasis added), since a modified, prismatically re-evaluated, qualitatively new information state is created (see also points *(ix)–(xi)* in Section 4.1 in Kertész and Rákosi 2009, Section 10.4 in Kertész and 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 data or other kinds of information that should be regarded as relevant, it does not examine alternative interpretations, it leaves relevant factors unclarified which could decrease the reliability of the results (data, hypotheses) stemming from them, or it contains statements which are plausible but which constitute an inconsistent set while no attempt is made at the resolution of this inconsistency, then we are facing 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 and Rákosi 2012), then it is the *experimenter's cycle* – an effective and fruitful enterprise of gaining new information about the world.[24]

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

---

[23] See also Pickering (1989).

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

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.

3. A case study: an experiment on metaphor processing

In Section 2, we have seen that on the basis of the study of the *praxis* of physical experiments, the current 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 in scientific argumentation. 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 linguistics as well. Therefore, with the help of a case study about a psycholinguistic experiment carried out by Keysar, Shen, Glucksberg and Horton we will examine whether there is an *analogy* between physical and psycholinguistic experiments. In Section 3.1, we will present the sketch of the experimental design of Keysar et al. (2000). Then, in Sections 3.2-3.8, 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.

3.1. Experimental design

Keysar et al. intended to test whether metaphorical expressions are comprehended by relying on conceptual mappings as Lakoff and Johnson's Conceptual Metaphor Theory 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, p. 579-580)

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;[25]
2. *no-mapping scenario*: conventional instantiations of the supposed mapping are replaced by expressions not related to the given mapping;[26]
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;[27]
4. *literal-meaning scenario*: renders the target expression as literal.[28]

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, p. 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.

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.[29] 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.

---

[25] 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***.

[26] 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***.

[27] 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***.

[28] 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***.

[29] 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***.

3.2. 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, 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 isolated 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.

3.3. 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.

3.4. 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:

- Target expressions were created as novel instantiations of conceptual metaphors listed in Lakoff and 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 as well (see Figure 1).

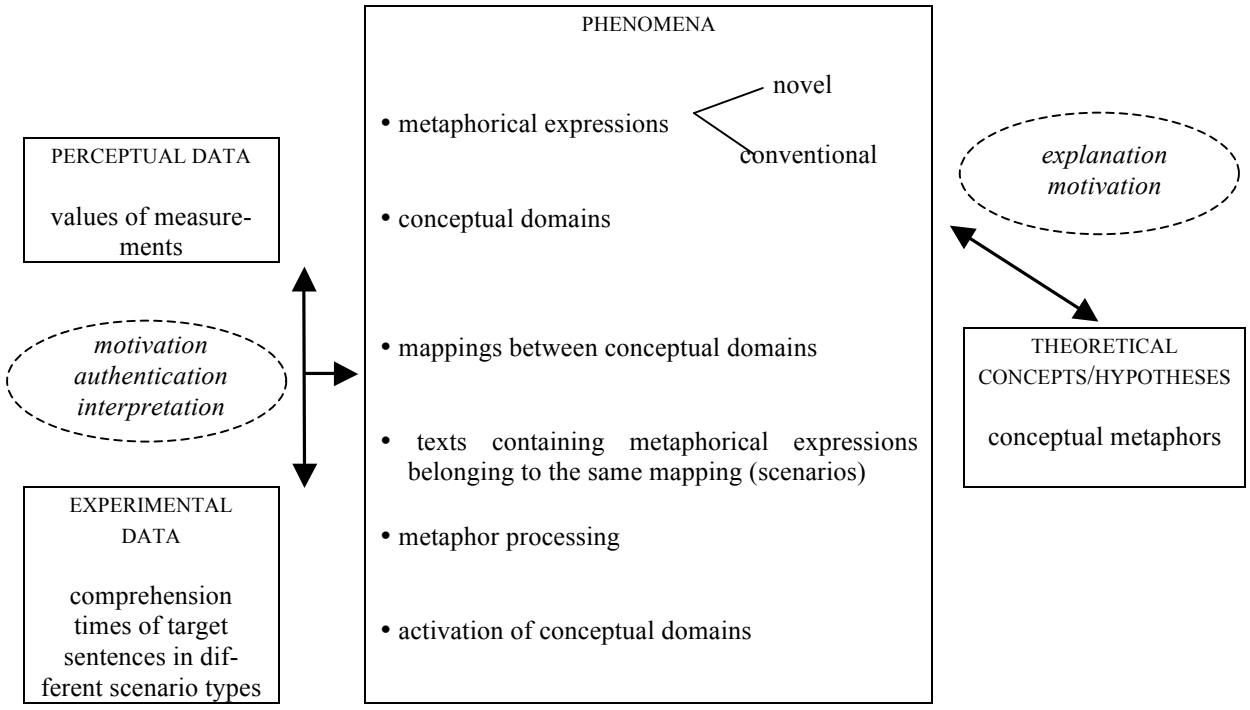

Figure 1: Phenomena and their relationship to data and hypotheses

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 can 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 and 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, p. 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 and 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, p. 588; emphasis added)

This means that they regard the activation of the source domain of metaphors as part of the processing of novel metaphors, but they do not consider the mapping between the two conceptual domains involved to be an activation of a stable conceptual metaphor but 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, p. 591)

For more on this, see Section 3.7.

3.5. 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, p. 583);

– whether novel-mapping scenarios contain significantly less conventional metaphorical expressions than implicit-mapping scenarios do (cf. Keysar et al., 2000, p. 585-586);
– 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, p. 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 and 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 in every scenario.

For further problems, see Section 3.7.

3.6. 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, clearly significant difference between the mean reading times of implicit-mapping scenarios in the two experiments, while with no-mapping scenarios, the difference is appreciable but may be not significant, and in literal-mapping condition, the values are almost identical.

b) Thibodeau and 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, respectively, indicates that the experiment is sensitive enough 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 raises some problems as well:

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 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 and Bowdle (2008, p. 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 and Johnson (1980), the wording of the scenarios was questioned by many researchers from different points of view.[30] 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 in Keysar et al. (2000, p. 582) and in Thibodeau and Durgin (2008, p. 525) fortify 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 and 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.

---

[30] "[…] 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 and Durgin, 2008, p. 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, p. 286; see also Gibbs and Lonergan, 2007, p. 78-79)
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 judged as correct.

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 and Gentner (2005, p. 204-206) who refer to earlier results as well as their own experiments which indicate that comprehension times of metaphors are influenced by familiarity and aptness besides conventionality. These factors should also be accounted for by the planning and evaluating of the experiments.

f) Thibodeau and 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 and 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 other considerations than was the case with the original experiment. Namely, Keysar et al's main motivation was to make sure that "participants would not anticipate or notice a particular pattern" (Keysar et al., 2000, p. 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 and Durgin, 2008, p. 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 and Durgin (2008, p. 529-531) 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 is still no 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.[31] Third, the similarity in reading times of literal targets and metaphorical ones should be accounted for in this case, too. Fourth, Gentner and Boronat's (1992) experiments were in accord with Keysar et al's findings and not with Thibodeau and Durgin's (see also Gentner and 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 and 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.

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 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*.[32] 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

[31] 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.

[32] "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 and Bowdle, 2008, p. 124; emphasis as in the original)

scenarios.[33] 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 diverse 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 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.

3.7. Interpretation of the perceptual data

As we have seen in Section 3.4 (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 – that is, establishing the link between them and the phenomena – is clearly theory-laden, that is, it involves many hypotheses that cannot be checked empirically in a direct way. Thus, for example, metaphoricity of expressions, the classification of expressions into metaphor families or metaphorical mappings involves subjective, arbitrary elements stemming from the intuitive judgements of the experimenters' that 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 the participant *X*, the value 1632 ms was gained in the 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 enough to explain why the value 1632 ms was gained in the 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.

---

[33] 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.

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

3.8. 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, p. 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.

Since 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.[34] At this point, the theoretical model

---

[34] 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

should have been improved, and with this, the experimental design should have been developed.

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.

4. Summary

The case study presented in Section 3 revealed, among other things, the following *similarities* between experiments in physics and in psycholinguistics:

- 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-ladeness 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 other ones 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.

---

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 would be akin to parallelprocess 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 and Gentner, 2005, p. 199; emphasis as in the original)

– 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.
– 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 and Boroditsky, Keysar et al. and Thibodeau and Durgin, as well as the papers cited where 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; it cannot be described as a continuous evolution of the results and theories, either. Rather, it indicates that previous and already rejected hypotheses or explanations may revive and be improved.

These similarities support (H2)(a) to a considerable extent, making our proposal a plausible alternative in contrast to (GR) and (SR):

(H2) 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 on hand, and insights gained by philosophers of science studying experiments in other disciplines:

(a) Results of methodological reflection on experiments carried out by philosophers of physics, psychology, social sciences, biology etc. have to be analysed to determine whether there is *analogy* between them and the situation in linguistics.

There are, of course, some *differences* between experiments in physics and psycholinguistics as well. The most important are perhaps the following:

– Physics divides into experimental and theoretical branches, while in 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, in favour of the theory at issue or against a rival theory.
– In physics, raw experimental data (perceptual data) are often publicised; in linguistics, that is practically never the case.
– In physics, experiments are almost always repeated; in linguistics, this happens only if the results are questioned. Moreover, during the replication the experimental setting applied is often not the same, but only similar.
– In 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-ladeness in linguistics means in most cases also theory-dependence. In physics, however, experimental data usually contain lower-level theoretical concepts.

– In 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.

These differences indicate that psycholinguistic experiments are not identical with physical experiments – there is only a *strong analogy* between them and between the epistemological problems they raise as well. This finding underlines the importance of the elaboration of detailed case studies in metalinguistic research, and reinforces (H2)(b):

(H2) (b) *Research practice* as well as the *self-reflection of linguists* has to be taken into consideration.

Since in 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 have to study and model linguistic theorising as well. In this way, we obtain (H2)(c):

(H2) (c) Methodological guidelines or principles have to be in accord with *a general account of linguistic theorising* that covers not only specific issues related to the treatment of experimental data but comprises the whole process of theory formation.

A theoretical framework which might be suitable for fulfilling the task of modelling both linguistic theorising and linguistic experimenting is the *p-model* by Kertész and Rákosi (cf. Kertész and Rákosi 2012). Since the p-model describes scientific theorising as a cyclic and prismatic, heuristic argumentation process, and current literature on experiments renders experiments as cyclic and prismatic processes, the integration of the two models seems to be natural. Metascientific reflection along these lines may make linguistic research more effective. It may contribute to the elaboration of a workable methodology which is not alien to the practice of linguistic research but lays down clear criteria. This methodology does not consist of mechanically applicable, general rules but requires the active and (self-)reflective contribution of linguists. If the experimenter in linguistics elaborates the experimental design carefully, conducts the experiment, then compares the results with his/her theoretical expectations, reveals inconsistencies, examines the widest range of alternative explanations, and tries out diverse modifications of these elements – that is, if he/she looks actively for the weak points in his/her experiment –, then experiments will become more reliable and more comprehensive. Of course, the more one knows about the nature of experiments and scientific theorising and about problem solving techniques related to their components, the better chances one has. To achieve this, an important step would be to break with the custom that experiments function as *argumentum ad verecundiam*. Therefore, perceptual data and the whole stimulus material should be made public – but also the doubt, uncertainty, the unclarified points, etc. The revealing and overt discussion of such seemingly unimportant or irrelevant issues would not indicate the immaturity of linguistics or the incompetence of the researcher but could result in a rapid development of the field.

Acknowledgement
Work on the present paper was supported by the Research Group for Theoretical Linguistics of the Hungarian Academy of Sciences, the projects OTKA NI 68436 and TÁMOP 4.2.1./B-09/1/KONV-2010-0007 as well as the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Literature

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. Routledge, London & New York, pp. 159-170.

Bogen, J., 2002. Experiment and observation. In: Machamer, P., Silberstein, M. (Eds.), The Blackwell guide to the philosophy of science. Blackwell, Malden & Oxford, pp. 128-148.

Bogen, J., Woodward, J., 1988. Saving the phenomena. The Philosophical Review XCVII(3), 303-352.

Borsley, R.D., 2005. Introduction. Lingua 115, 1475-1480.

Bowdle, B.F., Gentner, D., 2005. The Career of Metaphor. Psychological Review 112(1), 193-216.

Cantor, G., 1989. The rhetoric of experiment. In: Gooding, D., Pinch, T., Schaffer, S. (Eds.), The uses of experiment. Studies in the natural sciences. Cambridge University Press, Cambridge, pp. 159-180.

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 University Press, New York, pp. 51-94.

Chomsky, N., 2002. On nature and language. Cambridge University Press, Cambridge.

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

Consten, M., Loll, A., 2010. Circularity effects in corpus studies – why annotations sometimes go round in circles. This issue.

Deignan, A., 2008. Corpus linguistics and metaphor. In: Gibbs, R.W. (Ed.), The Cambridge handbook of metaphor and thought. Cambridge University Press, Cambridge, pp. 280-294.

Franklin, A., 2002. Selectivity and discord. Two problems of experiment. University of Pittsburgh Press, Pittsburgh.

Franklin, A., 2009. Experiments. Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/physics-experiment/.

Galison, P., 1987. How experiments end. Chicago University Press, Chicago.

Geeraerts, D., 2006. Methodology in cognitive linguistics. In: Kristiansen, G., Achard, M., Dirven, R., Mendoza Ibáñez, F.J.R. (Eds.), Cognitive linguistics: Current approaches and future perspectives. Mouton de Gruyter: Berlin & New York, pp. 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 University Press, Cambridge, pp. 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. MIT Press, Cambridge, pp. 199-253.

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. Peter Lang, Frankfurt a.M. & New York & Oxford, pp. 71-90.

Glucksberg, S., 2001. Understanding Figurative Language: From Metaphors to Idioms. Oxford University Press, Oxford.
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.
Gooding, D.C., 2000. Experiment. In: Newton-Smith, W.H. (Ed.), A companion to the philosophy of science. Blackwell, Malden & Oxford, pp. 117-126.
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. University of Chicago Press, Chicago, pp. 29-64.
Kepser, S., Reis, M., 2005. Evidence in linguistics. In: Kepser, S., Reis, M. (Eds.), Linguistic evidence. Empirical, theoretical and computational perspectives. de Gruyter, Berlin & New York, pp. 1-6.
Kertész, A., 2004. Philosophie der Linguistik. Narr, Tübingen.
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. Lang, Frankfurt am Main & Bern & Bruxelles & New York & Oxford & Wien, pp. 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. Lang, Frankfurt am Main & Bern & Bruxelles & New York & Oxford & Wien, pp. 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. Lang, Frankfurt am Main & Bern & Bruxelles & New York & Oxford & Wien, pp. 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 University Press, Cambridge.
Keysar, B., Shen, Y., Glucksberg, S., Horton, W.S., 2000. Conventional language: How metaphorical is it? Journal of Memory and Language 43, 576-593.
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. Blackwell, Malden & Oxford, pp. 1-17.
Nickles, Th., 1989. Justification and experiments. In: Gooding, D., Pinch, T., Schaffer, S. (Eds.), The uses of experiment. Studies in the natural sciences. Cambridge University Press, Cambridge, pp. 299-333.
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 University Press, Cambridge, pp. 275-297.
Pullum, G.K., 2007. Ungrammaticality, rarity, and corpus use. Corpus Linguistics and Linguistic Theory 3, 33-47.

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.
Sampson, G., 1975. The form of language. Weidenfeld & Nicholson, London.
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.
Schütze, C.T., 1996. The empirical base of linguistics. Grammaticality judgments and linguistic methodology. The University of Chicago Press, Chicago & London.
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.
Stefanowitsch, A., Gries, S.Th. (Eds.), 2007. Grammar without grammaticality. Special issue of Corpus Linguistics and Linguistic Theory.
Sternefeld, W. (Ed.), 2007. Data in generative linguistics. Theoretical Linguistics 33(3), 269-413.
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.