

# *CLCWeb: Comparative Literature and Culture*

ISSN 1481-4374 <<http://docs.lib.purdue.edu/clcweb>>

Purdue University Press ©Purdue University

*CLCWeb: Comparative Literature and Culture* (ISSN 1481-4374), the peer-reviewed quarterly of scholarship in the humanities and social sciences, is published by Purdue University Press ©Purdue University online in full text and in open access. The journal publishes scholarship following tenets of the disciplines of comparative literature and cultural studies designated as "comparative cultural studies" in a global, international, and intercultural context and with a plurality of methods and approaches: papers for publication are invited to <<http://docs.lib.purdue.edu/clcweb/submit.html>>; for the aims and scope of the journal consult <<http://docs.lib.purdue.edu/clcweblibrary/clcwebaims>>; for the journal's style guide consult <<http://docs.lib.purdue.edu/clcweblibrary/clcwebstyleguide>>. In addition to the publication of articles, the journal publishes review articles of scholarly books and publishes research material in its *Library Series* <<http://docs.lib.purdue.edu/clcweblibrary/library>>. Work published in the journal is indexed in the Annual Bibliography of English Language and Literature, in the Arts and Humanities Citation Index, in Humanities International Complete, and in the International Bibliography of the Modern Language Association of America. *CLCWeb* is member of The Council of Editors of Learned Journals <<http://www.celj.org>> and it is listed in the Directory of Open Access Journals. *CLCWeb* is mirrored on the website of the British Comparative Literature Association <<http://www.bcla.org/clcweb/>>, it is preserved at research libraries in the Stanford University lockss system <<http://www.lockss.org/lockss/>>, and it is archived in the *Electronic Collection of Library and Archives Canada* <<http://www.collectionscanada.ca/electroniccollection/>>. *CLCWeb: Comparative Literature and Culture* is affiliated with the Purdue University Press hard-copy monograph series in Comparative Cultural Studies and selected papers of the journal are published in thematic annuals in the series <<http://www.thepress.purdue.edu/comparativeculturalstudies.html>>. Contact: <[clcweb@purdue.edu](mailto:clcweb@purdue.edu)>

---

## ***CLCWeb* Volume 2 Issue 3 (September 2000) Article 4**

### **Johan F. Hoorn, "The Hazard of Hidden Interactions: A Reanalysis of Designs in Reaction-Time Studies on Metaphor"**

<<http://docs.lib.purdue.edu/clcweb/vol2/iss3/4>>

---

## **Contents of *CLCWeb: Comparative Literature and Culture* 2.3 (2000)**

<<<http://docs.lib.purdue.edu/clcweb/vol2/iss3/>>>

---

**Abstract:** In his article, "The Hazard of Hidden Interactions: A Reanalysis of Designs in Reaction-Time Studies on Metaphor," Johan F. Hoorn argues that research designs in empirical literature and the psychology of aesthetics often include unanalyzed factors. The nature of these factors may be linguistic such as word frequency or lexical ambiguity or technical such as presentation order, repeated measures, etc. By not correctly analyzing an experiment, higher-order interactions may go unnoticed, while interfering with results. Hoorn reviews a sample of reaction-time experiments on metaphors, some of which are considered key studies in the area. Because the quality of an argument depends on the quality of the experiment, Hoorn places emphasis on designs and statistics. He then discusses the consequences of improper analysis for the theory of metaphor processing.

**Johan F. HOORN**

## **The Hazard of Hidden Interactions: A Reanalysis of Designs in Reaction-Time Studies on Metaphor**

### **Introduction**

The study of literature attaches much worth to putting theory to the test. In metaphor theory also a great deal of experimentation is found, particularly among cognitive psychologists with an interest in (literary) language and among literary scholars with an interest in psychological methods. Although the experiments may obtain valuable information, much of its value is lost or interpreted wrongly by improper use of the statistical tools of analysis. In this article, I explain that many experiments that supposedly have a 2x2 design, actually used more factors or factor levels, and thus contain a higher-order interaction. For the research question, such interactions seem unimportant. They often are resulting from trivial manipulation: The order of presenting the stimuli, or the order of using the preferred hand in a reaction-time (RT) choice task. Yet, I will point out that higher-order interactions may cast a completely different light on what seems a straightforward result.

The rationale of my article is quite simple. Suppose a group of expert readers is contrasted with a group of novices to see whether education improves the understanding of postmodern poetry. It turns out that experts find postmodern poems better to comprehend than novices. Of course, experts are better educated, and practice facilitates comprehension. However, this interpretation is based purely on analysis of a main effect (experts vs. novices). Suppose, however, that the results show no difference for sex in novices but do show that female experts comprehend postmodern poetry much better than expert men do. It now turns out that comprehensibility only increased for expert *female* readers. In other words, education only facilitates comprehension for women. Men stay as stupid as they were. Thus, the interaction between expertise and sex strongly limits the generalization that education improves comprehensibility. Naturally, most researchers do not neglect the possible effects of such important factors as sex and age. However, less obvious factors are ignored all too often, although they may limit findings in quite the same way. It may be objected that no matter how refined an analysis is, there is always one extra factor one can come up with that limits the external validity of a study. True as this may be, it is not too much asked if factors that are explicitly mentioned in the design or that can be derived from it are treated in a full factorial model. The suggestion of eclecticism should be avoided.

### **The Object: Time Course of Metaphor Processing**

First-order interactions (e.g., education x sex) may modify main effects, but may be modified themselves by higher-order interactions (only old, fair-haired expert women comprehend postmodern poems better: Education x sex x age x hair x color). In metaphor research, similar constellations occur. Gibbs investigated whether the presence or absence (factor 1) of a literal or metaphoric context (factor 2) facilitated the literal or metaphoric paraphrases (factor 3) of idioms. Glucksberg, Brown, and McGlone (1982) measured the comprehension speed for two different idioms (factor 1) in two different story contexts (factor 2) with two different actors (factor 3), and two different referents (factor 4) as *dramatis personae*. However, such designs are not always analyzed exhaustively. Nevertheless, the results derived from incomplete analyses count as evidence in developing metaphor theory (see Hoffman and Kemper). The question put forth by RT-studies on metaphor is whether the literal interpretation of a metaphor ("man is a machine") is executed before the figurative interpretation. Different from literal expressions ("man is a mammal"), metaphors would arouse an "anomalous moment" when the literal stage fails and the figurative stage is initiated (for counterevidence, see Hoorn 1997). Therefore, the idea behind many reading-time and reaction-time studies is that literal expressions are processed faster than metaphors. My aim is to evaluate certain experimental attempts made in this field. It is certainly not an exhaustive review nor does it underestimate the theoretical contributions that some of these studies make. Obviously, examples also exist of proper analysis in RT-research on metaphor. Nonetheless, the following sample shows that uncarefully analyzed designs are not uncommon.

## Examples

In an illuminating review of the pitfalls of RT-studies on metaphor, Hoffman and Kemper tick off hazardous matters such as neglect of subject strategies, task biases, individual differences, and natural settings. There may be one more added: Improper analysis of research designs. Ortony, Schallert, Reynolds, and Antos (1978) carried out two RT-experiments, in which contexts primed the literal or metaphoric meaning of metaphoric and idiomatic expressions. Idioms (e.g., "a pain in the neck") are metaphors that have become so conventional that their literal meaning is subordinate. In experiment 1, the time was recorded to understand an expression. Context length (short vs. long) was compared with prime type (cuing the literal or metaphoric meaning of an expression). According to a serial two-stage anomaly model, expressions primed as metaphors were supposed to be processed more slowly than those primed as literals. Moreover, it was expected that this difference would only occur in short contexts. Longer contexts were supposed to provide so many cues, that the relevant stage was evoked directly and, hence, any difference would be eliminated. Thus, the seriality of the two-stage model was supposed to be confirmed in short contexts, not in long ones. The design combined two lists of expressions, two presentation orders and two contexts (short vs. long) as between-subject factors. Literal or metaphoric priming was a within-subject factor. Presentation order had no effect, and was excluded from further analysis. Unfortunately, a potential effect of expression list was not reported. Analysis of variance of the remaining factors indicated that metaphoric interpretations took longer than literal interpretations in the short contexts, whereas they took about equal time in the long contexts. So far, so good.

However, Ortony et al. conducted a second experiment in which the manipulation was more evasive. Again, the time taken to understand an expression was recorded. Three between-subject factors of expression type order (idioms first or last), expression list and presentation order were employed in combination with two within-subject factors: Context type (long vs. short) and expression type (idioms vs. literal). The between-subject factors proved insignificant, and need not concern us here. The numbers of expressions in the two within-subject factors of context type and expression type were not counterbalanced, so that the effects were confounded with practice. Two stories served as contexts (story A and B), while expressions could be idiomatic or literal. However, the idioms were connected to both stories, whereas the literal expressions were connected only to story A. Thus, the missing cell was literal expressions in story B. The metaphoric or literal priming was also unbalanced. Story A was constructed such that it primed the metaphoric meaning of the idioms, and the literal meaning of the literal expressions. Story B, however, only primed the literal meaning of the idioms. In other words, apart from prime type, another factor of story type should have been devised, so that story A and B would have primed the literal or metaphoric meaning of idioms and literal expressions. Another solution would have been to omit one of the stories, or to weigh the results of story A as half.

These studies by Ortony et al. initiated a series of RT-experiments on metaphor and idiom processing. For instance, Gibbs (1980) conducted an experiment on idioms ("he's singing a different tune"). As mentioned above, idioms are ambiguous stimuli, because they have a literal and a metaphoric meaning. The metaphoric meaning has become conventional in the standard language, whereas the literal meaning is hardly ever used. Gibbs investigated whether the conventional metaphoric meaning was processed faster than the unconventional literal meaning, even when the literal meaning was primed by the context. In one subject group, idioms were used as prime, which were followed by literal or metaphoric paraphrases of these idioms. In the second subject group, idioms were embedded in contexts, cueing either the literal or metaphoric meaning of the idiom. Here also, literal or metaphoric paraphrases followed the idioms. To make sure that the first group would not interpret the idioms only in their metaphorical sense, the idioms were mixed with literal expressions, which also received literal or metaphoric paraphrases. RT was registered for true-false decisions after paraphrase presentation. The results suggested that conventional (metaphoric) uses of idiom were processed faster than unconventional (literal) uses, despite their metaphoric origin and despite priming the literal meaning. To summarize, idioms and literal expressions without context preceded paraphrases that could be either literal or metaphoric,

and idioms in a literal or metaphoric context preceded paraphrases that could be literal or metaphoric. Gibbs analyzed this design as a standard 2x2 of condition (context vs. no-context) and paraphrase (literal vs. metaphoric). Yet, the design was a 2x2x2 of condition, paraphrase (literal vs. metaphoric), and prime type (literal prime vs. metaphoric prime). To know whether the prime type of the contexts was literal or metaphoric, a new group of subjects rated the priming of the contexts, whereas the priming of the idioms and literal expressions was not rated. Thus, the prime type in no-context might not have been the same as the prime type in context. Yet, they were treated as comparable in the design, whereafter only the effects of idioms - not of the literal expressions -- were analyzed. Important information was lost concerning idioms with literal primes and metaphoric paraphrases, and those with metaphoric primes and literal paraphrases. These could have told the power -- within subjects -- of the literal and metaphoric primes in the interaction of (condition by) prime type by paraphrase.

Glucksberg, Gildea, and Bookin (1982) administered a within-subject experiment on metaphors and literals in a standard sentence verification task (true/false). Literals could be correct ("Standard True") or incorrect ("Standard False"); accordingly, metaphors could be correct ("Metaphors") or incorrect ("Scrambled Metaphors"). The latter can be viewed as instances of anomalies. Subjects judged the truth value of the expressions, and the RT from expression onset to button press was measured. It was argued that if the figurative meaning of correct metaphors can be ignored, 'false' decisions should be equally fast for correct and incorrect metaphors. Conversely, if the figurative meaning is accessed automatically, "false" decisions for correct metaphors should be slower than for incorrect metaphors, because an extra figurative check is needed. It was found that RT for correct literals was fastest -- followed by incorrect literals and incorrect metaphors -- whereas correct metaphors were slowest. It seemed that metaphors were evaluated on their figurative truth, which is consistent with the two-stage model. Incorrect literals and incorrect metaphors were perceived as equally anomalous and could be rejected as not literal. However, the study has a number of pitfalls. The number of literals and metaphors were not balanced. The correct literal expressions had 80 items, the incorrect 40, the incorrect metaphors had 20 items, and the correct metaphors also 20. As indicated above, incorrect literals and incorrect metaphors took about an equal time to be rejected as "false." In other words, the incorrect expressions may all have been perceived as anomalies. If so, it could be that subjects perceived 80 literals, 60 anomalies, and 20 metaphors. This range coincides with the ordinal pattern that was found in the mean RTs. Thus, reactions may become slower as expressions occurred less frequently. The authors suggested that the semantic relationship between terms was probably less strong for metaphors than for literals, so that metaphors were processed more slowly. If so, the authors argued, the RT pattern should remain unchanged even in different contexts. In experiment 2, expressions were provided with context by introducing the quantifiers *Some* and *All*. '*Some* surgeons are butchers' was supposed to be more plausible than "*All* surgeons are butchers." The more plausible or "correct" metaphors were rated for "goodness," whereas the other expression types were not. In the interaction of expression type (literal vs. metaphor) by quantifier, the *Some* metaphors (rated as "good") showed the pattern of experiment 1 that RT for correct literals was fastest -- followed by incorrect literals and incorrect metaphors - whereas correct metaphors were slowest. The *All* metaphors (rated as "less good") did not show this pattern. Thus, context had effect, and it was inferred that metaphor comprehension was not merely a matter of semantic relatedness. This experiment was analyzed as two quantifiers with four expression types. The expression types (Standard True, Standard False, Metaphors, Scrambled Metaphors) were treated as four independent conditions. Yet, an important interaction with correctness was overlooked. Standard True and False are literally correct and incorrect expressions. Metaphors and Scrambled Metaphors are figuratively correct and incorrect. Thus, the design is a 2x2x2 MANOVA for quantifier by expression type by correctness. Differences attributed to expression type may be due to correctness in the interaction. The same is valid for experiment 3, which duplicated this design and analysis.

Gildea and Glucksberg (1983) administered three variations on the above study, which combined three prime types (literal prime, figurative prime, no-prime) with four expression types

(Standard True, Standard False, Metaphors, Scrambled Metaphors). Despite this 3x4 design, Standard True and False were not analyzed in experiments Ia, Ib and II, while the remaining 3x2 was analyzed only by planned comparisons (*t*-tests). Ib was analyzed in a 2x2 ANOVA, ignoring the higher-order interactions outlined earlier.

Estill and Kemper (1982) surveyed idioms (e.g., "climbing the walls") for effects on RT with three cue types and four context types. Subjects were asked to identify the last word ("walls") in an expression, given a particular precue. Cues could be words ("walls") identical to the last word in the expressions, they could be rhyming ("falls") with the last word or they formed the semantic category "part of a building" of the last word. The idioms appeared in literal, figurative, or ambiguous contexts. Expressions that were not idioms but did use the idiom's last word ("knocking out the walls") were presented in non-idiomatic contexts. Subjects reacted as soon as the last word of an expression was encountered in the context. This 3x4 design was properly analyzed and it was found that the main effects of cue type and context type were significant, whereas the interactions were not. The Word Identity cue yielded faster responses than the Rhyme cue, which was faster than the Semantic Category cue. More importantly, non-idiomatic contexts slowed down RT, compared with all other contexts. Thus, the last word was differently processed when it was part of an idiom than when it was not. The authors suggested that idioms were processed as discrete lexical entries. There is a slight inconvenience, however, concerning the last words of the expressions. In the non-idiomatic context, the last word came from an idiom - but was not used in an idiom - as opposed to the other contexts. Thus, the stimulus ratios were not counterbalanced, because three idioms were contrasted with one non-idiom. This may explain the slower RTs for non-idiomatic contexts. Because the argument of automatic access of idiomatic meaning was entirely based on this RT-difference, the effects should have been treated more carefully. Unfortunately, the authors did not provide the exact *p*-values for the *t*-tests on context types. However, the definition of the alpha-level is critical for the acceptance of -- in this case -- automatic access of idiomatic meaning. Context type had four levels (literal, figurative, ambiguous, non-idiomatic), and thus established six related main effects. In other words, the *p*-value should not have been tested with  $\alpha = .05$ , but rather with  $\alpha = 8.33^{-03}$ . Likewise, if the interactions had been significant, three cue types by four context types would have resulted in 18 related interactions, and should have been tested with  $\alpha = .05 / 18 = 2.77^{-03}$ .

Gerrig and Healy (1983) examined within-subject effects of prior and ensuing contexts on reading time for metaphors. Metaphors rated as "good" were contrasted with 'bad' metaphors in active and passive voice sentences, which coincided with prior and ensuing context: "The train followed the parallel ribbons" vs. "The parallel ribbons were followed by the train." Subjects pushed a button as soon as they understood the sentence. Only the main effect of context was significant, indicating that prior context yielded faster reading times than ensuing contexts, or -- as a confounded alternative -- confirming the classic finding that active voices were processed faster than passive voices. Each cell in the analysis contained 8 metaphors. However, the experiment also employed 16 fillers. In other words, the stimuli were not counterbalanced (32 metaphors against 16 fillers). Fillers were not rated on potential metaphoricalness, so that their effects were unpredictable. Filler effects were not analyzed together with the metaphor effects in a (weighted) MANOVA. Because only one random order was used for the stimuli, all subjects received stimuli in fixed order, which may easily affect the data. The same applies to the second experiment. As indicated by the authors, the manipulation confounded ensuing context with passive voice. Therefore, another within-subject design contrasted literals with metaphors in active and passive voice. It was argued that if slow reading was induced by passive voices rather than ensuing contexts, this effect should equally occur for metaphors and literals. If, on the other hand, metaphors increased reading times in passive voice -- whereas literals would remain equal -- the effects in experiment 1 could be attributed to ensuing context, not to passive voices. Significant interactions between (literal-metaphor) and (active-passive voice) that increased metaphor reading times would underscore the idea that ensuing context played a special role in metaphor comprehension. Indeed, such a significant interaction was found. However, the authors indicated that the literal expressions had shorter sentences than metaphors. They asserted that this might

have influenced the relative speed with which literals and metaphors were read, but that this was immaterial for their argument. Thus, the main effect of expression type was irrelevant, and only the interaction counted. Nonetheless, the significant interaction may not only indicate that reading times for metaphors were increased by the ensuing contexts, but also that short sentences (literals) canceled the elongating effect of passive voices. If so, the interaction of expression type and context is confounded again.

Inhoff, Lima, and Carroll (1984) verified the results of Ortony et al. with three experiments. In experiment 1, short contexts primed either the literal or metaphoric meaning of metaphors, whereas in experiment 2, long contexts did. Both experiments also employed unrelated contexts, which were inappropriate combinations of context and metaphor. Sentence reading time and total viewing time on critical words were measured with an eye tracker. The authors claimed that metaphorical meanings were understood as quickly as literal ones in long contexts, but slower in short contexts. Stimulus lists were varied between subjects, while related vs. unrelated context was varied within subjects. Context could prime the literal or metaphoric meaning. Moreover, short and long contexts were compared across experiments. Thus, the complete design was a 2x3x2x2 MANOVA for experiment by list by relatedness by prime type. Yet, two separate analyses for experiment 1 and 2 merely tested three means for the main effects of prime type: Literal vs. metaphoric vs. unrelated. The discussion suggested that there was an interaction of prime type with context, but without any supporting statistical tests. Additionally, the metaphors were presented in the same serial order for each list, so that practice or boredom effects were not recognized. Furthermore, the priming effect of literal and metaphoric contexts was scored for the appropriate metaphors. However, it was not for the inappropriate ones. Experiment 3 investigated the thematic relatedness between context and metaphor on reading time. The short contexts of the previous experiments served as the thematically related condition, whereas newly created "associated-words" contexts served as the unrelated condition. Prime type was literal or metaphoric in both conditions, while expression type could be literal or metaphor. Six lists were a between-subject factor, whereas relatedness, prime type and expression type were varied within subjects. Thus, the design was a 6x2x2x2 MANOVA for list by relatedness by prime type by expression type. Nonetheless, the design was diagnosed as 6 lists by 3 prime types (literal vs. metaphoric vs. associated-words) by 2 expression types, and was yet analyzed by 2x2 ANOVAs for e.g., prime type (literal vs. metaphoric) versus expression type (literals vs. metaphors).

Paivio and Clark (1986) explored the importance of imagery and intelligibility for the processing time of metaphors. Three prime types were used between subjects: A cue for the *A-term* (topic or tenor) a cue for the *B-term* (vehicle or image), and no cue. Metaphors were scored within subjects for the imagery value of the *A-term*, of the *B-term*, and for the intelligibility of the metaphor. Subsequently, these metaphors were orthogonally distributed over high or low *A-term* imagery, *B-term* imagery, and intelligibility. Subjects read the metaphors after the presentation of a prime, and released a button when they were ready to paraphrase the metaphor. In principle, this is a 3x2x2x2 MANOVA for prime type by *A-term* imagery by *B-term* imagery by intelligibility. Yet, prime type was analyzed as a main effect of three individual means, separately from the three scale factors, which were analyzed as a second-order interaction. The main effect of prime type was only significant with stimuli as random factor, so that it was merely an idiosyncrasy of the subject group. The authors noticed that the results were paradoxical. Because the *B-term* is the image in which the *A-term* is perceived, *B-term* imagery should be and actually was a stronger correlate of metaphor comprehension time than *A-term* imagery. However, the expected superiority of priming the *B-term* over priming the *A-term* or no priming did not appear. This result seems curious, given the fact that the third-order interaction was never analyzed.

Janus and Bever (1985) also verified the results of Ortony et al. Literal contexts primed literal meanings, whereas metaphoric contexts primed metaphoric meanings of idioms. The time was measured that subjects indicated that the idiom was understood. RT was slower for metaphoric than for literal meaning, which was interpreted as support for the two-stage anomaly model. Nonetheless, idioms are ambiguous expression types. Actually, they consist of two expression types, i.e., a literal expression and a metaphor, which are easily activated by the appropriate

context. The authors indicated that the idioms in the metaphoric contexts were consistently judged as less predictable than the idioms in the literal contexts. They attributed this effect to the literal meaning of the idioms. However, it was overlooked that the metaphoric meaning was systematically correlated with one set of texts, whereas the literal meaning was systematically correlated with another set of texts. In other words, because they used two different text sets to prime one of the meanings, they could not determine whether the effects were due to divergent idiom meaning or to different success in priming.

Glucksberg, Brown, and McGlone (1993) studied whether conceptual analogies motivated the use and comprehension of idioms in discourse. Two reading time experiments were performed, in which two contexts were combined with two idiom types. Contexts indicated which person (S or C) in the story was the referent of the idiom. The idioms differed for their consistency with the stories (consistent vs. inconsistent). Subjects read the stories for comprehension and indicated that they finished reading by pressing a key. However, the actual manipulation looked differently. Their Appendix B lists two different contexts, two different idioms ("blew top" vs. "bit head off"), two different actors ("S blew top" -- "S bit head off" vs. "C blew top" -- "C bit head off") and two different referents ("S blew S's top" -- "S bit C's head off" vs. "C blew C's top" -- "C bit S's head off"). Obviously, this is not the 2x2 ANOVA that the authors suggested, but rather a 2x2x2x2 MANOVA of context, idiom, actor and referent. Moreover, context 1 was systematically correlated with the combination SS and SC, whereas context 2 was systematically correlated with CC and CS. Thus, the reading time effects for context by idiom were biased by the systematically correlated interaction with actors and referents.

Johnson (1996) contrasted metaphors and similes on comprehension speed for priming sentences. He also recorded the response speed for verifying the appropriateness of target sentences. The assumption was that similes ("deserts are *like* ovens") take an extra cognitive operation to transform them into metaphors again ("deserts are ovens"), so that process time for similes should increase. Two similarly designed experiments were conducted. In the first experiment, three groups of subjects were exposed to a counterbalanced mix (within subjects) of three prime types (similes, metaphors, and literals), which were followed by two target types (matching or mismatching the prime). Dependent of the prime, targets were similes or metaphors. Literal primes served as distractor or filler items. In experiment 1, subjects decided between "logical" or "illogical prime-target combination"; in experiment 2 between "yes" or "no appropriate prime-target combination." The time from prime onset to continuation-key press supposedly reflected prime-comprehension time, while target onset to decision-key press would indicate context-verification time. It was found that metaphoric primes were comprehended significantly faster than simile primes (experiment 1 and 2), although this effect differed significantly per group (experiment 1). For both dependent variables, the design was analyzed with a 3x2x2 repeated measures ANOVA of group (I, II, III) by prime type (simile vs. metaphor) by target type (match vs. mismatch). Note that although the experiments were almost identical, there was an effect of experiment with group. Yet, experiment was not treated as an extra factor. Additionally, the effects of literal expressions in contrast to similes merely were visually inspected in the discussion. Strangely, the target types were included into the analysis of prime-comprehension time although the effects of backward priming were not under investigation. Thus, in analyzing prime-comprehension time, the complete design was a 2x3x3 MANOVA for experiment (1 vs. 2) by group by prime type (simile vs. metaphor vs. literal). Regarding context-verification time, a 2x3x3x2 MANOVA of experiment by group by prime type by target type should have been performed. Moreover, F2-analysis with stimuli as the random factor (see Clark, 1973) was not administered (in multivariate designs, quasi-F would be in place). Bonferroni correction to avoid alpha-level inflation was ignored (see Neter, Wasserman, and Kutner 1990, 160-62, 579-89) and in misreading Cohen, magnitudes of .20 and .34 were presented as large effect sizes instead of small (Cohen considers ES= .20 a small, ES= .50 an average, and ES= .80 a large effect). Put differently, probably none of the reported effects can be considered reliable.

Gentner and Wolff (1997) pointed out the interesting difference between two types of metaphor models: Abstraction-first and alignment-first. In the metaphor "my job is a jail," abstraction-first

models presume that an abstraction (e.g., "confinement") is derived from "jail" and projected onto "job." By matching "confinement" with the representation of "job," the metaphor is verified. Alignment-first models assume that metaphors are understood by the intersection of features and relations between "job" and "jail." Thus, nonidentical features may be connected by fitting relations. Within-subjects, four types of priming stimuli were used. In the *both* condition, both "job" and "jail" were present in the metaphor; in the *base* condition, only "jail" was present; in the *target* condition, only 'job' was present, and in the *blank* condition, both terms missed. Four between-subject groups were used to counterbalance the assignment of metaphors to conditions. Each subject received all 32 metaphors in random order. After the prime, the complete metaphor was presented and subjects started typing an interpretation. Because abstraction-first models assume that subjects start with interpreting the base, the *base* condition should yield faster comprehension times than the *target* condition. Alignment-first would predict that seeing *both* base and target should be faster than all other conditions. In experiment 1, interpretation time was measured as the time to begin typing the interpretation after stimulus onset. In experiment 2, subjects pressed the spacebar before starting to type. Moreover, the inter-stimulus-interval (ISI) was increased. Evidence was found for the alignment-first model, because seeing *both* terms first led to the fastest interpretation times. However, this is also the condition that supplies most information altogether, so that any information theory would predict the same facilitation.

Further, the within-subjects factor of prime type had four levels (*both*, *base*, *target*, *blank*). The four between-subjects groups had different blocks of stimuli (four times eight). Only if no interactions of group by prime type can be expected, this design may be analyzed by a one-way repeated measures ANOVA. In experiment 2, however, the authors also were concerned with increasing ISI and the way interpretation speed should be measured. Moreover, the *blank* condition utilized the word "something" instead of blanks. Yet, no interactions with experiments were calculated to test these concerns. Therefore, the complete design was not a one-way ANOVA, but a 2x4x4 MANOVA of experiment (1 vs. 2) by subject group (1-4) by prime type (*both*, *base*, *target*, *blank*). Note that for experiment 1,  $N=60$  is likely a misprint. The  $F$ -distribution for the interaction between prime type and subject groups shows that  $df_1=3$  and  $df_2=117$ . The value of 117 is probably based on 3 times (40 minus 1) subjects, so that similar to experiment 2 and 3,  $N=40$ . Experiment 3 used the same design but different stimuli. *Bases* were highly conventional, and low in relational similarity to the *target* (subject ratings). This manipulation favored the abstraction-first model in that base-primers accelerated interpretation speed more than *target* primes. Yet, again this design was analyzed with a one-way repeated measures ANOVA, whereas the complete design was a 4x4 MANOVA of subject group (1-4) by prime type (*both*, *base*, *target*, *blank*). It may even be argued that the effects of stimulus types should have been tested between experiments. In experiment 4, metaphors were rated by norm groups for conventionality and relational similarity. Interpretation speed was measured for four subject groups in which stimuli were counterbalanced (see above). A 2x2x4 analysis of variance was run for conventionality (high vs. low) by relational similarity (high vs. low) by prime type (*both*, *base*, *target*, *blank*). These factors were within subjects, so that multiple analysis of variance should have been in place but the report is not clear here. Again, the interactions of group by treatment were not examined. Luckily, the Bonferroni method of error protection was used to further explore the differences implied by the significant interactions. This makes the choice for the Neumann-Keuls procedure in the previous three experiments rather odd. In repeated measures with more than three conditions, the error variances usually are unequal. In such cases, therefore, the Neumann-Keuls procedure is not recommended (see Hsu 1996, 127).

### Discussion

As a general remark, many studies reviewed above mainly explored the two-stage model with idioms. However, idioms are both literal and metaphoric expressions that share the surface form, which makes them improper stimuli to test the two-stage model. First, using idioms presupposes that the two-stage model is context-dependent, otherwise the manipulation with literal and metaphoric primes would be ineffective. Second, *any outcome based on idioms is congruous with a context-dependent two-stage model*. If idioms are fast with literal primes, they were supposedly

perceived as a literal expression, so that a second figurative stage was not necessary. If idioms are slow in literal contexts, they were perceived as metaphors, so that figurative interference took place. If fast in metaphoric contexts, they were metaphors, the figurative stage of which was entered immediately. If they are slow in metaphoric contexts, it mainly shows that context was unimportant to the two-stage model, because the literal stage was completed before the second was initiated. Thus, before including idioms in the stimulus set, it should be demonstrated that the two-stage model is sensitive to context. This could be done by crossing purely literal and purely metaphoric expressions with literal and metaphoric contexts (an extra level of no-context would be even better). If the two-stage model is context-independent, one effect should be significant: A main effect of expression type should underscore that literals are always faster than metaphors. If the model is context-dependent, the literal context shows that literals are faster than metaphors, which effect should change in metaphoric contexts. The direction of the change in the metaphoric context is the interesting part. If metaphors are less slow in metaphoric than in literal contexts, then indeed there are two stages, the second of which is facilitated by the proper prime. If metaphors are as fast as literals in metaphoric contexts, then the two serial stages in literal contexts become parallel in metaphoric ones. If metaphors are faster than literals in metaphoric contexts, there are not two stages within one process, because either a literal stage or a figurative stage was executed, dependent on the context prime. All other patterns are also evidence against the two-stage model. Idioms can be used only in combination with purely literal and purely metaphoric expressions. They are not suitable to test a metaphor model, because they are ambiguous. Idioms may only be used to investigate the conventionality of their literal or metaphoric meaning. Dependent of the pattern they follow in literal and metaphoric contexts, they may be processed as literal or metaphoric expressions and pass through one or more stages. If they behave as literal expressions, idioms have become conventional uses of figurative meaning. If they behave like metaphors (perhaps in metaphoric contexts), the conventional use is affected by figurative interference. Similes (comparisons with the preposition *like*) are a more suitable expression type (see Johnson, 1996). In contrast with metaphors (*A is B*), the preposition (*A is like B*) may immediately launch the figurative or relation stage, so that the fixed seriality of stages is limited by the linguistic indicator *like*. In that case, decisions for literal expressions in simile version ("the sun is *like* a star") should be more confusing and should take longer than normal. Decision errors for literals and anomalies with the preposition should be directed to "metaphor" more than without.

To enable theoretical conclusions, I think that, apart from using similes, literals and metaphors ought to be contrasted with anomalies. Anomalies set apart the first stage (literal interpretation) from the second (figurative interpretation). A serial two-stage model predicts that literal expressions yield fast processing times. If the expression is not literal, it is an anomaly (intermediate processing times). If the expression is literal nor anomalous, it is a metaphor (slow processing times). The introduction of a third expression type induces a major problem, namely the choice of choice task. If the three expression types are tested in combination, a 3-choice task (literal-metaphor-anomaly) causes difficulties that are not found in a 2-choice task, in that subjects may decide to ignore one option and concentrate on the other two. A 2-choice task for three decisions requires three experimental runs: L-M, L-A, M-A. In that case, metaphors may be faster or slower simply as an effect of the combination with literals *or* anomalies, whereas they would not if they were accompanied by literals *and* anomalies, as in the 3-choice task. Moreover, three 2-choice tasks repeat stimuli, whereas the 3-choice task does not. If repetition affects processing time, the 2-choice tasks may fail to show evidence for the model, whereas the 3-choice task does. Either kind of choice task is arbitrary. The results of an experiment may be fully limited to the specific design. Choice task should be treated as a factor in the analysis. Metaphor research has devoted a great deal of attention to context effects, which requires a baseline condition of no-context. However, little attention has been paid to the effects of condition order (presenting *context* before *no-context* or vice versa). The same applies to the order of choice task (3-choice before 2-choice and vice versa). When repeated presentation infringes on metaphor processing (1-

trial learning), any model should be discharged from an overall claim on the time relations, or it may simply be that RTs are not the best means to investigate them.

In this article, I tried to evaluate reaction-time studies into metaphor processing on the use of research design and statistics. I am afraid that most of the results are not too reliable, owing to design artifacts and careless analysis. In the discussion, I also argued that similes and anomalies (together with literals and metaphors) rather than idioms should be exploited to examine the metaphor models. Ultimately, we should not be over-optimistic about what is known about the time aspects of metaphor processing. Much has yet to be learned. Exactly because metaphor processing seems such a delicate matter, precise evaluation of the designs and statistics in metaphor research is not just a drill in methodological hair-splitting. The metaphor models structure the mental chronometry of figurative information processing, which probably is a highly sensitive operation. It could easily be wiped out by effects of repetition, order of presentation or choice task. From these observations, it follows that many of the RT-experiments on metaphor are strongly circumscribed by the experimental set up.

### Works Cited

- Clark, H.H. "The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research." *Journal of Verbal Learning and Verbal Behavior* 12 (1973): 335-59.
- Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press, 1977.
- Estill, R.B., and S. Kemper. "Interpreting Idioms." *Journal of Psycholinguistic Research* 11. 6 (1982): 559-68.
- Gentner, Deirdre, and Peter Wolff. "Alignment in the Processing of Metaphor." *Journal of Memory and Language* 37 (1997): 331-55.
- Gerrig, Richard J., and A.F. Healy. "Dual Processing in Metaphor Understanding: Comprehension and Appreciation." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 9 (1983): 667-75.
- Gibbs, Raymond W. Jr. "Spilling the Beans on Understanding and Memory for Idioms in Conversation." *Memory and Cognition* 8.2 (1980): 149-56.
- Gildea, P., and S. Glucksberg. "On Understanding Metaphor: The Role of Context." *Journal of Verbal Learning and Verbal Behavior* 22 (1983): 577-90.
- Glucksberg, Sam, M. Brown, and M.S. McGlone. "Conceptual Metaphors are not Automatically Accessed During Idiom Comprehension." *Memory & Cognition* 21, 5 (1993): 711-19.
- Glucksberg, Sam, P. Gildea, and H.B. Bookin. "On Understanding Nonliteral Speech: Can People Ignore Metaphors?" *Journal of Verbal Learning and Verbal Behavior* 21 (1982): 85-98.
- Hoffman, R.R., and S. Kemper. "What Could Reaction-Time Studies be Telling us about Metaphor Comprehension?" *Metaphor and Symbolic Activity* 2 (1987): 149-86.
- Hoorn, Johan F. "Electrocortical Evidence for the Anomaly Theory of Metaphor Processing: A Brief Introduction." *The Systemic and Empirical Approach to Literature and Culture as Theory and Application*. Ed. Steven Tötösy de Zepetnek and Irene Sywenky. Edmonton: Research Institute for Comparative Literature, University of Alberta and Siegen: Institute for Empirical Literature and Media Research, 1997. 67-74.
- Hsu, J.C. *Multiple Comparisons: Theory and Methods*. London: Chapman & Hall, 1996.
- Inhoff, A.W., S.D. Lima, and P.J. Carroll. "Contextual Effects on Metaphor Comprehension in Reading." *Memory & Cognition* 12 (1984): 558-67.
- Janus, R.A., and T.G. Bever. "Processing of Metaphoric Language: An Investigation of the Three-Stage Model of Metaphor Comprehension." *Journal of Psycholinguistic Research* 14.5 (1985): 473-87.
- Johnson, A.T. "Comprehension of Metaphors and Similes: A Reaction Time Study." *Metaphor and Symbolic Activity* 11.2 (1996): 145-59.
- Neter, J., W. Wasserman, and M.H. Kutner. *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*. Burr Ridge: Irwin, 1990.
- Ortony, Anthony, D.L. Schallert, R.E. Reynolds, and S.J. Antos. "Interpreting Metaphors and Idioms: Some Effects of Context on Comprehension." *Journal of Verbal Learning and Verbal Behaviour* 17 (1978): 465-77.
- Pavio, A., and J.M. Clark. "The Role of Topic and Vehicle Imagery in Metaphor Comprehension." *Communication and Cognition* 19.3-4 (1986): 367-88.

Author's profile: Johan F. Hoorn works in general and comparative literature at Vrije Universiteit in Amsterdam. He is author of *Metaphor and the Brain: Behavioral and Psychophysiological Research into Literary Metaphor Processing* (Ph.D. Dissertation, Vrije University, 1997) and articles in the psychology of literature, computer analysis of literature, and interdisciplinary studies. He has published in, among others, in Kreuz and MacNealy's *Empirical Approaches to Literature and Aesthetics* (1996), in Tötösy de Zepetnek and Sywenky's *The Systemic and Empirical Approach to Literature and Culture as Theory and Application* (1997), and *Literary and Linguistic Computing* (1999). He published previously "How is a Genre Created? Five Combinatory Hypotheses" in *CLCWeb: Comparative Literature and Culture* 2.2 (2000): <<http://docs.lib.purdue.edu/clcweb/vol2/iss2/3/>>. He is now completing, with Elly A. Konijn, "Perceiving and Experiencing Fictional Characters: I. Theoretical Backgrounds" and "Perceiving and Experiencing Fictional Characters: II. Building a Model": Two studies about how the appreciation of fictional characters in books, theater, film, and TV is mediated through involvement and distance processes in an interdisciplinary context of literature, aesthetics, media studies, emotion psychology, social psychology, memory psychology, and mathematics. E-mail: <[jf.hoorn@let.vu.nl](mailto:jf.hoorn@let.vu.nl)>.