The Humour of Exceptional Cases:

Jokes as Compressed Thought Experiments

Tony Veale

1. Introduction

As much as one might seek certainty and simplicity in life, the category boundaries that shape our perception and guide our behaviour are neither fixed nor certain. Rather, these boundaries are frequently the subject of examination, renegotiation and sometimes, even outright rejection, by creative individuals ranging from philosophers to artists, and jokers to scientists. In this paper we consider two of the cognitive activities that can influence these boundaries. Both are, we argue, remarkably similar in terms of the conceptual manipulations and strategies that they employ, yet both are used in very different domains, one primarily for scientific discourse, the other for social intercourse.

The first of these cognitive activities is a powerful conceptual tool for probing the underbelly of received scientific wisdom. The thought experiment, or Gedanken experiment (the latter term is often attributed to Ernst Mach; see Mach 1960,1976; Kuhn, 1964), is armchair science at its most cerebral, presenting a purely conceptual means of probing the limits of a theory not with any physical apparatus, but wholly in the mental laboratory of the imagination. This cerebral quality notwithstanding, physical intuition about the world still plays a key role in most thought experiments. As Mach notes, the goal of a good thought experiment is to construct a conceptual scenario that dredges up, from the realm of the intuitive and the instinctive, previously unarticulated knowledge

that can be manipulated at the level of concepts and categories. A thought experiment is a form of embodied reasoning that brings not just concepts, but instincts, intuitions and emotions to bear on a problem, motivating a sceptic to want to accept the conclusions of the experiment's logical argument.

The second of these activities is humour. Thought experiments and jokes both take aim at the limitations of received wisdom, often employing the same high-level strategies to provoke an audience – perhaps even a hostile audience – into accepting an alternative conceptual perspective. In each case the inconsistencies of habitual thinking are exposed, frequently with a hint of derision, satire or superiority. And in each case, imagination is vital, for jokes ask us to imagine scenarios that are so out of the ordinary that conventional modes or rules of behaviour appear to break down, in much the same way that thought experiments ask us to imagine scenarios for which conventional scientific theories fail to offer a consistent explanation. We shall argue then that many jokes are, in fact, humorous thought experiments, in which the theories under revision are social norms, genre conventions and taboos. Even off-hand witticisms and one-liners can possess the argumentative force of a good thought experiment, and we shall consider here examples that demonstrate a remarkable density of implicit argumentation. Conversely, we shall see that many thought experiments are philosophical jokes, in which the subversive logic of humour is used to induce a contradiction in an opponent's theory.

1.1. Category-Juxtaposition and Category-Subversion

It is perhaps not surprising that thought experiments and jokes should appear so similar when viewed from the appropriate level of abstraction. Both employ a tightly structured narrative to guide an audience to a particular, and often shocking, conclusion. More

generally, science and humour each thrive on insight and innovation, which in turn require a high degree of creativity. With this in mind, Koestler (1964) describes a psychological mechanism he calls 'Bisociation' to explain both, going as far as to suggest that bisociation is implicated in all varieties of creativity, from science to humour to art. Koestler's influential theory can be considered the fore-runner of Attardo and Raskin's (1991) General Theory of Verbal Humour (GTVH), Fauconnier and Turner's (1998) Conceptual Integration Theory (or blending theory), and Coulson's (2000) frame-shifting theory of humour. De Mey (2002) in turn presents an account of thought experiments in terms of conceptual blending theory.

Bisociation, blending, frame-shifting and the script-switching of the GTVH are all multi-space juxtaposition theories. Each describes a mechanism whereby multiple input categories (whether the matrices of bisociation, the mental spaces of blending and frame-shifting, or the scripts of the GTVH) can be integrated, and through which the oppositions between these inputs can be identified and resolved. Since juxtaposition can only meaningfully apply to a plurality of inputs, humour and thought experimentation are consequently viewed as combination operations: given the appropriate input categories to combine, the desired output category can be generated. Juxtapositional theories are capable of describing a good many instances of humour, and as De Mey illustrates, a variety of celebrated thought experiments also. At a trivial level, of course, humour appreciation is inherently juxtapositional, since some form of comparison will always be required to detect deviation from that which is normative or expected to that which is innovative (Giora, 2002). The key issue concerns the nature of the structures that are actually juxtaposed: do they comprise different input structures that are to be cross-

mapped or blended, or are they variations (normative and creative) of a single input structure? We reserve the label "juxtapositional theory" for theories that non-trivially presuppose the former view. However, many creative jokes, specifically of the kind to be discussed in sections 3 and 4, appear to presuppose the latter view, as do many creative thought experiments. In these cases, it seems that what is juxtaposed is not a pair of different but overlapping categories, but a single category and a creatively subverted variation of this category. For all intents and purposes, comprehension of these constructs does not involve the juxtaposition of multiple inputs, but manipulation of a single input.

1.2. Structure of this Paper

It is our goal in this paper to explore the role of subversion in thought experimentation and humour. We begin by considering the structure of thought experiments in section two, where we elaborate on Gendler's (1998, 2000) notion of an *exceptional case* and consider how one might be constructed for a given theory. In section three we explore the role of subversion in verbal humour, and demonstrate how exceptional cases can be constructed from the raw lexico-conceptual components of conventional linguistic constructions. We also describe here a particular genre of humour called 'trumping' whose form more clearly echoes the adversarial dialogue that lies at the heart of thought experiments. In section four we look to inter-personal considerations in both jokes and thought experiments to better understand why humour arises from some exceptional cases and not others. In section five we offer a case study of a particular linguistic form, the stereotypical simile, to see how (and how often) commonplace stereotypical associations are subverted by exceptional cases. Finally, we conclude in section six with a consideration of the implications of this synthesis of thought experiments and humour.

2. Thought Experimentation

Thought experiments derive their distinctive argumentative force from a grounding in physical reality. Most describe activities that are physically realizable, if given enough time or resources, and thus engage the corresponding physical intuitions and instincts. Scenarios that do not require us to delve into the level of physical intuition and exploit what Mach (1976) calls "our store of instinctive knowledge" are not truly thought 'experiments'. It is not enough then to merely ask "what if", for the thought process must engage in some simulacrum of an experimental activity. Consider Euclid's demonstration of the infinitude of prime numbers. This proof by contradiction neatly demolishes the notion of a largest prime number by first asking us to conceive of such a number, before then demonstrating how an even larger prime can be constructed. Yet this can hardly be called an experiment, since there is little here to actually visualize, and the key mental operation has no experimental analogue in the physical world. In contrast, the argument provided by Lucretius for the infinitude of space is an experiment, of sorts. Imagine yourself lobbing a spear at the boundary of known space, he says. If the spear bounces back, then the boundary is real, but there must be something on the other side for the boundary block access to; but if the spear passes through, then the boundary was not real to begin with. In either case, we can conclude that space is not bounded at this particular point, and since we can repeat the experiment everywhere, it is not bounded anywhere.

For thought experiments to preserve a definite philosophical function of their own (in the sense of Kuhn, 1964), and be seen as more than a conventional logical argument in fancy-dress (as has been argued by critics such as Norton, 1996), they must bring more

than pure logic to the table. The sense of physical embodiment that comes from imagining a real world action, whether the throwing of a spear at an imagined boundary or the dropping of stones from the top of a tower, is intrinsic to how we interpret the experiment and its outcomes. For one, this embodiment serves a psychological purpose to be sure (as Norton concedes). But more importantly. the appropriate physical descriptions can engage the image schemas that best support the sceptic's arguments (see Johnson, 1987). For instance, the image schema of boundary is a symmetric one, where we asked to imagine a wall-like barrier separating two regions of space[1]. By successfully invoking this schema in the mind of the reader, Lucretius establishes not just an experiential basis for his argument, but a strong conceptual basis also, for his very language presupposes the existence of space on the other side of the boundary.

A strong emotive basis is also important in nurturing the desired response to a thought experiment, much as it is in humour, which often seeks a visceral reaction to a joke. As such, the mocking tone of some thought experiments is intended not just to persuade, but to deride and lampoon, as though to make an opponent embarrassed to espouse the theory under attack. It seems clear, for instance, that Galileo is having fun at the expense of his opponents in his Dialogue Concerning The Two Chief World Systems when he puts Aristotle's opposing theory into the mouth of a character he derisorily calls "Simplicius" (see Gendler, 1998). Likewise, Searle (1980) creates a *mise en scene* for his Chinese Room argument against Artificial Intelligence (as typified by the work of Schank and Abelson, 1977) that is so vaudevillian that it elevates this particular thought

[1] More specifically, because Lucretius depicts the boundary as acting as a potential obstacle to the flight of his imaginary spear, the Blockage schema becomes activated (Johnson, 1987: 45). Inherent in this schema is the idea that one can "get past the blockage", and hence the idea that there is something beyond the blockage.

experiment to the level of parody. Searle asks us to imagine a man much like himself who is locked in a room and who receives a series of incomprehensible Chinese markings through a slot. By following a set of complex mapping rules in a huge rulebook, the man assembles a response that is returned through another slot. Though the response may seem germane and intelligent to the outside observer, who may well believe the room to be occupied by a native Chinese speaker, those who are privy to the experiment know that the man inside remains ignorant of Chinese and the meaning of the symbols that he processes. Were the man to be replaced by a computer, and the rule book by a program, Searle concludes that the computer would likewise be ignorant and would thus not exhibit true intelligence. In a display of verbal mastery, Searle formulates the perfect exceptional case of an abstract computer (a man using a rule book) to subvert the idea of an intelligent computer.

2.1. Exceptional Cases

Gendler (1998, 2000) claims that all thought experiments describe the construction of exceptional cases, since it is exceptional cases that best expose the limits of the conventional uses of categories. In this view, the history of thought experiments is the history of exceptional cases that beg the right questions at the right time in the development of science. For instance, the "ship of Theseus" experiment asks us to imagine a ship in which every single piece of wood has been replaced over time, raising an important question about the nature of identity. The Chinese room experiment of Searle (1980) asks us to imagine a book of rules with which a person ignorant of Chinese can meaningfully process any Chinese query. But Gendler's view begs an important question of its own, namely, what is an adequate definition of an exceptional case?

Gendler begins by assuming that not all of the features of a category are of equal importance, but that some will be of primary or central importance in defining membership, while others will be secondary and peripheral. Therefore, the primary features are either necessary for category membership (in a classical view of category structure) or prototype-defining (in a non-classical view), while those features that one tends to habitually associate with primary features, are considered secondary. This dichotomy of features is consistent with both Lakoff's (1987) notion of a *radial category* (which has a prototype member at its centre, and less representative members on its periphery) and Sowa's (2000) *egg-yolk theory* of meaning (where primary features occupy the yolk, and secondary features occupy the egg-white).

Consider Aristotle's theory of falling objects, which claims that all objects fall at a speed proportional to their weight. Paradoxically, Aristotle provides no consistent explanation of how entities are to be individuated into discrete objects, yet offers a theory of falling speed that crucially depends on the nature of this individuation. Galileo's classic thought experiment (perhaps the most celebrated example of the genre), exposes this theory as contradictory by imagining an exceptional case where the issue of individuation is foregrounded, by asking us to imagine a composite object comprising two stones, one large and one small, flexibly connected by a rope. This exceptional object is simultaneously a single entity (a system of connected stones) and two individual entities falling in concert. This simultaneity reveals a fundamental confusion in the Aristotelian world view, since the composite object should fall both faster than the heavier stone alone (because the composite is heavier still), *and* slower (because the lighter stone would act as a drag on the heavy stone). Implicit in Aristotle's theory is the

expectation that the objects of interest are either atomic (non-composite) or rigid. However, Galileo recognizes that atomicity and rigidity are not primary but secondary features of the category that can safely be contradicted to form an exceptional case.

Galileo demonstrates that to create an exceptional case, one must first have an appreciation of what constitutes the *un*exceptional or stereotypical examples of a category, so one can perceive where the category is most vulnerable to criticism. One can then choose a highly conventional example of the category to subvert (e.g., a physical *and* non-composite object), separating those components that are primary and central to membership in the category (e.g., physical) from those that are secondary and merely habitual or accidental (e.g., non-composite). One can then reassemble these components to arrive at an example that is, technically at least, a member of the category while contradicting certain of the habitual expectations that have been stripped away (e.g., an object that is physical *but* composite).

2.2. Exceptionality and Consistency

As this example demonstrates, an exceptional case will successfully subvert a category only when it forms a valid yet uncomfortable fit with this category. This is, it must possess enough primary characteristics to be recognized as a valid category member, if only technically so, yet once admitted, it must prove itself to be an inconsistent member of this category. But consistency is often relative in thought experiments, especially those that rely on instinctive knowledge. In a purely logical argument, contradiction is defined in absolute terms via negation, whereas an embodied argument must pit one qualified belief against another. An inconsistency will arise then whenever a theoretical belief is shown to imply a conclusion that contradicts a more deeply entrenched belief or physical

instinct. A subverted theory thus faces a serious dilemma when presented with a suitably exceptional case: either the case should be excluded from the theory, in an admission that the theory is necessarily incomplete, or admitted into the theory, where its presence forces the theorist to reject an even more fundamental and treasured belief. For example, Galileo forces Aristotle (via Simplicius) to either reject the idea that weight determines falling speed, or to accept the bizarre outcome that two different speeds can be simultaneously ascribed to the same falling entity. Ultimately then, consistency is judged against a set of baseline beliefs that one is more loathe to reject than the theory itself.

These baseline beliefs can be a matter of simple common sense (e.g., that an object has a single speed at any given moment), or a scientific belief that one takes as a near-absolute. For instance, Mach (1960) recounts a thought experiment in which a chain is draped over a frictionless triangular wedge, and demonstrates that the chain must be in a state of rest by asking us to further imagine joining the loose ends of the chain so that they form a loop that hangs around the wedge. If the chain is not to reach a state of equilibrium, its circular shape ensures that it must slip endlessly around the wedge, and in doing so, form a perpetual motion machine. Mach suggests that we instinctively find such an outcome impossible, though it is perhaps truer to claim that the conservation of energy is a principle that physicists are instinctively driven to defend.

3. Verbal Humour

The categories of most interest in verbal humour, in particular the narrative humour of "story" jokes, are those that pertain to event structure, social convention and genre expectations. This realization has lead computationally minded theorists of humour to

view the *script* (a notion given computational form by Schank and Abelson, 1977) as the most appropriate level of categorization for resolving the meaning of narrative jokes (e.g., see Raskin, 1985). Indeed, one cannot discuss theories of humour without granting centre-stage to the script-based General Theory of Verbal Humour (or GTVH) of Attardo and Raskin (see Attardo and Raskin, 1991; Attardo *et al.* 2002). The GTVH is a juxtapositional theory of humour that is an modular reworking of Raskin's (1985) Semantic Script Theory of Humour (or SSTH). Like the SSTH, the GTVH views a joke as a narrative that is compatible with multiple scripts, one of which will at first appear primary until the punch-line contrives a *incongruity* that must be resolved (see Suls, 1972; Ritchie, 1999; Veale, 2004). Resolution is achieved, either partially or fully, by a special logical mechanism that analyses the nature of the mismatch between the primary script and the text, before switching the thrust of interpretation from this script to another.

GTVH scripts can be activated by a text in one of three ways: lexically (by association with a single word, called the *lexical handle* of the script); sententially (by a pattern of words and lexical scripts); and inferentially, as a by-product of common-sense reasoning (e.g., as when one intuits that a joke is racist and activates a Racism script). Furthermore, since certain elements in a script will be more salient and foregrounded than others, these elements are marked to distinguish them from less salient background elements. More recently, Attardo *et al.* (ibid) augment this view with a graph-theoretic account of script representation that views scripts as arbitrarily complex symbolic structures, to which juxtapositional processes like sub-graph isomorphism can be applied. This representational shift allows the GTVH to encompass even punning as a script-level operation, provided the notion of script is sufficiently generalized to accommodate

phonetic as well as semantic information. With this generalization, the GTVH moves further from Schank and Abelson's vision of a script, toward a generalized data structure that perhaps buys its increased descriptive flexibility at the cost of explanatory power.

The GTVH views the process of incongruity-resolution as the work of a particular logical mechanism (LM) that operates across script representations. Understandably, LMs have proved to be the most enigmatic elements of the GTVH, prompting Attardo *et al.* (2002) to enumerate a taxonomy of 27 different LMs. For instance, it is suggested that an LM called *false-analogy* is central to jokes whose humour derives from ill-judged comparisons, as in the old joke where a mad scientist builds a rocket to the sun but plans to embark at night to avoid being cremated. Here a false analogy is created between the sun and a light-bulb, suggesting that when the sun is not shining it is not "turned on", and hence, not hot. Different LMs may be employed in different jokes, bringing a distinctive logical flavour to each. Indeed, insofar as jokes that employ the same LM may possess the same identifiable character, LMs resemble the *Ur-jokes* of Hofstadter and Gabora (1989). These are joke skeletons that can be re-instantiated in different settings with different characters while preserving a distinctive character that runs through each of their manifestations (in this respect, Ur-jokes are productive humour schemas that in turn resemble the metaphor schemas of Lakoff and Johnson, 1980 and Johnson, 1987). The individuation of different LMs make the GTVH a highly modular theory of humour in which research can proceed on many different fronts simultaneously. Nonetheless, such extreme modularity, when combined with the GTVH's juxtapositional view of humour, tends to reduce jokes to the level of particular dishes as defined by standard recipes. In this view, the GTVH begins to resemble a kitchen appliance, in which logical

mechanisms are little more than the optional whisks and cutting blades that can be attached in different contexts to meet different production needs.

3.2 Subversion of Verbal Meaning

As noted in the context of thought experiments, one constructs an exceptional case by stripping away the layers of conventionality and habitual thinking that have accreted around a category. Fauconnier and Turner's (1998, 2002) theory of blended concepts explains why integrative ideas can have so many accreted meanings that do not directly derive from their individual parts. In this view, the integrated concept occupies its own mental space, a special blend space, in which recruitment of additional concepts and a process of gradual elaboration can occur. To undo the after-effects of blending, one must dismantle the chosen concept into its fundamental parts, so that it may be reconstructed devoid of these layers of recruited and elaborated meaning. Consider the following witticism from serial divorcee Zsa Zsa Gabor, which shows that in verbal humour, these fundamental parts are often directly accessible as individual words and morphemes.

> "Darlink, actually I am an excellent housekeeper. Whenever I leave a man,
> I keep the house!"

While the GTVH entreats us to view jokes like this as a juxtaposition of scripts, this merely begs the question of where these scripts originate. For while can expect to find a conventional housekeeping script in the lexicon, indexed by its lexical handle "housekeeper", it is unlikely that any *a priori* structure expressing the meaning "a taker and keeper of houses" can be found so readily. As conventionally defined, scripts "are not subject to much change, nor do they provide the apparatus for handling totally novel

situations" (Schank and Abelson, 1977). It follows that this new meaning must arise not from a script, but as a creative product in itself, via an exceptional reading of the phrase "housekeeper" that describes an exceptional member of the category of housekeepers. The subsequent inclusion of this exceptional reading in the housekeeper category, whose prototypical members are thrifty and hard-working rather than spendthrift and pampered, is a creative act of category subversion that in turn undermines the tacit value system to which Gabor is responding, In doing so, she pointedly (via "actually") undermines the suggestion that conventional housekeeping skills are the valid measure of a woman.

Derived categories are those that depend on a logically prior category for their conventional meaning. For instance, the category Hypotenuse depends on the category of triangles, for without right-angled triangles there would be no hypotenuses, or, in the language of Cognitive Grammar (Langacker, 1991), Hypotenuse is a profiled element of the base concept Right-Angle-Triangle. However, this dependency can be subverted by witticisms such as "Hypotenuse seeks two straight lines to form love triangle". Likewise, the category Meat is conventionally conceived as logically dependent on the category Animal, since instances of Meat are derived from instances of Animal. Conventional wisdom thus holds that without animals there can be no meat, and without meat there could be no vegetarians. However, this is a form of habitual thinking that can be wittily subverted, as in the following one-liner:

"If God wanted us to be vegetarians he wouldn't have made animals out of meat".

Which categories are subverted here, Vegetarian, Animal or Meat? The answer appears to be all three, for we seem to be presented with three quite exceptional objects that simultaneously subvert three different categories. First we are directed to imagine an

exceptional member of the Animal category, the animal as meat machine, from which all non-utilitarian aspects are divorced; if such an animal were not sentient, there could be no moral basis for vegetarianism. Secondly, we are asked to imagine an exceptional kind of meat, one that possesses all the biological properties of conventional meat yet one that may not derive from an animal source. Thirdly, we are directed to imagine an exceptional kind of Vegetarian, one that would eat meat if it did not derive from an animal source. All three subversions together lead to a subversion of the category Vegetarianism, for what moral force would this lifestyle preserve if vegetarians could freely eat meat yet remain a vegetarian? The above joke is, in fact, a highly compressed thought experiment that attempts to undermine the conventional theory that vegetarianism is a morally superior way of life, while justifying a moral laissez faire on the part of the meat-eaters.

Some of the most effective uses of subversion aim for a more visceral effect:

"Eating is over-rated. Remember, food is just excrement waiting to happen."

This witticism succeeds in constructing a quite exceptional member of the category Food, that of excrement-in-waiting, for prototypical members of the category Food are expected to be edible and tasty, while excrement is neither. We can thus view this joke as another compressed thought experiment, one that uses a time-shifted view of food to subvert the conventional wisdom that to eat well to live well. It achieves this subversion through a visceral form of metonymic *tightening* (e.g., see Fauconnier and Turner, 1998; Veale and O'Donoghue, 2000), a compression of relations that strengthens the connection between Food and Excrement to uncomfortably suggest that when one is eating the former, one is simultaneously eating the latter (see also Fauconnier and Turner, 2002).

3.3 Subversion of Conceptual Mappings

One might well argue that while thought experiments involve a deep form of conceptual subversion, wherein a conceptual construct like a scientific theory in undermined, jokes merely subvert the semantic (or script-based) expectations of an audience. But this kind of expectation-subversion often goes hand-in-hand with a subtle subversion of conceptual viewpoints, as illustrated by the following exchange between two vagabonds:

Tramp #1: ‘ave you seen yesterday’s newspaper?

Tramp #2: Can’t says that I ‘ave. What’s in it, anyhow?

Tramp #1: My lunch, that’s what!

Listener expectations are here based on a number of common metonymies: “to see” a newspaper is usually taken to mean “to read” a newspaper, while the “contents” of a newspaper usually refers to news stories rather than physical objects. But these metonymies do not, in themselves, provide the humour of the exchange; non-sequitors and nonsense behaviour will also thwart listener expectations, but to produce bafflement rather than humour. The subversion we find here is also conceptual, and works at several levels simultaneously. First we see the subversion of the concept Newspaper, which is demoted from its usual standing as a container of knowledge (an organ of the truth) to a lowly container of food: “today’s news is just tomorrow’s wrapping for fish and chips”. Secondly, we see the subversion of the speaker himself, who is transformed from a consumer of “high” knowledge to a consumer of “lowly” food. And thirdly, we see a subversion of the concept News, or Knowledge in general, where “food for thought” is seen as less important to human existence than food itself.

3.3 Trumping: Subversions of Figures of Speech

The subversion of verbal meaning allows a witty speaker to overtly agree with a critic while simultaneously subverting the critic's argument. Veale *et al.* (2006) describe this particular combination of category subversion and overt agreement as conversational 'trumping', and note that trumping is not a form of deliberate *mis*-understanding, but is actually a form of *hyper*-understanding, in which the respondent exhibits a greater understanding of the verbal meaning than does the critic. Trumping heightens the social dimension of category subversion by requiring that the parties to the dialogue do not overtly disagree, but this in turn heightens the creative demand placed on the respondent. The result is not only an ideational subversion of the initiator's pragmatic goals (via a given category), but a highly effective (and thus humour inducing) interpersonal subversion of the initiator as a social agent.

Trumping is a form of subversive humour that thrives on idioms and conventional figures of speech, since close analysis often reveals these figures to be built upon on a foundation of active conceptual metaphors. From this perspective, even the most frozen of idioms can be appreciated as a manifestation of metaphoric reasoning (e.g., see Gibbs, 1993). So where one finds metaphors, one also finds theory-like assumptions that can be subverted for humorous ends. The joke below nicely demonstrates that figures of speech, like theories, make claims that can be subverted using a potent exceptional case:

CEO: *(indignantly)* I do the work of two people for this company!

Chairman: Yes, Laurel and Hardy.

The idiomatic expression "to do the work of two people" makes the theory-like claim that two typical workers can achieve more than one alone (or, proverbially, that "many hands

make light work"). They key assumption here is typicality: two unexceptional workers might do more work than one alone, but one unexceptionally competent worker is surely preferable to two exceptionally incompetent ones. The goal of trumping is not simply to invalid the initiator's use of a particular category, but to validate this usage while simultaneously undermining the tacit assumptions that determine the effects of the usage. Since the pairing of "Laurel and Hardy" serves as a recognizable prototype of the bumbling duo, it also serves as the ideal *exceptional case* for the assumption of typicality that governs the use of this idiom, to the detriment of the initiator.

4. Interpersonal Dimensions of Conceptual Subversion

It is not the act of subversion in itself that gives rise to a humorous effect, but the pragmatic, largely social, uses to which the subversion is put. Conventional scientific thought experiments can employ conceptual subversion without being overtly humorous, but may be humorous to the extent that they can be understood in social terms (for example, as attacks on the originating theorist). That is, to the extent that a given subversion has a strong social and interpersonal dimension, we should expect the effect to be perceived as more humorous, *ceteris paribus*[2], whether the context is an explicit joke or a scientific thought experiment. Consider the following exchange between the boxer Muhammad Ali (at the height of his sporting and verbal prowess) and a flight attendant:

Flight attendant: Buckle your seat belt, Mr. Ali, we're about to take off.

Muhammad Ali: Superman don't need no seat belt!

Flight attendant: Superman don't need no airplane neither.

[2] Other factors, like aptness, topicality, conciseness and intellectual depth, are also extremely important.

Though an off-the-cuff remark rather than a considered scientific claim, Ali's description of himself as "Superman" constitutes a world-view that one can either defend or attack., much as one might react to a novel scientific claim. But what makes Ali's claim particularly ripe for humorous subversion is that the claim concerns *himself*: few scientists advance scientific claims about themselves, so most thought experiments only tangentially represent an assault on the proponent of the claims themselves. In undermining Ali's use of the concept Superman, the attendant simultaneously undermines Ali himself in a way that demonstrates her authority over a disruptive passenger and her verbal mastery over a smart-mouthed aggressor. Ali's particular world-view of himself as a "Superman" is thereby shown to be inconsistent, much as Aristotle's world-view was posthumously shown to be untenable by Galileo.

The distinction between funny and unfunny thought experiments is not at all a binary one. As we have seen, scientific thought experiments are not the abstract expression of a socially disinterested thought process, but the forceful expression of a criticism that can be rich in satire and humorous intent. Galileo, for instance, was less interested in attacking Aristotle than he was his contemporaries, who he labels simpletons through his none-too-subtle naming of the interlocutor "Simplicius". Likewise, Searle's Chinese room argument has many humorous qualities, not least the extreme caricature he paints of A.I. research. He lampoons the rules of conventional A.I. programs by stating that his imaginary Chinese rulebook contains rules of the form "*squiggle squiggle* means *squoggle squoggle*" (Searle, 1980). His goal is to show that A.I. is not just inconsistent with our conception of human intelligence, but quite ridiculous in its own right. The baroque elements of Searle's argument place his experiment into the realm of a joke but

his exposition never strays so far that the butt of the joke, symbolic A.I., becomes inaccessible or obscured. As such, Searle strives to construct an argument that is *optimally innovative* in the sense of Giora (2002).

In these thought experiments we see an attempt to subvert not just a theory but the proponents of the theory themselves. For instance, Searle (ibid) claims that only those "in the grip of an ideology" would attempt to critique his Chinese room argument via the "system's reply" (which states that it is the combination of *man* + *rulebook*, rather than the man alone, that truly understands Chinese). This suggests that thought experiments, like jokes and metaphors, can have both an ideational and an interpersonal function (e.g., see Halliday, 1985). Thought experiments with a pronounced interpersonal dimension are more likely to engage the social instincts of a listener and achieve both an emotive and an intellectual effect. So to the extent that our social instincts lead us to enjoy the humbling of the mighty (e.g., Aristotle), the pompous (e.g., the religious supporters of Aristotle), the arrogant (e.g., Roger Schank, as perceived by John Searle) or the stupid (e.g., perhaps all of the above), we may tend to find humour in such thought experiments (see Freud, 1938). The precise degree of humour will depend, of course, on other factors, such as the quality of execution (e.g., cleverness, conciseness) and the social connection of the listener to the category and its proponents.

5. Subverting Cultural Stereotypes: A Case Study

Language is the primary means through which cultural knowledge is preserved and transmitted from one generation to the next, yet it does so in a manner that is not always maximally informative. Indeed, idioms and other stock phrases represent this cultural

knowledge in a fossilized form that can often appear inscrutable to modern speakers. For instance, Charles Dickens opens "A Christmas Carol" by stating that "Old Marley was as dead as a door-nail" before going on to wonder what it is about door-nails that makes them so suited to the evocation of death. Suggesting that "coffin nails" might make a more suitable metonymy (noting "I might have been inclined, myself, to regard a coffin-nail as the deadest piece of ironmongery in the trade"), Dickens admits defeat by concluding "the wisdom of our ancestors is in the simile, and my unhallowed hands shall not disturb it, or the Country's done for" (Dickens, 1843/1984, page 1). In raising the inscrutability of commonplace similes, Dickens is, in fact, having fun at the expense of received wisdom as it is encoded in language. By subverting listener expectations in the way such stock phrases are used creatively, a witty speaker can also subvert the received wisdom that underpins these phrases, to point out the limitations of this wisdom.

We take our cue from Dickens in this section, and demonstrate how commonplace similes can be subverted for humorous ends. We first describe how a large corpus of such similes is automatically harvested from the texts of the World Wide Web (in section 5.1). before showing how the simile pattern can be used to ironically subvert listener expectations in section 5.2, and less frequently, to subvert the underlying conceptual stereotypes themselves in section 5.3.

5.1 Acquiring a Corpus of Commonplace Similes

For our current purposes we consider stock similes of the form *"as ADJ as a|an NOUN"*, and we attempt to collect all commonly-used values of ADJ for a given value of NOUN. To do this, we first extract a list of antonymous adjectives, such as "hot" or "cold", from the lexical database WordNet (see Fellbaum, 1998); the intuition here is that explicit

similes will tend to exploit properties that occupy an exemplary point on a scale. For every adjective ADJ on this list, we then send the query "*as ADJ as **" to Google and scan the first 200 snippets returned for different noun values for the wildcard *. From each set of snippets we can ascertain the relative frequencies of different noun values for ADJ. The complete set of nouns extracted in this way is then used to drive a second phase of the search. In this phase, the query "*as * as a NOUN*" is used to collect similes that may have lain beyond the 200-snippet horizon of the original search, or that hinge on adjectives not included on the original list. Together, both phases collect a wide-ranging series of core samples (of 200 hits each) from across the web, yielding an initial set of 74,704 potential simile instances (of 42,618 unique types) relating 3769 different adjectives to 9286 different nouns.

However, many of these instances are not sufficiently well-formed for our purposes. In some cases, the instance does not convey a stereotypical association, but a highly contingent one that only holds in a specific, ephemeral context. In other cases, the noun value forms part of a larger noun phrase: it may be the modifier of a compound noun (as in "bread lover"), or the head of complex noun phrase (such as "gang of thieves"). A human judge is thus used to annotate those instances that correspond to bona-fide similes, by which we mean, similes that associate an adjectival property with a noun concept for which that property stereotypically holds (such as "hot" for "oven", "humid" for "sauna", or "busy" for "beaver"). Overall, 30,991 of these simile instances are accepted as bona-fide expressions of a stereotypical association, yielding 12,259 unique adjective-to-noun associations, from 2635 adjectives to 4061 different nouns. As such, this collection represents the largest resource of its kind for the study of similes.

5.2 Subverting Listener Expectations

When one uses the syntactic pattern "as ADJ as a …" there is a clear listener-expectation that what follows is a noun description that is highly evocative of the property ADJ. But among those instances of the simile pattern that our judge rejects, we find 4685 instances (or 2798 unique associations of an adjective to a noun) that can be classified as *ironic*. An ironic simile is here taken to be any comparison of the form "as ADJ as a NOUN" for which NOUN is not merely un-stereotypical for ADJ, but for which NOUN actively evokes the opposite property *not*-ADJ. Examples include "as subtle as a freight-train", "as bullet-proof as a sponge-cake" and "as private as a shopping-mall". Surprisingly then, the number of ironic similes constitutes a larger proportion of well-formed similes than one might have previously imagined, with bona-fide similes (those that express a stereotypical association) out-numbering ironic similes just 3 to 1. Of course, the bona-fide simile remains the norm, as one would expect, but these findings nonetheless point to a large-scale subversion of the simile frame to achieve humorous effects in web texts.

Similes like "as hairy as a bowling-ball" do an effective and humorous job of subverting listener expectations by promising one thing (e.g., a stereotype of hairiness) and delivering another (e.g., a stereotype of baldness). In doing so, they doubly accentuate the lack of an expected property and re-create in the mind of the listener the surprise initially experienced by the speaker. In other words, speakers use such similes when they themselves expect to find the property ADJ but are surprised to instead perceive *not*-ADJ. As such, these similes are clearly amenable to analysis by the GTVH, wherein the syntagmatic pattern "as ADJ as …" triggers a script that expects a stereotype of ADJ, but a stereotype of *not*-ADJ is discovered instead. The conflict between ADJ and

the implicit evocation of *not*-ADJ yields the semantic incongruity demanded by the GTVH, while this incongruity is resolved by recognizing that "as ADJ as …" can have two meanings: ADJ can denote an extreme point on the scale of ADJ-ness (e.g., "hairy" may denote a state of extreme hairiness) or the scale of ADJ-ness itself (e.g., "hairy" may denote the extent to which something is more or less hairy).

5.3 Subverting Arguments based on Stereotypes

Simple ironic similes such as these are amenable to the GTVH because they merely succeed in subverting expectations, and expectations are eminently conducive to formalization as scripts. Because these similes do not exhibit any traction at the conceptual level, they cannot succeed in changing our views of the concepts concerned. For instance, the ironic simile "as ruthless as a bunny-rabbit" does not make us think of rabbits as any more callous, nor does it make our stereotypes of ruthlessness, such as sharks and wolves, seem any less cold-blooded. In short, while such similes can be novel, they are not *optimally innovative* in the sense of Giora (2002), since they fail to identify a specific stereotype to which any innovation can apply.

To be optimally innovative, an ironic simile should both evoke a stereotype and simultaneously subvert it. For instance, while someone who is "as fast as a cheetah" is very fast indeed (since speed and agility are properties of the stereotypical cheetah), someone who is "as fast as a *three-legged* cheetah" is, conversely, remarkably slow. Examples like the latter do not simply subvert listener expectations, but subvert the very logic of stereotype-based reasoning itself. Just as ironic similes are less frequent than their bona-fide counterparts, we expect optimally-innovative ironic similes to be less frequent than ironies that simply challenge listener expectations.

We test this hypothesis by harvesting similes from the web that have the syntactic form "as ADJ_1 as an ADJ_2 NOUN", since – as illustrated by the cheetah example above – this form can both evoke a salient stereotypical association ("as ADJ_1 as a NOUN") and innovatively subvert this association (insofar as "ADJ_2 NOUN" denotes a sub-kind of entities that strongly suggest the opposite property, *not*- ADJ_1). Taking the 12,259 stereotypical adjective-to-noun associations harvested in section 5.1, we use these to construct queries of the form "as ADJ as a|an * NOUN" that retrieve elaborations of these basic similes from the web. We find 5729 elaborations in total, such as "as invulnerable as an *armoured* tank", "as ugly as a *shaved* mule" and "as delicious as a *fresh* peach".

Unsurprisingly, most elaborations serve to augment and reinforce the stereotypical association on which they are based. Thus, we find "as bright as an *exploding* star", "as supple as a *young* willow" and "as blue as a *cloudless* sky", where, in each case, the elaboration prompts the listener to construct a more detailed mental picture of the stereotype underlying the comparison. In some cases these reinforcing elaborations add humour to an otherwise unfunny association, such as "as white as a *frightened* ghost", "as green as a *pickled* toad" and "as dry as a *Syrian* martini". Interestingly, these examples do not subvert listener expectations (a pickled toad is still green) but do subvert the stereotypes underlying these expectations (a frightened ghost is not a stereotypical ghost, a pickled toad is a quite exceptional toad, and a Syrian martini is an unlikely concoction at best). However, we have as yet no formal basis for deciding whether a non-ironic elaboration presents an exceptional case or not, as this depends crucially on the perceived absurdity of the mental image that is constructed (is a "shaved mule" more or less exceptional than a "pickled toad" or an "exploding star"?). In each case, these reinforcing

elaborations have the following argumentative form:

Speaker 1: If you think a NOUN is ADJ_1, an ADJ_2 NOUN is even more ADJ_1

Speaker 1: X is not just as ADJ_1 as a NOUN, but as ADJ_1 as an ADJ_2 NOUN!

When humorous, these examples just stop short of subverting the associated stereotype, but do show how the stereotypical category can be elaborated to include increasingly bizarre category members. The more bizarre (yet valid) these members, the funnier the elaboration is likely to appear. One might consider this a mild form of subversion, of a kind that does not undermine the stereotype but which nonetheless exploits the stereotype to achieve ridiculous ends. As such, these examples can be seen as a weak form of thought experiment, one that does not strive for contradiction or inconsistency, but one that cleverly explores the limits of what is possible and what is meaningful.

We also find a number of ironic elaborations that more obviously correspond to the stronger formulation of a thought experiment discussed earlier: these subvert both the listener's expectations (by offering a comparison for which *not*- ADJ_1 is most salient than ADJ_1) and the explicitly provided stereotype: we find, for instance, "as white as a *bloody* sheet", "as edgy as a *dulled* razor", "as explosive as a *wet* firecracker" and "as accurate as a *drunken* archer". Each can be seen as a compressed dialectical argument, or trumping:

Speaker 1: X is as ADJ_1 as a NOUN

Speaker 2: Yes, an ADJ_2 NOUN!

Speaker 1's claim can here be seen as a culturally received theory about NOUN (as well as a personal theory about X), to which Speaker 2's trumping elaboration can be seen as

an exceptional case for which this theory *should* hold, but does not.

Our study reveals that just 2% of the elaborations we harvest from the web (or 109 cases among 5079) have this doubly subversive form. While we make the strong prediction that these forms, which subvert both listener expectations and stereotype-level expectations, should be perceived as funnier than corresponding forms that subvert neither, or that subvert just one (expectations *or* stereotypes), this prediction still remains to be empirically validated.

6. Conclusions

By examining the similarities between thought experiments and jokes, we see that both forms of discourse are similar by virtue of their subversive role in the undermining of habitually-held world-views. As such, we argue that rather than view humour as a juxtapositional mechanism that combines different scripts, frames or mental spaces, it is often more fruitful to view humour as a rather pointed use of a more fundamental cognitive mechanism: the ability to probe the boundaries of existing categories, to illuminate the unspoken limitations of these categories, and to offer exceptional cases that expose these limitations to ridicule. In this view, jokes and thought-experiments can be seen as language games that posit exceptional cases to subvert or undermine patterns of conventional thinking and the people that exploit them (e.g., see Veale, 2002). The construction of exceptional cases, as advocated by Gendler (1998, 2000), relies on an ability to dismantle concepts into their component parts, to reveal what is truly central to the workings of a category. In this way, jokes and thought experiments provoke a sceptic to undertake a radical re-analysis of a category, one that reveals how a category can be

used and, perhaps more importantly, misused.

Jokes and thought experiments can both be used to prompt a reappraisal of a particular mindset or mode or behaviour, but it is important to note that jokes are granted a special licence in this regard, one that is off-limits to the scientific thought experiment. Crucially, jokes are free to exploit hyperbolae, irony, sarcasm and metaphor, while the effectiveness of a thought experiment is predicated upon the perceived fairness, factuality and descriptive plainness of the arguments used. Jokes are not expected to be either fair or factual, and their effectiveness is measured in terms of their ability to provoke laughter. People rarely, if ever, laugh at semantics: humour is a social phenomenon, and semantics can provoke laughter only when it is given a pragmatic social dimension, explicitly or otherwise. We laugh not just because categories are subverted and their boundaries shown to be fuzzier than previously believed, but because there are social consequences of this subversion that we find psychologically satisfying (Freud, 1938).

Crucially then, we do not claim that all jokes can be interpreted as thought experiments in the strongest sense, for shoehorning all of verbal humour into a single philosophical paradigm would inflict a serious injustice on both phenomena. For instance, there exists a substantial body of jokes – such as those that nurture stereotypes of women and ethnic minorities – that attempt to promulgate conventional thinking and buttress existing belief structures. While such jokes can be subversive to the point of being socially corrosive, they do not fit comfortably into the Gedanken mould. However, they do have a corresponding form in scientific discourse, since most scientific experiments are designed to buttress rather than undermine a particular theory. Jokes then, like experiments, can be used to bolster or to undermine, and just as thought

experiments cause us to question the unspoken assumptions that surround a theory, jokes can prompt us to question the habitual associations that surround a word or category.

This subversion view is not antagonistic to, but complementary to, the juxtaposition view of humour as embodied in mechanisms like bisociation, script switching and frame shifting. Subversion explains how, and why, new categories are created from old: to demonstrate the limitations of conventional wisdom as captured in conventional categories. We thus believe that the subversion view, as illuminated by a comparison to the workings of scientific thought experiments, goes further to explain (rather than simply describe) the creativity inherent in both joke production *and* joke understanding.

References

Attardo, S. and Raskin, V. (1991). Script theory revis(it)ed: joke similarity and joke representational model. *Humor: International Journal of Humor Research* 4-3, 293-347.

Attardo, S, Hempelmann, C. F. and Di Maio, S. (2002). Script oppositions and logical mechanisms: Modeling incongruities and their resolutions. *Humor: International Journal of Humor Research* 15-1, 3-46.

Brône, G., Feyaerts K. and Veale, T. (eds.) (2006). Special issue on 'Cognitive Linguistic Approaches to Humor'. *Humor: The International Journal of Humor Research* **19**(3).

Coulson, S. (2000). *Semantic Leaps: Frame-shifting and Conceptual Blending in Meaning Construction*. New York/Cambridge: Cambridge University Press.

De Mey, T. (2002). Thought experiments, conceivability arguments, and conceptual blending. *Odense Working Papers in Language and Communication* 23. A. Hougaard and S. Nordahl (eds.).

Dickens, C. (1843/1984). A Christmas Carol. Puffin Books, Middlesex, UK.

Fauconnier, G. and Turner, M. (1998). Conceptual Integration Networks. *Cognitive Science*, 22(2):133–187.

Fauconnier, G. and Turner, M. (2002). *The Way We Think*. Basic Books.

Fellbaum, C. (ed.). (1998). *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Freud, S. (1938). Wit and its relation to the unconscious, in *The Basic Writings of Sigmund Freud*, ed. A. A. Brill. New York: Modern Library.

Gendler, T. S. (1998). Galileo and the Indispensability of Scientific Thought Experiment. *The British Journal for the Philosophy of Science*, 49(3), pp 397-424.

Gendler, T. S. (2000). *Thought Experiment: On the Powers and Limits of Imaginary Cases*. London, UK: Garland Publishing.

Gibbs, R. W. (1993). Why idioms are not dead metaphors. In *Idioms: Processing, Structure and Interpretation,* C. Cacciari and P. Tabossi (eds.). NJ: Lawrence Erlbaum.

Giora, R. (2002). Optimal innovation and pleasure. In: O. Stock et al. (eds.), *The April Fools' Day Workshop on Computational Humou*r: Proceedings of the Twentieth Twente Workshop on Language Technology (Series TWTL 20). Enschede: UT Service Centrum.

Halliday, M. A. K. (1985). *An Introduction to Functional Grammar*. Edward and Arnold.

Hofstadter, D. and Gabora, L. (1989). Synopsis of the Workshop on Humor and Cognition. *Humor: International Journal of Humor Research* 2(4),417 –440.

Johnson, M. (1987). *The Body in the Mind*. Chicago, IL: Chicago University Press.

Koestler, A. (1964). *The Act of Creation*. New York: Macmillan.

Kuhn, T. (1964). A Function for Thought Experiments. Reprinted in Kuhn, T. *The Essential Tension*. Chicago: University of Chicago Press.

Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. Chicago University Press.

Langacker, R. W. (1991). *Concept, Image, and Symbol: The Cognitive Basis of Gramma*r. Cognitive Linguistics Research. Berlin and New York: Mouton de Gruyter.

Mach, E. (1960). The Science of Mechanics. J. McCormack (translator). LaSalle, Illinois: Open Court Press.

Mach, E. (1976). On Though Experiments. *Knowledge and Error*. Dordrecht: Reidel.

Norton, J. (1996). Are Thought Experiments Just What You Always Thought?. *Canadian Journal of Philosophy*.

Ortony (ed.) (1979). *Metaphor and Thought*. Cambridge: Cambridge University Press.

Ritchie, G. (1999). Developing the Incongruity-Resolution Theory. *In the proceedings of 1999 AISB Symposium on Creative Language: Stories and Humour*, Edinburgh, Scotland.

Raskin, V. (1985). *Semantic Mechanisms of Humor*. Dordrecht: D. Reidel.

Schank, R. C. and Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding*. New York: Wiley.

Searle, J. (1980). *Minds, Brains and Programs*. Behavioural and Brain Sciences 3(3), pp 417 – 457.

Suls, J. M. (1972). A Two-Stage Model for the Appreciation of Jokes and Cartoons: An information-processing analysis. The Psychology of Humor, eds. J. H. Goldstein and P.

E. McGheen, pp 81-100.

Veale, T. (2004). Incongruity in Humor: Root-Cause or Epiphenomenon? *The International Journal of Humor* 17/4, a Festschrift for Victor Raskin.

Veale, T. and O'Donoghue, D. (2000). Computation and Blending. *Cognitive Linguistics*, 11(3-4), special issue on Conceptual Blending.

Veale, T. (2002). Compromise in Multi-Agent Blends. *Odense Working Papers in Language and Communication* 23. A. Hougaard and S. Nordahl (eds.).

Veale, T., Feyaerts K. and Brône, G. (2006). Cognitive Mechanisms of Adversarial Humour. *Humor: The International Journal of Humor Research*, special issue on 'Cognitive Linguistic Approaches to Humor'. In preparation.