Evidence for online repair of Escher sentences[1]

Ellen O'Connor – *University of Southern California*
Roumyana Pancheva – *University of Southern California*
Elsi Kaiser – *University of Southern California*

Abstract. Illusory "Escher" sentences (*More people have been to Russia than I have*) are a unique puzzle for theories assuming full and thorough grammatical analysis and semantic composition during processing, since people generally appear to accept them before noticing that the meaning is incoherent. Prior work has shown, however, that meaning *is* extracted from the illusion: in particular, reactions to the illusion are sensitive to whether a comparison of events is possible (Wellwood et al. 2009, 2012), suggesting that the perception of grammaticality is related to the fact that determiners may sometimes count events (Krifka 1990). The series of self-paced reading experiments reported here suggest the illusion is ambiguous between an individual and a (less salient) event quantification interpretation, since reading times are generally slower for comparatives with plural predicates. Reading times indicate that the meaning of the illusion is plausibly shifted to a comparison of events at the point of anomaly – a costly process that causes observably slowed reading times. The success of this operation determines how felicitous an interpretation comprehenders can obtain for an essentially ungrammatical sentence.

Keywords: semantic illusions; shallow processing; comparatives; event-related determiner quantification.

1. Introduction

Perceptual illusions afford unique opportunities to study nonveridical aspects of cognition, as they can reveal inherent properties of the system in relative independence from the input. Illusions in language are important for many of the same reasons: they allow us to investigate the nature of the linguistic system, independent from the input (Phillips, Wagers, & Lau 2011). Such an approach is similar in spirit to that taken by generative grammar, which has made great advances by studying aspects of linguistic competence not readily observed in communication, namely the rejection of ungrammatical sentences.

"Escher sentences" like (1) (Montalbetti 1984) are a particularly robust illusion, so-named because of their resemblance to the famous M.C. Escher lithograph, *Ascending and Descending* (Liberman 2004). As with Escher's impossibly infinite staircase, itself based on the Penrose stairs (Penrose and Penrose 1958), people seem to accept (1) in spite of its global incoherence, at least initially. It is usually only upon closer examination that the infelicity of this sentence becomes apparent. The logical form of the main clause, constructed by the rules of syntax and

[1] Many thanks to Alexis Wellwood, Valentine Hacquard, Colin Phillips, Chris Barker, Jim Higginbotham and the audiences at Sinn und Bedeutung, AMLaP, University of Maryland and the USC Psycholinguistics Lab for their valuable feedback on various versions of this project. This research was supported by a National Science Foundation Graduate Research Fellowship.

compositional semantics, should require comparison of cardinalities of sets of individuals. The *than*-clause therefore needs to contain a bare plural noun phrase in subject position, from which degree abstraction will be possible (Chomsky 1977, Heim 2000, and others). (1) does not meet this requirement: the *than*-clause subject, *I,* does not contribute a cardinality of a set of individuals, and there is no other appropriate constituent to host the degree variable after ellipsis is resolved. Nevertheless, the sentence is largely accepted by English speakers.

1. a. More people have been to Russia than I have.
   b. More **λd . d-many people** have been to Russia
   than ***wh λd . d-many I** have been to Russia

A grammatical continuation of the first clause – as shown in (2) – will contain a bare plural noun phrase, either overt or covert, e.g. elided *people* in (2b). This is required by the determiner *more* and its incorporated *many,* a gradable determiner incorporating a measure function, whose semantics in (3) require a semantically plural NP for an orderly, non-trivial mapping of individual sums to degrees of cardinality (Hackl 2001).

2. a. More people have been to Russia than to Berlin.
   b. More **λd . d-many people** have been to Russia
   than **wh λd . d-many people** have been to Berlin

3. $[[\text{many}]] = \lambda d\ .\ \lambda P_{<e,t>}\ .\ \lambda Q_{<e,t>}.\ P(x) = 1\ \&\ Q(x) = 1\ \&\ |x| = d$ (Hackl 2001)

Meanwhile, the second clause of the illusion, *than I have,* is an acceptable continuation – but only for a comparative of another type, such as (4a). As with Escher's impossible staircase, the parts of the illusion are independently coherent, yet cannot compose with each other in a globally coherent way.

4. a. People have been to Russia more often than I have.
   b. More **λd .** People have been to Russia **d-much**
   than **wh λd .** I have been to Russia **d-much**

Given that the parser so rapidly and efficiently implements a range of grammatical constraints, it is a mystery why speakers regularly fail to notice this problem (see Phillips, Wagers & Lau 2011 for more discussion on this point). The common impression of (1) is that the incoherence of the input is ignored by the linguistic system, and the parts are integrated seamlessly, at least until closer consideration. This conclusion, however, is at odds with a tacit assumption in sentence processing that the semantic representations built online are largely accurate, fully specified and built compositionally. If the illusion is somehow interpreted, how can its semantics be derived compositionally without the comprehender becoming aware of the grammatical problem? And if no interpretation at all is assigned to the illusion, then how can it be perceived as coherent?

Indeed, semantic anomalies like Escher illusions raise interesting questions about the relationship between semantic composition and processing, providing potential motivation for "dual-route"

processing theories in which interpretations may arise online from simple interpretive heuristics in addition to full grammatical analysis, as mediated by the demands of the task or situation. Theories of "shallow" semantic processing often draw on evidence from semantic anomalies of various types (Sanford & Sturt 2002; Sanford & Graesser 2006). For example, the "Moses" illusion in (5) suggests that lexico-semantic integration may be less thorough when a target mismatching word fits closely into the global context of the sentence (Erickson and Mattson 1981, Oostendorp and Kok 1990, Oostendorp and De Mul 1990, Kamas et al. 1996, Hannon and Daneman 2001). Though it was not Moses who put animals on the ark, most speakers will answer the question in (5) by saying "two," even when warned in advance about the possibility of the question containing an anomaly.

5. How many animals of each kind did Moses put on the ark?

Proponents of 'Good Enough' theories have also argued for shallow processing at the syntax-semantics mapping. For example, Ferreira (2003) found that participants tended to incorrectly select *the dog* as the agent of (6) – such mistakes were more common for passive sentences, especially those where the syntactically faithful interpretation conflicted with world knowledge. This result is interpreted as evidence that the processor accesses heuristics, but in most cases their results converge with the syntax. In implausible passives, on the other hand, the use of an "agent-verb-theme" template (Bever 1970) and real-world knowledge to assign thematic roles may yield an outcome in conflict with the syntax, yielding occasional misinterpretations.

6. The dog was bitten by the man.

Escher sentences are good candidates for integration into a dual-route processing theory because of the clear mismatch between the output of the grammar and the general impression of the sentence, or the illusion's "percept". Townsend & Bever (2001) suggest that the illusion arises because it triggers "plausible sentence templates" (p. 184), leading people to accept the string before it has been sent to the grammar for analysis, and before any initial meaning has been assigned. This matches the informal observation that people only consciously notice the oddity of the illusion when they are asked to explain what it means. However, in a series of offline studies, Wellwood, Pancheva, Hacquard, Fults & Phillips (2009) established that listeners are sensitive to semantic properties of the sentence, suggesting that the illusion is interpreted and repaired. In particular, illusions containing predicates that can be repeated for a given subject (e.g. *called their families* in 7a) are consistently rated higher than those containing non-repeatable predicates (e.g. *graduated from high school*, 7b). Since comparison of events is compatible with the former but not the latter, they infer that speakers extract a comparison of events from the illusion.

7. a. Repeatable: More undergrads called their families during the week than I did.
   b. Non-repeatable: More New Yorkers graduated from high school this semester than I did.

Wellwood et al. (2009, 2012) note that event readings for determiners are already allowed by the grammar for sentences like (8) (Krifka 1990, Doetjes and Honcoop 1997, Barker 1999). Although the determiner *4000* in (8) most saliently counts individuals – yielding a one-to-one pairing of ships and lock-passings – an alternative reading is possible in which there are 4000 events of lock-passing involving possibly fewer than 4000 ship entities. To obtain these readings, Krifka (1990) posits the null determiner in (9), which combines with a quantized predicate such as *4000 ships* and an event relation (a VP denotation). This yields an *object-induced event measure relation* (OEMR) that measures events in terms of their participants, as in (10).

8. 4000 ships passed through the lock.

9. a. $[[D_{\varnothing}]] = \lambda P_{<e,t>} . \lambda R_{<e,<v,t>>} . \lambda e_{v} .$ OEMR (R)(e)(P)
   b. OEMR (R) is the smallest relation between events and quantity predicates, such that for any event *e* and quantity predicates P and Q:
      i. if *e* is not iterative with respect to R (i.e. there is no object that stands in R relation with respect to different parts of *e*), then OEMR(R)(e)(P) iff $\exists x . P(x) \& R(e)(x)$.
      ii. if *e* is iterative with respect to R (i.e., there is an object that stands in R relation with respect to different parts of *e*), then for any non-overlapping sub-events e_1 and e_2 if OEMR (R)(e_1)(P) and OEMR (R)(e_2)(Q) then OEMR (R)($e_1 + e_2$)(P+Q)

10. a. $[_{DP} [D_{\varnothing} [4000 \text{ ships}]] [_{VP} \text{passed through the lock}]]$
    b. $\exists e . \text{OEMR}([[\text{passed through the lock}]])(e)(\lambda x . \text{ships}(x)=4000)$
       if *e* is iterative, and has *n* non-overlapping, non-iterative sub-events, then e1 $n = 4000$

Krifka's analysis can be extended to comparatives, as he himself notes, and thus to the main clause in the illusion sentence in (1a). The *than*-clause will still be ill-formed, however, since the quantity *wh*-determiner, needed to create a degree predicate (Heim 2000 a.o.) does not compose with singular or definite NPs, here, the *than*-clause subject *I*. Given the absence of a quantity expression that can be a suitable argument for Krifka's null determiner, this determiner cannot be posited for the *than*-clause. The illusion sentence remains ungrammatical even under a Krifka-style event-related interpretation, as illustrated in (11) (with the problem in boldface).

11. More $\lambda d . [\exists e . \text{OEMR} ([[\text{went to Russia}]])(e)(\lambda x . \text{people}(x)=d)]$
    than *wh $\lambda d . [\exists e . \text{OEMR} ([[\text{went to Russia}]])(e)$**$(\lambda x . \mathbf{I(x)=d)]}$**

However, event-related readings may still be a factor in the relative acceptability of the illusion sentences. An event-related reanalysis of the main clause of the illusion, licensed by the grammar (possibly via Krifka's null determiner[2]), may be combined with an event-quantification reanalysis of the embedded clause through positing an event measure function *much,* along the

[2] The proper analysis of event-related readings of adnominal quantifiers remains a topic of continued debate (see Doetjes and Honcoop 1997, Barker 1999). We are not necessarily committed to the specifics of Krifka's account, but have incorporated his semantics to be as concrete as possible about how a reanalysis of illusions could proceed.

lines of (12) below. Such a reanalysis of the *than*-clause would not be licensed by the syntax of the matrix, which employs a determiner *more* and would require a corresponding determiner *how many* – this is why the illusion does not have a stable interpretation.

12. More λd . [∃e . OEMR ([[have been to Russia]])(e)(λx . people(x)=d)]
than wh λd . I have been to Russia d-much

The first aim of our experiments was to test for signs of reanalysis using online measures. Using self-paced reading we examined reading times at the auxiliary of the *than*-clause in sentences like (1), the point at which all remaining grammatical continuations are ruled out (e.g., *than I expected*). We reasoned that, under a shallow processing or heuristic-based approach to Escher sentences, we should observe no processing difficulty for illusions, because the sentence has not received thorough syntactic analysis at the time that people judge the illusion as acceptable. By contrast the event comparison approach predicts not only that the illusion is grammatically analyzed but that a process of reanalysis has been triggered in response to the problem.

Our secondary aim was to probe the mechanisms underlying the repair procedure by testing whether illusions that differ in offline acceptability are associated with different reading patterns. In particular, a slowdown at the critical region will only fully rule out shallow processing if it can be shown to occur even in cases where people seem to be "fooled" by the illusion. It is in principle possible, for example, for some people to detect the ungrammaticality of the illusion, slow down in reading times, and ultimately assign a low rating, while others read illusions and controls at equal paces, and then assign the illusions relatively high ratings. This pattern of responses would be compatible with shallow processing, under the plausible assumption that some participants processed items more deeply than others. To test this, we modulated offline ratings by relying on factors that are known to affect acceptability, namely the repeatability of the predicate and the plurality of the *than*-clause subject (Wellwood et al. 2009, 2012).

Although it is known that repeatability affects ratings for illusions, the question of why this is so is not fully resolved. Under a shallow processing account, perhaps repeatable predicates are easier to process, and thus the *than*-clause is not deeply parsed or thoroughly integrated into the matrix comparative. For example, predicates that can be repeated are permissible in a wider range of environments – both nominal and event comparatives – and therefore may be more frequently encountered in comparatives. Such an account leads us to expect faster reading times for repeatable predicates overall, even before the anomaly arises, while the event comparison approach predicts that a difference in reading times will show up only later, if at all, as a result from the relative ease of reanalysis.

2. Experiment 1: Effects of repeatability on illusion reading times

2.1. Methods and materials

In a combination self-paced reading, rating, and recall study, we tested whether real-time processing of illusion comparatives differs from that of non-illusion comparatives. We used a

within-subjects design that manipulated two independent variables: PRESENCE OF ILLUSION (illusion vs. control, shown in red vs. black in Table 1) and PREDICATE TYPE (repeatable vs. non-repeatable, shown underlined with solid vs. dashed lines in Table 1).

	REPEATABLE PREDICATE	NON-REPEATABLE PREDICATE
CONTROL	More judges vacationed in Florida **than lawyers did** because of the beautiful beaches and warm weather.	More judges retired to Florida **than lawyers did** because of the beautiful beaches and warm weather.
ILLUSION	More judges vacationed in Florida **than the lawyer did** because of the beautiful beaches and warm weather.	More judges retired to Florida **than the lawyer did** because of the beautiful beaches and warm weather.

TABLE 1. Conditions manipulated in Experiment 1 (predicate type X presence of illusion)

Illusion sentences were semantically and syntactically infelicitous; control sentences were parallel comparatives in which the common perception of the sentence's meaning matched that conveyed by its syntax. The illusion was created by substituting a singular definite noun phrase for a bare plural one: *than {the lawyer; lawyers} did*. The determiner type was counterbalanced, such that half of the illusions contained the definite article *the,* and half first-person possessive pronouns *my/our*. For reasons unrelated to the goals reported here we also counterbalanced quantifier type: half of the items contained comparative *more* and half the equative *as many*.[3]

Predicate type differed on the basis of whether the event could occur only once or multiple times per subject. Non-repeatable predicates preclude event comparison (*#The judge retired to Florida more than the lawyer*) and are known to reduce the acceptability of the illusion (Wellwood et al. 2009, 2012). The predicates were normed in an offline ratings study where a different set of 20 participants judged whether each predicate was compatible with frequency modifiers, e.g.: *#The lawyer retired to Florida three times* vs. *The lawyer vacationed in Florida three times*. Repeatable and non-repeatable predicates were matched for length, syntactic complexity, and were semantically as parallel as possible except for their repeatability. All predicates were simple past tense. The critical word was the auxiliary, *did/were,* where grammatical continuations of the illusion are no longer possible. The following spillover region was always eight words long and was the same for all conditions (e.g., *because of the beautiful beaches and warm weather*).

Eight lists were created, each containing 48 target items and 96 fillers. The lists rotated in a Latin Square design so that each participant saw only one condition of each item. Lists 5-8 were identical to lists 1-4 but were presented in reverse order to control for ordering effects. Each list had four blocks separated from each other by a rest period to reduce fatigue.

24 undergraduates from the University of Southern California (Los Angeles, USA) completed the experiment and were each paid $10 for their participation. The experiment was administered

[3] The fact that *as many* is longer than *more* will not affect the reading times we are interested in, since the critical region was defined as the portion of the sentence following the auxiliary *did/were*.

using Linger (Doug Rohde, MIT) and took approximately forty minutes to complete. Participants were instructed that the experiment investigated their "first impressions" of a variety of sentences. The sentences were presented one word at a time, masked by a series of dashes; pressing the space bar revealed one word and hid the preceding one, allowing us to measure how much time people spend reading each word before moving on to the next. After each sentence, acceptability ratings were assigned on a seven-point scale, using "1" for sentences that were "very bad" or that they "couldn't imagine an English speaker saying" and a "7" for sentences that "sounded perfectly fine or natural."

In order to gather production data, and to require participants to attend to the task at hand, participants were occasionally asked to recall the "gist" of an item out loud after assigning a rating, as much as they could recall; a similar recall task was used to probe production of illusions by Wellwood et al. (2012). This paraphrase task occurred on one third of the trials; participants did not know which sentences they would be required to paraphrase.

2.2. Predictions

Prior work has established that illusions are in fact rated lower than non-illusory control sentences, and that the predicate type determines the extent of the penalty (Wellwood et al. 2009, 2012). Illusions with a singular VP cannot be shifted to a comparison of events, which is presumably why they are rated lower. We expected to find the same pattern in our ratings, even with the two conditions matched very closely in length, syntactic structure, and semantic context. The question we were primarily interested in was whether the mechanisms leading to these ratings could be detected as disruptions in reading time. If comprehenders initiate semantic reanalysis or coercion in response to the grammatical problem, we expect to find evidence of slowdown at the critical region of the illusion but not the control. If on the other hand the illusion can go largely undetected by the parser, with the ratings patterns arising in an offline process, then we expect to find no differences in processing difficulty for illusions and controls.

We were also interested in investigating whether the offline ratings for illusions are predictive of the amount of difficulty experienced. One possibility is that illusions involve a process of anomaly detection and an attempted (and sometimes untenable) repair strategy; it is also possible that only illusions with non-repeatable predicates cause difficulty associated with anomaly detection, while highly acceptable illusions pass by wholly undetected and so may still be a byproduct of shallow processing. Finally, if illusions with repeatable predicates are somehow more acceptable because they are simply easier to process and therefore do not require in-depth grammatical analysis, we should expect to find faster reading times for repeatable items before the anomaly in the illusion is apparent.

2.3. Results

2.3.1. Ratings

The mean ratings by condition are given in Figure 2. Ratings were standardized prior to analysis based on each participant's mean ratings for all experimental items, including fillers. Subject and item means were entered into separate two-way repeated-measures ANOVAs with two factors: presence of illusion and predicate repeatability. These tests revealed that illusions were rated significantly less acceptable than controls ($F_1(1, 23) = 99.29$, $p < .001$; $F_2(1, 47) = 201.10$, $p < .001$). There was no main effect of repeatability; however, there was a significant interaction between predicate repeatability and presence of illusion, such that illusions with non-repeatable predicates were rated lower than illusions with repeatable predicates ($F_1(1, 23) = 10.57$, $p = .004$; $F_2(1, 47) = 8.23$, $p = .006$).

	ILLUSION: *…than the lawyer did*	CONTROL: *…than lawyers did*
NON-REPEATABLE: *More judges retired to Florida …*	4.30 (1.79)	5.82 (1.33)
REPEATABLE: *More judges vacationed in Florida …*	4.63 (1.72)	5.70 (1.52)

TABLE 2. Mean ratings by condition (standard deviation is given in parentheses).

Illusions with *as many* ($M = 4.38$, $SD = 1.78$) were rated similarly to those with *more* ($M = 4.53$, $SD = 1.74$, $ps > .2$). Illusions with definite determiners ($M = 4.61$, $SD = 1.68$) were rated higher than those with possessive determiners ($M = 4.31$, $SD = 1.83$), according to post-hoc t-tests ($t_1(23) = -2.77$, $p = .011$, $t_2(46) = -1.81$, $p = .077$).

2.3.2. Reading times

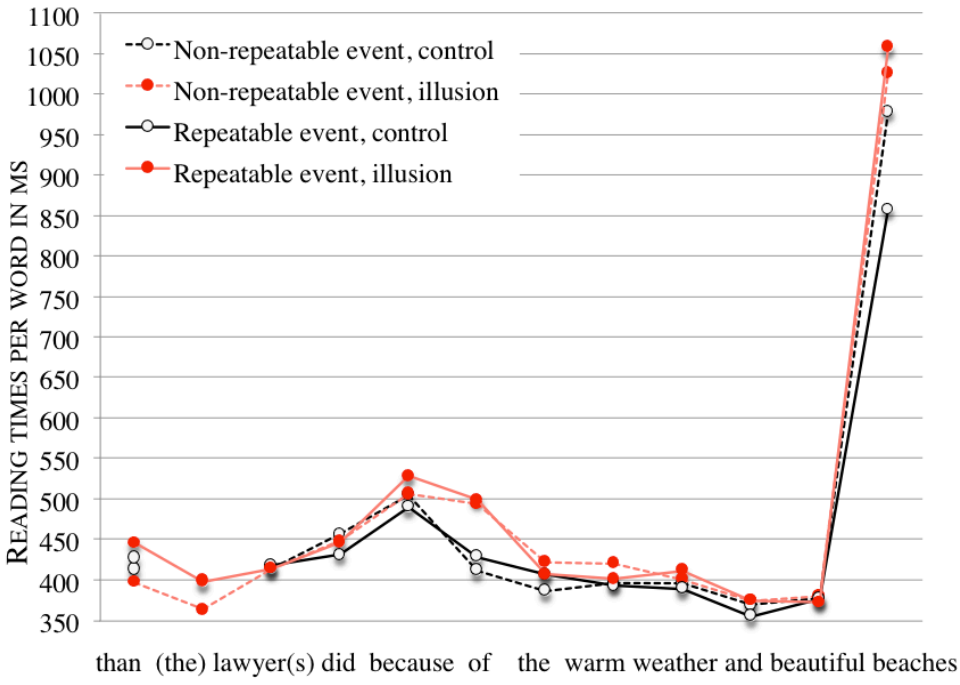

FIGURE 1. Reading times by word position

The reading time patterns are shown in Figure 1 above. Prior to analysis, extreme outliers (<100 ms or >4000 ms) were adjusted, followed by values over three standard deviations from the mean reading time per word position, calculated separately for each condition. This affected 1.3% of the data overall. For each word position, starting at *than* (or *as,* if the item contained the quantifier *as many*) through the end of the sentence, subject and item means were entered into two-way repeated-measures ANOVAs.

At *than/as,* comparatives with repeatable events were read 30.8 ms more slowly than items with non-repeatable predicates ($F_1(1, 23) = 8.44$, $p = .008$; $F_2(1, 47) = 6.09$, $p = .017$). Only the illusion conditions had determiners *my/our/the,* which were read 33.7 ms slower when the predicate was repeatable. Paired t-tests showed that this difference was significant by subjects and marginal by items ($t_1(23) = -2.17$, $p = .041$, $t_2(47) = -1.89$, $p = .065$). In sum, the repeatability of the predicate affected processing of all comparatives early on, before it was apparent whether or not the item was an illusion.

No effects were found at the noun phrase or auxiliary (e.g. *lawyers did*). Illusions and controls were read at equal speed, as were items with repeatable and non-repeatable predicates; there were no interactions.

Let us now examine the spillover region, consisting of the eight words following the auxiliary *did* (the point at which the illusion becomes ungrammatical). At the first word of the region, there was no main effect of the illusion or the predicate, and no interaction (all Fs < 1). At the second word, however, a main effect of the illusion condition was detected (in Figure 1, at *of*). At this position, illusions were read 76.9 ms more slowly than controls, a difference which was highly significant both by items and subjects ($F_1(1, 23) = 24.61$, $p < .001$; $F_2(1, 47) = 15.97$, $p < .001$). The repeatability of the predicate had no effect on the reading times, and did not interact with the illusion (Fs < 1).

In order to check whether slowdown at *of* was caused by conscious detection of the illusion, we examined whether there was any relationship between the amount of slowdown observed at this position, and the average amount the illusion was penalized in offline ratings. This was done by computing the difference between reading times for illusions and controls, and the difference in ratings for illusions and controls, averaged first across items and then across participants. The results were not significant in either case (ps > .4), suggesting there was no relationship between the two measures. Thus, the amount of slowdown observed at this position was no different for participants who rated illusions highly, or for items that were especially illusory. More broadly, illusions seemed to cause processing difficulty regardless of whether they were consciously detected and rated down.

In the following word position (in Figure 1: *warm*) there were no main effects of the illusion or the predicate type. However, we find indications of an interaction between repeatability and presence of illusion: non-repeatable illusions were read 35.6 ms slower than repeatable illusions, while controls were read at equal speed in both conditions. This numerical difference was

marginally significant by subjects, but was not by items ($F_1(1, 23) = 3.84$, $p = .062$; $F_2(1, 47) = 2.26$, $p = .14$). There were no other significant effects throughout the end of the sentence.

2.3. Discussion

The experiment outlined here investigated how deeply participants process Escher illusions. We explored two hypotheses: one hypothesis was that the illusion is caused when the processor builds a representation of the sentence at a level that is incomplete enough for the anomaly to pass undetected. This shallow representation may involve only detection of two broadly acceptable syntactic "templates" corresponding to each clause of the comparative, potentially before deriving any meaning (Townsend & Bever 2000).

The second hypothesis was that illusions are rated highly because they receive a felicitous interpretation via reanalysis. This hypothesis is motivated by prior work showing that comprehenders are sensitive to factors influencing the semantic coherence of event comparison (Wellwood et al. 2009, 2012). Experiment 1 replicated these findings. Although illusions seem to be highly acceptable, they are in fact rated significantly lower than controls. Non-repeatable illusions were rated the lowest of all, again suggesting that the availability of event comparison is an important contributor to the acceptability of the illusion.

The results here show, however, that regardless of offline acceptability, the processor is sensitive to the presence of the illusion. At the beginning of the spillover region, both illusion conditions were read more slowly than controls. This slowdown was very robust and all available measures indicated that it occurred almost completely independently of the rating eventually assigned to the illusion, or more broadly, the extent to which participants were consciously aware of the illusion. This fact seems to strongly rule out a shallow processing explanation of the phenomenon, suggesting instead that the parser can generally differentiate between the ungrammatical illusions and grammatical controls, and for this to happen, the illusion must be processed at a relatively deep level.

The perception of grammaticality likely arises from a relatively easy switch to event comparison, as hypothesized by Wellwood et al. (2009, 2012). Reading times suggest that event comparison is an available analysis that the processor considers when the predicate is repeatable. Slower reading times at and around *than* – before the illusion even arises – can be interpreted as evidence that comparatives with plural VPs are systematically ambiguous between individual and event quantification readings; maintaining both analyses takes processing effort and surfaces as slowed reading times. Importantly, this added cost for repeatable predicates is inconsistent with a shallow processing account of the ratings distribution. If illusions with repeatable predicates are rated highly because repeatable predicates are simply easier or more frequent, this added processing cost would be highly unexpected.

When individual quantification is made impossible there will be a shift to event comparison if such an option is available; otherwise the sentence will be perceived as unacceptable. Perhaps the failure of reanalysis is especially costly, since illusions with non-repeatable predicates caused

a marginally more prolonged slowdown than illusions with repeatable predicates, after the main slowdown.

Below we present a second experiment, which was designed to address several potential concerns related to our experimental design. First, it is possible that the added processing cost for illusions is related to the fact that participants were asked to paraphrase some of the experimental items. Such a task may force participants to process sentences more carefully than usual and thus to confront the illicit meaning more often than they normally would. To address this concern we changed the task such that participants only had to repeat as much of the item as they could remember, since this task is less likely to trigger deep processing of the meaning.

Second, we wanted to ensure that the slowdown was not caused by superficial properties of the illusion condition, which differed only in the presence of an extra determiner and the singularity of the *than*-clause subject NP. One possibility is that participants are sensitive to the number mismatch between two otherwise semantically parallel noun phrases (e.g., *judges… lawyer*). For example, it is known that in ellipsis constructions, including comparatives, the processor uses information about the prosodic and lexical parallelism of noun phrases to recover the ellipsis and correspondingly, the thematic role of the remnant (Carlson 2001). Therefore, to make the lexical properties of the noun phrases maximally similar we used all plural *than*-clause subjects (*than my lawyers did*). Since plurality in the *than*-clause also increases acceptability ratings for illusions (Wellwood et al. 2009, 2012), this allowed us to further test the finding that even maximally acceptable illusions are difficult to process.

3. Experiment 2: Effects of subject plurality on illusion reading times

3.1. Methods and materials

As before, we used a self-paced reading task combined with offline acceptability judgments and production, except that participants were required to *repeat,* not paraphrase, as much as they could remember of the target items on certain trials. The design was the same, with two independent variables: PREDICATE REPEATABILITY (repeatable vs. non-repeatable) and PRESENCE OF ILLUSION (illusion vs. control).

The 24 target items and 60 fillers were adapted with minimal changes from the previous experiment. First, all of the *than*-clause subjects were pluralized, such that the illusion condition always contained definite plural NPs, while the controls contained bare plurals: *…than* {*lawyers; my lawyers*} *did*. With plural NPs, however, there is a possibility that participants find illusions acceptable for an uninteresting reason, namely that they read the sentence very quickly and fail to notice it contains a determiner, parsing it as a regular nominal comparative. This repair would be unrelated to the process of interest, which persists even with pronouns and singular NPs and therefore is not caused by dropping or ignoring determiners. To avoid this possibility we used only the items from the first experiment that had possessive determiners, which by virtue of having more semantic import are less likely to be ignored. We also hoped that this change would

reduce the likelihood that reactions to illusions would be affected by complications associated with the uniqueness presupposition or discourse requirements of the definite article.

The experiment was conducted online using Ibex (designed by Alex Drummond, University of Maryland, http://spellout.net/ibexfarm). The participants (n=40) were recruited from Amazon Mechanical Turk, and were paid $2.25 for their participation. All participants were native monolingual English speakers with a U.S. IP-address and with a task approval rating of 97% or higher. Participants who correctly answered items that tested that they were using the appropriate end of the rating scale (for sentences such as *The salad build a fork ten times* or *Mary went to the store yesterday*) or items ensuring their attention (*If you understand this sentence, assign it the lowest possible rating*) were included in the analysis. Due to the different experimental settings – online using Mechanical Turk, as opposed to the lab environment in Experiment 1 – we caution that the reading times between the two experiments may not be directly comparable to each other. Our goal for this experiment was primarily to see if the main findings from Experiment 1 could be replicated, rather than to examine whether or how reaction times would change in magnitude or timing across the two experiments.

The procedure was similar to that of Experiment 1. Participants were told that the repetition task would be difficult and that they should not worry about minor inaccuracies but should focus on providing reasonable ratings to the sentences they were rating. On a third of the trials participants were prompted to type into a text box as much as they could remember of the previous item. The repetition task was, as before, placed at random intervals such that participants could not predict which items they would have to repeat.

3.2. Results

3.2.1. Ratings

Prior to analysis the ratings were again standardized to z-scores based on mean ratings across all items for each participant. The illusion condition was rated lower than the control condition; this difference was statistically significant both by subjects and items ($F_1(1, 39) = 22.069$, $p < .001$; $F_2(1, 23) = 9.218$, $p = .006$). There was no main effect of repeatability, and no interaction. The raw mean ratings are given in Figure 4.

	ILLUSION: *…than my lawyers did*	CONTROL: *…than lawyers did*
NON-REPEATABLE: *More judges retired to Florida*	5.09 (1.61)	5.54 (1.44)
REPEATABLE: *More judges vacationed in Florida*	5.09 (1.71)	5.40 (1.64)

TABLE 3. Mean (standard deviation) of raw ratings from Experiment 2.

3.2.2. Reading times

Analysis proceeded as before; resulting reading times are plotted in Figure 2.

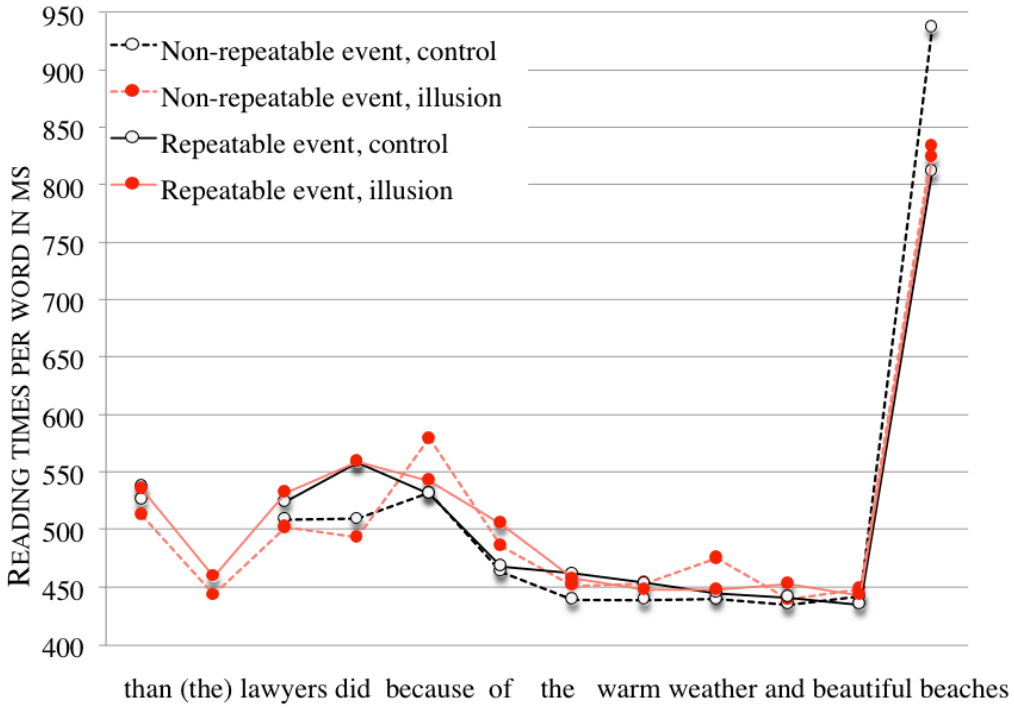

FIGURE 2. Reading times by word position (in ms)

Prior to the critical word, *did,* there were no main effects and no interactions. At *did,* items with repeatable predicates were read 57 ms more slowly than those with non-repeatable predicates, a difference that was significant by subjects ($F_1(1, 39) = 4.64$, $p = .037$), though not by items ($F_2(1, 23) = 2.32$, $p = .136$).

At the word following *did,* illusions were read more slowly than controls ($F_1(1, 39) = 4.53$, $p = .04$, $F_2(1, 23) = 6.46$, $p = .018$), while there was no main effect of repeatability. Numerically, illusions with non-repeatable predicates seem to incur a larger processing cost than those with repeatable predicates; however, this interaction was not statistically significant ($F_1(1, 39) = 1.47$, $p = .23$, $F_2(1, 23) = 0.90$, $p = .35$). At the following word (*of*), illusions were again read marginally more slowly by subjects $F_1(1, 39) = 3.96$, $p = .054$, $F_2(1, 23) = 2.06$, $p = .17$) by 31 ms; there was no main effect of repeatability. At this position, it was repeatable illusions that were read numerically slower than non-repeatable ones, but again the interaction was not significant (Fs < 1).

There were no other main effects or interactions until five words after *did* (in Figure 5, at *disapproved*), where illusions were again read more slowly than controls by subjects ($F_1(1, 39) = 4.49$, $p = .04$, $F_2(1, 23) = 2.32$, $p = .14$). This effect was driven by a marginal interaction:

nonrepeatable illusions were read 40 ms more slowly than controls, while reading times for repeatable illusions were equal to those of controls ($F_1(1, 39) = 3.58, p = .066, F_2(1, 23) = 3.80, p = .063$). There were no other significant effects throughout the end of the sentence.

3.3. Discussion

In this experiment we sought to test whether the main findings from Experiment 1 would hold with several design changes. First, rather than asking participants to paraphrase the target items at random intervals, we asked them only to repeat as much as they could remember. This was to ensure that the experimental task allowed for shallow processing, particularly since the availability of shallow processing is thought to be mediated by task-specific demands (Sanford & Sturt 2002; Ferreira 2003, 2007). If a task requires thorough comprehension, full algorithmic analysis would be unsurprising. Since item repetition does not require thorough comprehension of the meaning of the item, it should allow for the use of processing heuristics. Second, we tested illusions of a different sort, namely those with plural *than*-clause subjects. Plural illusions differ first in that they are generally found to be more acceptable (Wellwood et al. 2009, 2012) and second in that they rule out superficial number mismatch between similar noun pairs (e.g., *judges*/*lawyer*). Thus, the items in this experiment should be both maximally illusory, and should also fully facilitate shallow processing.

The key findings from Experiment 2 were largely the same as those in Experiment 1, with some differences in timing. The effect of the illusion surfaced in roughly the same region. However, repeatability effects showed up later in Experiment 2: the main effect was detected at the end of the *than*-clause (instead of the beginning), and the marginal interaction near the end of the spillover region. We refrain from drawing any strong conclusions about why these effects surfaced in slightly different parts of the sentence, since the experimental conditions themselves (online vs. in person) may be responsible for the difference, as opposed to the *than*-clause subject. Most important for our purposes is the fact that the nature of the effects, as well as their relative order, was constant across both experiments.

As expected, plural illusions were rated numerically higher than the singular illusions in Experiment 1, a result fully in line with prior research. Wellwood et al. (2009, 2012) attribute this effect to the resulting plurality of the VP, i.e. the fact that – whether the predicate is repeatable or not – a plural subject will naturally allow for multiple events under a distributive reading, even if the action is completed at most once by each participant. A natural prediction that arises from this analysis is that subject plurality and repeatability will both affect the illusion, but potentially not additively: the illusion is rated highly whenever some mechanism is available to obtain an event comparison interpretation, and it is rated down when no such mechanism is available. Perhaps it makes no difference if one or two different routes to event comparison are available. This is in fact what our results suggest, since we observed a robust interaction between the presence of the illusion and the repeatability of the predicate with singular *than*-clause subjects, but not plural ones. While other experiments have always found non-repeatable illusions to be less acceptable than repeatable illusions, this was the first experiment to include all plural items. Since the items from Experiment 2 were adapted from

Experiment 1, the differences responsible for this are reduced to either the nature of the *than*-clause NP, or possibly, a peculiarity associated with Mechanical Turk.

Even though repeatability did not affect the ratings, it did continue to influence reading times, indicating that it still played an active role in the processing of comparatives. At the critical word (the auxiliary), we observed a slowdown for items with repeatable predicates – also found in Experiment 1 – which we have interpreted as evidence that the matrix clause of a comparative with a repeatable VP is compatible both with event and individual readings. This ambiguity is plausibly related to the fact that determiners tend to elsewhere allow for event-counting readings (Krifka 1990 and others).

As in Experiment 1, illusions with non-repeatable predicates were more difficult than those with repeatable predicates, though this effect was again only marginally significant and occurred very late in the sentence. Further research is needed to probe whether the result is in fact reliable, but if it is, we speculate that that it may be related to the generation of a distributive reading in which a plurality of events is obtained by applying the predicate once to each member of the set denoted by the *than*-clause subject NP. This is consistent with prior work showing that distributivity incurs a processing cost, possibly because distributive readings involve postulating an additional operator in the semantics (Frazier et al. 1999).

Finally but importantly, despite the changes in design to facilitate shallow processing and maximize the illusion, we again found that illusions were worse than controls, both in terms of ratings and reading times. Participants rated illusions reliably lower than controls, and read the region immediately following the critical word more slowly. The fact that this key finding remained stable across both experiments establishes that the facts of this illusion are more complicated than can be accommodated by a simple shallow processing account.

4. Conclusions

The two self-paced reading experiments presented here probe a curious phenomenon in natural language, namely the existence of sentences that seem to sound acceptable but yet have no coherent meaning. It is a puzzle to many (if not most) sentence processing theories how Escher sentences arise: acceptability judgments should logically arise from grammatical analysis, which in turn should alert the system to the presence of an anomaly. This issue has been integrated into e.g. the dual-route theory of Townsend & Bever (2000), who use Escher sentences as evidence for the use of superficial heuristics and grammatical analysis alike in sentence processing. Strings are first matched to syntactic templates using processing heuristics, and then later sent to the grammar for analysis. Because the illusion here contains two locally coherent clausal templates, the string is perhaps accepted before algorithmic parsing, and therefore before the incoherence of the meaning can be determined.

However, the evidence as a whole suggests that this phenomenon is more complicated than this initial analysis allows for. Of the various factors that have been suggested to contribute to the acceptability of the illusion, the two that are robustly confirmed by experimentation are (i) the

plurality of the *than*-clause subject NP, and (ii) the repeatability of the event described. Both are closely tied to the semantics of event comparison, implicating an erroneous repair of the illusion towards a comparison of cardinalities of events (Wellwood et al. 2009, 2012). The results we report are largely consistent with this account: illusions appear to cause significant slowing in reading times, representative of the difficulty the parser experiences when confronting the ungrammatical structure. Importantly, *all illusions* incur this processing penalty, even the most acceptable sounding ones, even in situations that should clearly allow for shallow processing. The evidence presented thus further argues that the illusion is processed in at least enough depth to distinguish between illusions and controls, with the perception of acceptability likely related to the availability of a repair, plausibly to a comparison of events.

Although the exact nature of the repair process remains mysterious, it does *not* likely consist of moving *more* to an adverbial position, since the illusion persists equally with degree quantifiers that are unambiguously nominal, namely *fewer* (Wellwood et al. 2009, 2012) and *as many* (in our experiments). This points to a shift that happens at the level of semantics and not syntax, a conclusion consistent with the timing of the slowdown – slightly *after* the critical word, in the spillover region, instead of the immediate effect typically observed for garden path sentences. Similarly late timing is reported, for example, by Traxler. Pickering and McElree (2002, Experiment 3) for the region *following* the target word in cases of complement coercion (e.g. *The boy started the puzzle/fight* ***after***…). Inverse scope interpretations are also reported to incur a processing cost that seems to manifest itself late in the spillover region (Anderson 2004:74).

A recent fMRI study also supports the conclusion that illusion sentences do not receive the same treatment by the parser as garden paths: Christensen (2010) found that garden path sentences activated regions of the left inferior frontal gyrus (LIFG), as well as premotor and posterior temporal cortices. Illusion sentences with singular *than*-clause subjects, on the other hand, elicited *less* activity in those regions than control comparatives. This effect is interpreted by Christensen (2010) as support for shallow processing, but we suggest instead that it is simply further evidence that the parser can and does differentiate between illusions and controls, and that the relevant processes happen at the level of semantics, not syntax. It is unsurprising then that the two constructions would activate different neural networks, since they involve different processes of recovery and repair.

Interestingly, the results reported here suggest that the event comparison reading is not derived in response to the ungrammaticality of the illusion but rather is already an available interpretation, as determined by the semantic context of the predicate. The availability of event comparison seems to clearly modulate reading times during the *than*-clause, perhaps an indication that the matrix clause ambiguously measures both cardinalities of sets of individuals as well as sets of events; maintaining both analyses incurs a processing cost. If this effect is related to the ambiguity that Krifka (1990) observes, as we suppose it is, then we also expect to find slowed reading times for non-comparative determiners with repeatable predicates.

In the case of illusions with plural *than*-clause subjects, an event reading is allowed whether or not the predicate is repeatable, so long as a distributive reading can be derived. Wellwood et al.

(2009, 2012) suggest that this fact explains why illusions with plural *than*-clause subjects receive higher ratings than those with singular subjects. In fact, we found that subject plurality completely wiped out the effects of repeatability on the illusion, further supporting this approach. Non-repeatable illusions seemed to still be slightly more difficult to process, but this would be expected assuming that distributive readings, when available, require additional grammatical processes – including postulating a null distributive operator – and so may be computationally costly to derive (e.g., Heim 1991, Frazier et al. 1999). If this account is correct, it crucially predicts that environments that rule out distributive readings will make illusions with plural *than*-clause subjects much worse. One way to test this is to look at the effect of the illusion on collective vs. distributive non-repeatable predicates, with the prediction that sentences like (13) are perceived as more acceptable than ones like (14):

13. More freshmen were expelled from school last semester than the seniors were.
14. More freshmen gathered in the quad last night than the seniors did.

It remains a mystery why even the worst illusions receive still relatively high ratings. This issue, however, is not exclusive to comparative illusions; it also arises in cases of syntactically unlicensed ellipsis. For example, it is widely known that people accept ellipsis constructions even when the antecedent mismatches in voice (e.g., (15)), and when the result violates syntactic constraints (16).

15. In March, four fireworks manufacturers asked that the decision be reversed, and on Monday the ICC did [reverse the decision]. (Darymple 1991)
16. Bill_i defended himself_i against the accusations because his lawyer_j couldn't [defend $\text{himself}_{i/j}$] (Darymple 1991)

An ongoing topic in the literature concerns how to account for the acceptability of cases like (15-16) without overgenerating the range of structures that people will accept. Interestingly, while the illusions in our experiment were rated much better than fillers that were plainly ungrammatical (e.g., comparatives with island violations), they received approximately the same range of ratings as fillers with antecedent mismatch in ellipsis. Our experiments were not designed to test this similarity explicitly, but the analogous ratings open up the possibility that these two puzzling cases can be grouped together, perhaps both as cases of "acceptable ungrammaticality" (Otero 1972, Staum and Sag 2007, Frazier 2008).

Semantic anomalies are often used as evidence that the structures and meaning considered online are not always constrained by the output of algorithmic, compositional semantics. On the face of it, Escher sentences seem to provide compelling support for the use of superficial heuristics to process meaning, potentially including ones that track how well a string maps onto familiar clausal templates (Townsend and Bever 2001). Yet so far, the facts of this phenomenon are most easily accounted for by drawing on the compositional semantics of comparison and plurality (Wellwood et al. 2009, 2012). Although the details remain to be worked out, the broad strokes of the semantic account explain a number of facts about how comprehenders react to illusions, both online and off. This suggests that a fruitful way of understanding semantic anomalies may

involve making connections with existing semantic analyses that are readily available in the formal literature, and thus that the processing mechanisms that give rise to illusion sentences may not be not be as grammar-independent as is sometimes thought.

References

Barker, C. (1999). Individuation and quantification. *Linguistic Inquiry* 30:4, 683-691.

Doetjes, J. and Honcoop, M. (1997). The semantics of event-related readings: a case for pair-quantification. In Szabolcsi, A. (ed.), *Ways of Scope Taking*. Dordrecht: Kluwer, 263-310.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology* 47: 164-203.

Ferreira, F., V. Ferraro and K. Bailey. (2002). Good enough representations in language comprehension. *Current Directions in Psychological Science* 11: 11-15.

Frazier, L., K. Pacht, and K. Rayner (1999). Taking on semantic commitments, II: collective vs. distributive readings. *Cognition* 70(1): 87-104.

Frazier, L. (2008). Processing ellipsis: a processing solution to the undergeneration problem? In *Proceedings of WCCFL 26,* Cascadilla Press.

Hackl, M. (2001). Comparative quantifiers and plural predication. In K. Megerdoomian and Leora Anne Bar-el (eds.), *Proceedings of WCCFL 20,* Cascadilla Press.

Heim, I. (2000). Degree Operators and Scope. In *Proceedings of SALT X,* Cornell University.

Liberman, M. (2004). "Escher sentences." *Language Log* post, http://itre.cis.upenn.edu/~myl/languagelog/archives/000862.html

Krifka, M. (1990). 4000 ships passed through the lock: object-induced measure functions on events. *Linguistics and Philosophy* 13: 487-520.

Montalbetti, M. (1984). After binding. PhD dissertation, MIT.

Penrose, L. and R. Penrose (1958). Impossible objects: a special type of visual illusion. *British Journal of Psychology* 49(1): 31-33.

Phillips, C., M. Wagers and E. Lau (2011). Grammatical illusions and selective fallability in real-time language comprehension. J. Runner (ed.), *Experiments at the Interfaces,* Syntax & Semantics, vol. 37. Bingley, UK: Emerald Publications.

Sanford, A. and A. Graesser (2006). Introduction: Shallow processing and underspecification. *Discourse Processes* 42: 99-108.

Sanford, A. and P. Sturt (2002). Depth of processing in language comprehension: not noticing the evidence. *Trends in Cognitive Sciences,* 6 (9): 382-386.

Townsend, D., and T. Bever (2001). *Sentence Comprehension: The Integration of Habits and Rules*. Cambridge, MA: MIT Press.

Wellwood, A., R. Pancheva, V. Hacquard and C. Phillips (2009). The role of event comparison in comparative illusions. Poster at CUNY, University of California, Davis.

Wellwood, A., R. Pancheva, V. Hacquard and C. Phillips (2012). Deconstructing a comparative illusion. Ms, UMD and USC.