COMPARATIVE ILLUSIONS AT THE SYNTAX-SEMANTICS INTERFACE

by

Ellen O'Connor

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(LINGUISTICS)

December 2015

For Matt

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Psycholinguistic research has focused much attention on the factors that influence structural ambiguity resolution, under the assumption that meaning is derived from a selected syntactic representation in a systematic, compositional way. Problematically, however, researchers have increasingly observed examples suggesting that perceptions of sentence acceptability and meaning are not always straightforwardly constrained by logical semantics. Most English speakers, for example, initially accept the sentence *More people have been to Berlin than I have* until asked to explain more clearly what it means, at which point its meaninglessness becomes obvious. Meanwhile, the sentence *No head injury is too trivial to ignore* is overwhelmingly perceived to mean exactly the opposite of its implausible grammar-based meaning, an error that is only readily detected with extended conscious effort.

The goal of this thesis is uncover what these "semantic illusions" tell us about semantic processing by identifying the locus of nonveridical processing. In spite of appearances I argue that it is impossible to explain perceptions of and reactions to these illusion sentences without referencing properties that influence their logical form; this suggests, contrary to existing proposals, that the illusion is generated by online computations associated with, not external to, the grammar, and that nonveridical perceptions are induced by processing mechanisms responsible for navigating the logical form of sentences that contain probable speech errors and/or those that fall at the outer boundaries of computational tractability. Because the source of these illusions is not well understood, a more general goal of this work is to establish their fundamental properties, including the nature of their percept(s) and the properties that modulate that percept, to pave the way for future research.

1   INTRODUCTION

1.1   Illusions in language

Perceptual illusions have long been a fruitful topic in the cognitive sciences, as they provide a rich source of information about how the brain builds representations of the world out of input that is noisy, incomplete and often massively ambiguous. For example, although humans consciously enjoy a stable, high-resolution, and full-field representation of visual scenes, in actuality our peripheral vision has only the resolution of a frosted shower door; we are unaware that we have blind spots on each retina, and are apparently unperturbed by the discontinuous visual experience that should arise from our continuous saccades. In addition, the visual system needs to infer a three-dimensional percept out of two-dimensional images projected onto the retina; this "inverse problem" cannot be solved unless the raw input is enriched with constraints for interpretation provided by the visual system – including, for example, knowledge that objects that are nearby tend to move further across our field of view than those that are far away (*motion parallax*). Illusions arise in situations where these assumptions about the visual environment yield an incorrect "hypothesis" – a perception that diverges from objective reality – thus affording the opportunity to investigate the "rules" of perception, including the "unconscious inferences" drawn from sensory data (Helmholtz 1866) in relative independence from the input.

The processing of language input presents many of the same challenges for the brain as other perceptual input. Linguistic processing in real time is extremely rapid, with normal fluent reading proceeding at a rate of three to six words per second; in addition, language input can be noisy or ambiguous along multiple dimensions, and thus must also be supplemented with additional "rules" for interpretation. It is therefore not surprising illusions can be observed in language as well as sensory perception. Much recent work has focused on cases where syntactic violations are fleetingly allowed during processing, such as agreement spreading errors (such as *The key to the cabinets <u>are</u> on the table;* Bock & Miller 1991, Vigliocco & Nicol 1998, Solomon & Pearlmutter 2004, Pearlmutter, Garnsey & Bock 1999, Wagers, Lau &

Phillips 2009), or the consideration of illicit binding configurations (e.g., *John thought that Bill_i owed him_i another chance to solve the problem*; Badecker & Straub 2002, Runner et al. 2006). Such illusions are unexpected if syntactic constraints are straightforwardly implemented during processing (e.g., Phillips 1996); thus, they raise questions about the nature of online structure-building, and the degree to which this process is influenced by the properties of the extralinguistic cognitive architecture, including properties of memory retrieval or attention (Phillips, Wagers & Lau 2011).

The work here focuses on illusions of a slightly different sort, namely those that can inform our understanding of online computations at the syntax-semantics interface. A common perception of (1)-(2) is that they are grammatically correct and felicitous English sentences. In fact, the sentence in (1) is grammatically ill-formed, while the sentence in (2) is well-formed but pragmatically anomalous; these facts will become clearer in subsequent chapters, where the actual and perceived meanings of the two sentences will be laid out more clearly, respectively. Both phenomena suggest that our perception of linguistic stimuli – not only in terms of acceptability, but also in terms of meaning – may robustly mismatch the output that the grammar provides. Because both illusions appear to involve nonveridical processing of the grammar of comparison, I use *comparative illusion* as an umbrella term for both phenomena.

(1)   a. More people have been to Berlin than I have.
       b. Grammar-based meaning: how many people have been to Berlin > how many me has been to Berlin.

(2)   a. No head injury is too trivial to ignore.
       b. Grammar-based meaning: All head injuries should be ignored, even the most trivial ones.

Note that the term *semantic illusion* is most frequently used in the psycholinguistics literature to refer to *Moses* illusions and to the lack of an N400 response to thematic role reversals (*At breakfast, the eggs would eat…*, Kuperberg 2007 a.o.). The former illusion is likely related to partial processing at lower levels of semantic analysis, such as word retrieval,

while categorization of the latter as an "illusion" depends crucially on certain (not universally accepted) assumptions about the nature of the N400 component itself. The studies here do not address these types of phenomena, but rather exclusively investigate comparative illusions that arise at the syntax-semantics interface. By investigating illusions with and without a well-formed LF, the experiments here can address grammar-parser relations at two different levels.

1.2   Implications for theories of semantic processing

To date, the little that is known about illusions like (1)-(2) comes from just a handful of studies (Wellwood et al. 2009, 2012; Wason & Reich 1979; Natsopoulos 1985), although a better understanding could have important implications for hypotheses linking language competence and performance. Much of the research that engages with these illusions thus far has used them as examples of "shallow" or heuristic strategies in sentence processing (Townsend & Bever 2001; Sanford & Sturt 2002; Sanford & Graesser 2006). Such theories propose, contra common and implicit assumptions in sentence processing, that linguistic input may receive only partial grammatical or semantic analysis, with comprehension potentially influenced by extra-grammatical processing strategies (Townsend & Bever 2001; Ferreira et al. 2002; Ferreira & Patson 2007).

For example, at lower levels, mechanisms for lexical access may allow for accessing a family of related word meanings, so that a word can be accepted on the basis of this "partial match" without more detailed analysis. This type of lexical underspecification is already commonly thought to underlie the processing of polysemy (e.g. *The dictionary is heavy/informative*; Frazier & Rayner 1990) and aspectual coercion (*The insect glided until it reached the garden;* Pickering et al. 2006), since comprehenders do not appear to show the usual behavioral patterns associated with resolving ambiguity. The task of disambiguating a particular word sense seems to be delayed until necessary. For shallow processing accounts, semantic underspecification is thought to be pervasive, and need not be temporary; certain aspects of meaning may never be resolved at all.

At higher levels of analysis, shallow processing mechanisms can be used to contend with the explosion of ambiguities in sentences with multiple quantifiers. A common technique

4

in computer science is to bypass the problem of disambiguation by using a pseudo-logical form that leaves this aspect of interpretation unspecified (Hobbs & Shieber 1987, Poesio 1996). Presumably, this type of massive ambiguity could also be problematic for a sentence processor with limited time and resources (Sanford & Sturt 2002; Sanford & Graesser 2006), and the details of how scope configurations differ may not always be crucial for communication. For example, although (3) can have multiple interpretations depending on the relative scope of the two quantifiers, Sanford and Graesser (2006) argue that a dialogue like (4) may not require resolution of this type of ambiguity; this explains why reading times for wide and narrow scope continuations, as shown in (5), do not substantially differ (Tunstall 1998; but see Andersen 2004): comprehenders are using underspecified representations and therefore are not committed to either scope ordering at the point of disambiguation.

(3)   a. Every kid is up a tree.

        b. ∀ > ∃: ∀x[kid(x) → ∃y[tree(y) & up(y,x)]]

        c. ∃ > ∀: ∃y[tree(y) & ∀x[kid(x) → up(y,x)]]

(4)   Mother: Did my kid manage to climb a tree?
        Babysitter: Right now every kid is up a tree!

(5)   Kelly showed every photo to a critic last month. The critic/critics were from a major
        gallery.

        Partial semantic analysis leaves the door open for anomalies to pass undetected. Whereas elements we attend to may receive a full interpretation, those outside of focus may not be fully interpreted and hence could be more error-prone. For example, work on the "Moses illusion" has shown that in response to the question in (6) people rarely point out the fact that it was Noah, not Moses, who put animals on the ark (Erickson & Mattson 1981), an error apparently related to the fact that *Moses* and *Noah* share many overlapping semantic features (Oostendorp & De Mul, 1990). However, when the focus structure of the sentence is changed and the same item is presented in an *it*-cleft as in (7) the error rate is greatly

diminished (Bredart & Modolo 1989; Bredart & Doquier 1989). This suggests that the word *Moses* can be analyzed to different depths, depending on the level of attention it receives. Focus structure presumably modulates the amount of attentional resources dedicated to lexical processing, and when a word is out of focus, comprehenders may fail to recognize that it has been replaced with a conceptually similar alternative.

(6)   How many animals of each kind did Moses put on the ark?

(7)   It was Moses who put two animals of each kind on the ark.

Other models, while similar in spirit to Sanford (2002) and Sanford & Sturt (2002), also allow for partial processing but additionally posit a set of "fast and frugal" interpretational heuristics (Townsend & Bever 2001, Ferreira, Bailey & Ferraro 2002, Ferreira & Patson 2007), which can fill in interpretational details when grammatical analysis is incomplete. Such heuristics hypothesize probable meanings based on statistical tendencies (for example, a bias towards interpreting noun-verb-noun sequences as agent-verb-patient) or real world "schemas" (including the ways that animate and inanimate entities typically interact), citing (among other things) unexpected misinterpretations of implausible passive sentences such as *The dog was bitten by the man* (Ferreira 2003). Ferreira argues that such misinterpretations arise because the meaning sanctioned by the grammar is in conflict with the output of a "NVN" heuristic, which typically assigns an agentive thematic role to the first noun phrase in the sentence, in this case *the dog.* Although in most cases the output of the grammar prevails, such conflict may occasionally lead to misinterpretations.

Probabilistic information and context are already often assumed to play a role in language processing, particularly to adjudicate structural ambiguities. However, in most sentence processing models probabilistic information is used to help *refine* grammatical processing by enabling selection among competing representations; if incoming structural properties weigh against a prior selection, that selection will be overridden. The central proposal of Ferreira, Bailey & Ferraro (2002) is that this is not necessarily the case, and that heuristics can *independently* determine sentence meaning if grammatical analysis is only partial, or if the two are in conflict. Thus, a crucial component of the proposal is the larger role

that heuristics are assumed to play in comprehension, and the resulting claim that it is occasionally possible to derive meaning independently from syntax and logical semantics.

Heuristics are well-motivated in domains such as decision-making, where foundational work by Tversky & Kahneman (1974) laid out a set of cognitive biases that can prevent humans from making fully rational decisions (for example, a tendency to overestimate the likelihood of an event that is easier to retrieve from memory). Heuristics or biases are thought to conserve resources and enable people to act rapidly in response to difficult tasks by ignoring part of the information available; however, they can also lead to systematic errors in judgment ("cognitive illusions"). Cognitive illusions are often taken to suggest the existence of two fundamentally different cognitive systems (Evans and Over 1996; Sloman, 1996; Stanovich and West, 2000; Evans, 2008): while "System 1" uses affective, heuristic or associative processes, "System 2" relies on abstract, serial or rule-based processes, monitoring the rapidly and unconsciously-generated output from System 1. Because System 2 is heavily constrained by working memory, individuals with better (analytical) cognitive resources will tend to perform better in reasoning and decision-making, relying more heavily on System 2, while individuals with fewer resources may fail to monitor or correct mistaken output generated by the heuristics employed by System 1, resulting in more errors.

Ferreira, Bailey & Ferraro (2002)'s proposal explicitly draws on this type of research in suggesting a similar use for heuristics in language comprehension, the over-application of which can lead analogously to comprehension errors. Ferreira et al. do not elaborate on this analogy in detail, although the topic merits closer consideration, since it is indeed an important question whether and how the processing strategies used in the context of language comprehension overlap with those used in the context of other cognitive domains. For example, the ambiguity of language input necessitates a series of parsing decisions, and in theory these decisions could be similarly informed by all, or only part, of the available information. In terms of semantic processing, the two routes to interpretation would include a less effortful route similar to System 1 that makes use of associative cues or heuristics (such as statistical tendencies), and a slower, more effortful route similar to System 2 that produces a grammatical analysis of the input. The former would either precede the latter or the two operate in concert, yielding a rapid perception of the sentence meaning that can clash with the output of the grammar.

One complication with extending the dual-system approach to language processing is that grammatical analysis is largely a rapid, automatic and unconscious process, making its association with System 2 potentially questionable: many grammatical violations are quickly and automatically detected (e.g., Hahne & Friederici 1999, Pulvermüller et al. 2008) and there is also evidence that abstract syntactic structures can be subconsciously primed (Bock 1986; Pickering & Branigan 1999). It seems unlikely that a structured sequence of words could be understood as a list of words, even with conscious effort. These types of facts have led some psychologists to casually cite language comprehension as an example of the types of procedures in the domain of System 1 (Kahneman 2011, Osman 2004), although the analytical and rule-based nature of the computations, together with its left lateralization, could be seen as evidence for associating grammatical analysis with System 2 (e.g. Evans 2008; see Golding 1981, Deglin & Kinsbourne 1996, Goel et al. 2000, Wharton & Grafman 1998 for the association between laterality and the two systems). Ferreira et al.'s analogy depends on categorizing grammatical analysis within the domain of System 2, which in turn depends on assumptions about what underlies the difference between systems. However, this issue is complicated and controversial even in the more well-established domain of reasoning (see, for example, Gigerenzer & Regier 1996).

Shallow language processing approaches (in particular, Sanford & Sturt 2002, Sturt et al 2004) are also at least in part inspired by research on visual illusions. Many researchers working on change blindness paradigms have shown that changes to visual scenes, sometimes quite large, can pass undetected by observers (Simons & Levin 1998). As a result, some have argued that the visual system does not in fact build complete and detailed representations of scenes, contrary to conscious perception (Noë, Pessoa & Thompson 2000; Noë 2002), but rather only maintains detailed representations of e.g. what we attend to at any given time[1]; the illusion of completeness arises in the mapping to consciousness. Likewise it could be claimed that language processing does not necessarily involve complete, detailed and fully disambiguated representations. Comparative illusions in particular could be seen as a variety of change blindness, such that at each "fixation" (clause, perhaps) a detailed representation is

---

[1] Interestingly, participants apparently fixate longer on elements of the scene that change, even in cases where they failed to overtly report the change (Henderson & Hollingworth, 2003), suggesting the possibility of covert detection and therefore relatively detailed visual representations.

built without ever projecting the full, incoherent global representation; the anomaly lies within our "blind spot", the point at which the two clauses are integrated.

The existence of semantic illusions is of course highly consistent with "partial" or abandoned grammatical analysis, but other equally interesting alternatives are not ruled out. Cognitive errors are often described by heuristics when they are systematic and widespread; however, they can theoretically arise from fleeting performance issues related to lapses in attention or memory (Stein 1996), or inherent computational limitations (Cherniak 1986), among other things. Systematic biases, unlike random performance errors, would yield strong correlations in individual performance across various reasoning tasks, as well as consistently similar performance within tasks, while performance errors might be associated either with random fluctuations, or with inconsistent reactions across various tasks. Although individual performance is apparently highly correlated among different reasoning tasks (Stanovich and West 2000), there is no data yet to suggest whether this is the case for semantic illusions. And if semantic illusions are associated only with exceptionally complex constructions, then it may be possible that errors are related to computational intractability rather than systematic use of heuristics; that is, it may be unreasonable to draw strong conclusions about how semantic processing proceeds in general, given the absence of an intact grammatical parse, if it would be computationally exceptionally difficult to produce it.

In broader terms, the problem that semantic illusions pose is that they challenge a widespread assumption that semantic processing is strictly compositional – i.e., they suggest that the brain might not constrain the meaning of a sentence to that generated by the meaning of its parts and the way the grammar combines those parts ("Frege's Principle"). At one extreme end of the possible solutions to this problem is the claim that noncompositional processing is pervasive – i.e. that there is a general ability to generate sentence meanings independently from grammar, using interpretational heuristics based on probabilistic information or world knowledge. At the other end is the possible claim that there is sufficient ambiguity in the bottom-up input that the perceived meaning *is* generated compositionally, albeit in some way that formal research has yet to describe. This approach requires enriching either the meaning of the "parts" or the rules for their combination so as to generate the

perceived meaning within the grammar[2]. Any approach of this sort, of course, needs to account for any temporal instability associated with semantic illusions, a property that would seem to set them apart from other cases of grammatically-sanctioned ambiguity: given sufficient time and consideration, the perception of the sentence seems to shift dramatically, with the meaning becoming increasingly *less* clear as it is considered in greater detail.

In between these two extremes lies a range of other compelling possibilities. For example, one could claim that meaning is generated compositionally, but that the "ingredients" passed to the compositional engine are different from those strictly suggested by the input – in other words, nonveridical processing is localized to lower-level lexical processing, while the global meaning of the sentence is still derived via the grammar. Errors in lexical retrieval are already observed in cases like (6), where the word *Moses* is perceived to mean *Noah,* but there is no obvious nonveridical processing at the level of the grammar. The input may also be misperceived because of a more fundamental property of the parser, namely that it needs to keep possible alternatives under consideration in light of the general noisiness of extralinguistic processing (arising from perceptual noise, speaker errors, or uncertainty about the input). For example, sentences like (8) elicit garden-path effects as the parser is distracted by the local coherence of *the player tossed a Frisbee* (Tabor, Galantucci and Richardson 2004)*;* it is unclear why this parse is considered, given that it is inconsistent with the global grammatical analysis. Levy et al. (2009) propose that perceptual uncertainties in sentence processing can lead the parser to question whether the input was misperceived as (8) – this requires processing resources and thus can be detected as an increase in reading times or in regressions in eye movement to earlier parts of the sentence. Thus, the existence of a perceptually similar and structurally more probable alternative analysis is weighted against the compatibility of the grammatical analysis with the input words. This leads to a correct prediction that garden path effects will be attenuated for variants like (9).

---

[2] Note, for example, that although veridical properties of visual input such as luminance can be determined in a theory-neutral way, the veridicality of our perception of linguistic "illusions" is less clear given that we have no direct and objective way to access and evaluate veridical grammatical representations.

(8)   a. The coach smiled **at** the player <u>tossed</u> a Frisbee

        b. The coach smiled **as/and** the player <u>tossed</u> a Frisbee.

(9)   The coach smiled **toward** the player tossed a Frisbee.

Even if the input is processed compositionally, comprehenders might simply fail to accurately report the outcome of this process, for example failing to retrieve the correct LF at wrap-up or during offline comprehension tasks. In this case, the illusion would not index failure during compositional processing but rather the failure of other systems to encode the outcome of that process accurately – an illusion that arises at the interface between logical semantics and extralinguistic cognition. For example, it has been repeatedly shown that misanalyses of sentences like (10) linger well beyond the time when garden path reanalysis would be expected to occur, with comprehenders answering "yes" to both types of comprehension questions (Christianson et al. 2001, Kaschak & Glenberg 2004, Sturt 2007, von Gompel et al 2006). However, by looking at patterns of reflexive binding and gender mismatch Slattery et al (2013) found that a faithful syntactic structure was indeed constructed following the point of disambiguation, arguing therefore that lingering misinterpretations arise primarily when intermediate representations are not fully pruned from memory, thus interfering with responses to comprehension questions.

(10)   While Anna dressed the baby that was cute and cuddly played in the crib.
        Did Anna dress the baby?
        Did Anna dress herself?

Finally, it is entirely possible that misperceptions of illusion sentences are related specifically to a nonveridical mapping between overt syntax and semantics. In this case meaning is still derived via logical form, but it is not clear that the logical form considered would be consistent with the surface syntax of the input. This might occur if, for example, the input were "corrected" implicitly in order to reinterpret sequences containing speech errors or other abnormalities. Illusions could arise in cases where such small-scale grammatical "repair" yields a sensical interpretation of an inherently anomalous sentence (see Frazier 2014 for

detailed discussion). For example, sentences like (11)-(12) are clearly interpretable, in spite of the fact that they contain instances of syntactically unlicensed VP ellipsis. This phenomenon is sometimes considered to be a processing error (e.g., Garnham & Oakhill 1987); however, Arregui et al. (2006) propose that the input is "repaired" by making small changes to the ellipsis site (e.g., *the ICC did* ~~*be reversed*~~ → *the ICC did* ~~*reverse it*~~) using the same mechanisms that are available for the "repair" of garden-path sentences. Large repairs that involve drastic changes to the grammar are more difficult to carry out, and are associated with a corresponding cline in acceptability, as they demonstrate across several experiments. Under this approach, the parser does interpret the input veridically and compositionally, and when the outcome of that process suggests a probable speech error, the input may be directly changed in such a way as to eliminate the error, and then reinterpreted to arrive at the likely intended meaning. Because the repair operations likely involve a tradeoff between generating a well-formed interpretation, while remaining as faithful as possible to the input at the relevant level of grammar, "repair" accounts will tend to penalize any changes that are dramatically unfaithful to the grammar, predicting instead that our misperceptions of semantic illusions differ in only minor ways from the actual output of the grammar – i.e. that the illusion percept is a close perceptual, syntactic, or semantic relative of the actual input.

(11)   In March, four fireworks manufacturers asked that the decision be reversed, and on Monday the ICC did [reverse the decision].   (Dalrymple et al. 1991)

(12)   Bill$_i$ defended himself$_i$ against the accusations because his lawyer$_j$ couldn't [defend himself$_{i/j}$]   (Dalrymple et al. 1991)

To summarize, while most of the work that engages with semantic illusions (especially those addressed here) can be categorized as dual-route approaches to semantic processing, the introduction of noncompositional processing mechanisms challenges widespread assumptions about the syntax-semantics interface; for this reason, the evidence for and against these models deserves critical attention. I have argued in this section that there are a number of alternative scenarios that could feasibly allow comprehenders to consciously report interpretations that

seem not to be supported by the grammar, outlined in (13).

(13)   *Potential loci of semantic nonveridicality*

a. Dual-route semantic processing: Meaning is generated both grammar-internally and grammar-externally.

b. Grammatical ambiguity: The grammar facilitates multiple possible interpretations of the input words.

c. Nonveridicality in lexical processing: Individual lexical items are perceived in a nonveridical way, but composed in the grammar.

d. Nonveridicality in extralinguistic processing: Meaning is generated grammar-internally, but not correctly retrieved in offline processing.

e. Nonveridicality in syntax-semantics mapping: The LF of the percept is generated nonveridically from the overt syntax.

1.3   Goals of these studies

Since little is known about the illusions in (1)-(2), a logical first step is to tackle two basic questions about them: first, in broad terms, how do comprehenders react to these illusion sentences, both online and off? Prior work has begun to identify the "profile" of each respective comparative illusion, and we will continue to flesh these findings out by looking at patterns of online processing as well as overall perceived acceptability and meaning. Second, what factors modulate illusion rates in a critical way? By probing the role of various bottom-up and top-down cues, we will begin to sketch out the range of environments where the illusion persists and those where it does not, with special attention devoted to the question to how these environments relate to the grammatical versus extragrammatical properties of the illusion sentences. The answers to these questions will be used to determine what comparative illusions tell us about semantic processing, and to identify the locus of nonveridicality – are illusion sentences in fact evidence of shallow processing at the syntax-semantics interface? If so, what do they tell us about these modes of processing? If not, how else might they inform models of computations at this level of the grammar?

We begin in **Chapter 2** by surveying the literature on the grammar of comparison, in order to gain some perspective on the "veridical" representations of comparative sentences – in other words, to understand what perceptions we might *expect* to observe for illusion sentences according to the interpretational patterns widely associated with comparatives elsewhere. We then consider in detail the way one might expect these representations to be generated in real time given the assumption of incremental compositionality, and outline some of the possible complications that could arise on the basis of construction-specific properties supported in the formal literature.

In **Chapters 3-4** we take a closer look at "Escher" sentences like (14) (Montalbetti 1984), 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 (14) in spite of its global incoherence, at least initially. The illusion is sometimes thought to constitute a "blending error" that combines the two coherent phrasal templates in (14); yet no globally coherent interpretation can be derived.
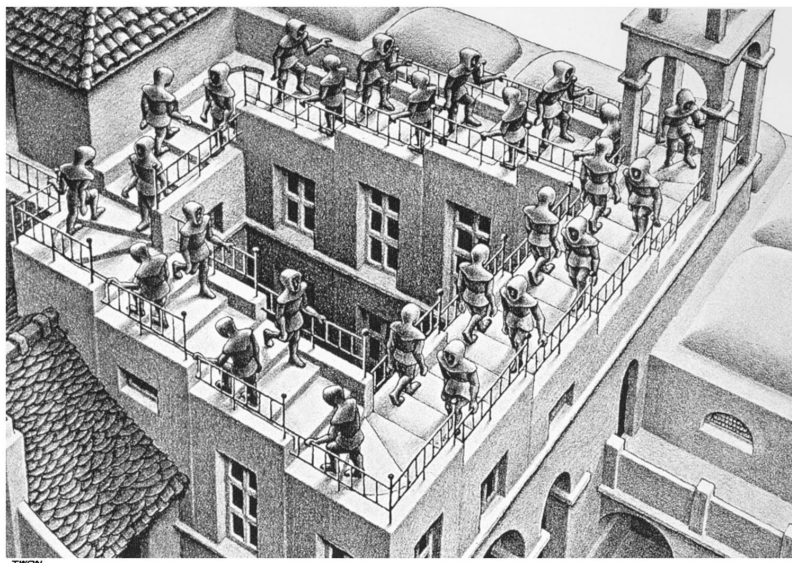
Figure 1. M.C. Escher's *Ascending and Descending,* an example of an impossible object

14

(14) a. More people have been to Berlin than I have.

  b. More people have been to Berlin [than zombies have]

  c. [People have been to Berlin more] than I have

Escher sentences were first used to argue for Townsend & Bever (2001)'s Late Assignment of Syntax Theory (LAST), which posits an initial, early phase of parsing that assigns to a string a pseudo-syntactic template (such as the NVN heuristic), using various superficial cues. Full syntactic analysis is initiated only later, and checked against the pseudosyntax for mismatch. Comparative illusions seem to provide initial support for the use of superficial "templates" in sentence processing, and Townsend & Bever (2001) have argued that the perception of grammaticality arises during an intermediate stage of parsing, where two locally coherent clausal templates have been identified but full syntactic analysis has not yet been initiated. These claims are evaluated against a competing "repair" theory that attributes the illusion more narrowly to issues arising because of the syntax and semantics of comparison and plurality (Wellwood, Pancheva, Hacquard and Phillips 2009, 2012).

For the shallow processing approach to Escher sentences, the illusion of acceptability arises when the grammatical anomaly passes by undetected, both implicitly and consciously. **Chapter 3** reports joint work with Roumyana Pancheva and Elsi Kaiser that investigates reading times associated with illusions that are consciously detected, versus those that are not consciously detected, in a way analogous to Bohan and Sanford (2008)'s investigation of reading patterns for *Moses* illusions. If the anomaly is registered implicitly by the comprehension system, this would suggest that its grammar has been processed veridically, but that something else has happened to give rise to the illusion of acceptability. In collaboration with Roumyana Pancheva, **Chapter 4** tests whether that "something" is necessarily related to a shift to a comparison of events, as suggested by Wellwood et al. (2009, 2012). The goal there will be to investigate how robust the illusion is by determining how flexibly it can be reinterpreted. We will provide some preliminary evidence that Escher sentences can be construed either as a comparison of cardinality of events *or* individuals.

From these chapters we conclude two important facts about Escher sentences. First, contrary to the shallow processing approach, the anomaly within the illusion is implicitly

detected. Comprehenders read illusion sentence more slowly irrespective of whether conscious, offline detection has occurred – in fact experiencing *more* difficulty on trials where the illusion is perceived to be especially acceptable. In other words, it seems to take discernible processing effort to arrive at a coherent interpretation of the illusion sentence. Second, contrary to the predictions of the event comparison approach, reinterpretation of illusion sentences is not strictly constrained to a shift to a comparison of events. In addition to a comparison of events, it seems possible to interpret illusion sentences as a comparison of individuals provided by either the subject or object noun phrase, so long as those elements of the sentence introduce a semantic plurality. I argue that these facts are generally most consistent with a repair approach to Escher sentences, such that the illusion arises due to implicit speech error reversal mechanisms – such mechanisms can flexibly generate various percepts by making small changes to the interpreted input.

Chapters 5-6 turn to the illusion in (15) (Wason and Reich 1979). In contrast to Escher sentences, the grammar can produce a globally coherent (if implausible) interpretation for (15)a; its logical semantics is not ill-formed, but is misinterpreted in the way shown in (15)c. The nature of the misinterpretation can be grasped more clearly through explicit comparison with sentences like (16) – which is not an illusion, but a well-formed and felicitous sentence – or (17), which contains only the local anomaly #*too trivial to ignore.*

(15)   a. No head injury is too trivial to ignore.
        b. Grammar-based meaning: #All head injuries should be ignored, #even the most trivial ones.
        c. Perceived meaning: All head injuries should be treated, even the most trivial ones.

(16)   No missile is too small to be banned.
        Grammar-based meaning: All missiles should be banned, even the smallest ones.
        Perceived meaning: All missiles should be banned, even the smallest ones.

(17)   John's head injury is too trivial to ignore.

Meaning: John's injury should not be ignored, #because of its high degree of triviality.

It is often suggested that the misinterpretation of (15) arises because of "pragmatic normalization" (Fillenbaum 1971, 1974) and that its interpretation is dependent on world knowledge. In a departure from prior work, I will use the label *inversion sentence* to describe the illusion, borrowing terminology from *depth inversion* illusions like the hollow mask illusion in Figure 2, where the hollow side of a rotated mask is perceived to be convex. This illusion, like the sentence in (15), raises serious questions about the way prior knowledge influences perception, including the extent to which object-specific knowledge or other world knowledge (such as the assumption that objects are usually convex) can supplement or override bottom-up visual information, including cues provided by shading or texture (see Hill & Johnston 2007 and references therein). In the case of the linguistic illusion, the meaning is also perceived as the opposite of its veridical meaning – i.e., it is 'inverted.'

Figure 2. In the hollow face illusion, a convex mask (upper left), when rotated so as to reveal the hollow back side (lower right), still appears to be convex.
(from http://www.richardgregory.org/papers/knowl_illusion/knowledge-in-perception.htm)

17

Chapter 5 sets forth a parallel line of inquiry for inversion sentences, outlining the possible ways that the illusion might be driven by top-down considerations versus bottom-up ambiguity. We sketch out three primary ways to understand the phenomenon of inversion sentences, which run the gamut from attributing the illusion to incomplete or "shallow" logical semantics, to incomplete or "shallow" lexical semantics, to deep processing in combination with pervasive grammatical ambiguity. Only by unpacking the possible hypotheses of the phenomenon suggested in prior work can we can begin to make concrete testable predictions.

Chapter 6 then tests each of these hypotheses in turn, by building out the profile of inversion sentences and disentangling the role of negation, plausibility, cloze bias, task type, and the possibility of grammatical dependency formation between negative elements. We conclude from these experiments that the logical force of the implicit negation in *too* can be neutralized under the negative determiner *no,* though crucially, only in environments that support negative concord or NPI-dependencies. Top-down information such as plausibility emerges as important primarily in terms of adjudicating between these two analyses. In other words, the illusion arises from a specific ambiguity associated with a specific element of the logical form, rather than a more general ability to interpret the sentence using world knowledge. We will leave open the question of whether this ambiguity arises because of specific properties of English grammar, or broader processing considerations, but provide some preliminary evidence in favor of the latter.

1.3.2   Data collection & analysis

The experiments reported here focus largely (though not exclusively) on responses from participants recruited from Mechanical Turk, an online crowdsourcing tool that facilitates access to large numbers of demographically diverse participants in an anonymized way. In addition to the practical benefits of crowdsourcing data collection, the massive size of the subject pool made available by this platform is an empirical advantage for studying phenomena such as semantic illusions, since it ensures that there is little to no overlap between participants across experiments, and accordingly avoids repeatedly exposing the same group of participants to the illusion sentences, a problematic scenario given that a fundamentally interesting property

of these sentences lies in their temporal instability. Mechanical Turk is widely used for the collection of psycholinguistic data, and has been shown to reliably elicit comparable results to more traditional laboratory research, both in terms of offline measures such as acceptability judgments, as well as more sensitive online measures, including reading times (see Gibson, Piantadosi & Fedorenko 2011, Sprouse 2011, Schnoebelen & Kuperman 2010, Keller et al 2009, Enochson & Culbertson 2015).

The data are analyzed throughout using mixed effects regression models (linear and logistic), which offer various advantages over e.g. traditional repeated measures ANOVA (see Baayen et al, 2008 for discussion). First, extraneous effects associated with variability across subjects or items, or noise due to e.g. fatigue, may be added to mixed models to increase the possibility of detecting patterns of interest in the data. The modeling of random effects for subjects and items in particular leads to conclusions about the data that are more readily generalizable across subject populations and across sentences. In addition, mixed effects models are able to deal with unbalanced designs, and their coefficients provide a basic way to understand qualitative (direction) and quantitative (effect size) properties of significant effects. The data are analyzed throughout using the lme4 package in R (version 1.1-7; Bates et al 2013) with the bobyqa optimizer to facilitate model convergence. There is no commonly accepted method for deriving p-values for mixed effects models; therefore we use a dual approach of examining the t-values for each model parameter, which can generally be treated as significant if they are under -2 or over 2 given the number of observations in these experiments (Baayen 2008; Gelman & Hill 2007), and generating more specific p-values using likelihood ratio tests that compare models with and without the effect of interest. It should be noted that the output of these two processes often, but do not always, neatly align[3] and so are generally interpreted within the context of one another.

---

[3] Likelihood ratio tests were performed as shown in (i). In cases where there was a significant X1*X2 interaction, this often gave rise to apparent significance of (ia) or (ib), likely primarily due to the main effect term accounting for some of the variability associated with the interaction term that was absent in the simpler models. These cases are interpreted with caution, especially where model output suggests that the interaction term is significantly different from zero, but the main effect term is not.

(i)   a. Likelihood ratio test, effect of X1:   Y ~ X1 + X2   Y ~ X2
        b. Likelihood ratio test, effect of X2:   Y ~ X1 + X2   Y ~ X1
        c. Likelihood ratio test, effect of X1*X2:   Y ~ X1 * X2   Y ~ X1 + X2

In the majority of experiments reported here a maximally specified random effects structure was used, per recommendations by Barr et al (2013). However, in some cases the complexity of the model resulted in serious difficulty obtaining model convergence. In most such cases, a data-driven approach was adopted, using forward model comparison to identify random slopes that significantly improved model fit (using a liberal threshold of $\alpha = .15$), against simplex models containing random intercepts only (or for interaction effects, models containing main effects only). In such cases details about the full model, including its random effects structure, can be found in the Appendix.

2   THE GRAMMAR OF COMPARISON

In order to form hypotheses about what issues comprehenders have interpreting expressions with *more* and *too,* and why these issues arise (and especially, whether or not they arise in the grammar), it is essential to have some understanding of their veridical grammar – what do these expressions actually mean, and how is that meaning built off of the syntax? Fortunately, there is a long and rich history of studying this question, which is reviewed briefly in this chapter to form a backdrop to the results and discussion on comparative illusions.

To describe the semantics of these and many other expressions, it is useful to introduce a domain of degrees into the semantic ontology (Cresswell 1976). Degrees are modeled as points (or intervals/extents; Seuren 1978, 1984, von Stechow 1984a, Löbner 1990, Kennedy 2001, Schwarzschild & Wilkinson 2002) that are totally ordered, forming a scale along some dimension, such as height or cardinality. Once the ontology contains degrees, we can then treat gradable predicates such as *tall* as relating individuals and degrees (Seuren 1973; Cresswell 1976; Hellan 1981; von Stechow 1984a; Heim 1985; Bierwisch 1989, 1991, Kennedy 1999, Heim 2000 a.o.). One standard way of implementing this intuition is shown in (18): the adjective introduces a degree argument and its lexical semantics contain a measure function that maps individuals to degrees associated with some property.[4]

(18) $[[tall]] = \lambda d \lambda x . height(x) \geq d$

The degree argument can be saturated by a family of degree morphemes, overt and covert, including measure phrases (19)a, referential degree words (19)b, degree operators (19)c, and intensifiers (19)d. Broadly speaking, these morphemes are involved in specifying the position of some degree with respect to a reference point on a scale. In cases like (19)a-b

---

[4] Not all such treatments assume the lexical semantics in (1). In particular, Kennedy (1999) treats adjectives as measure functions that map an individual to their degree along some dimension, as shown in (i) thus assigning them the type <e, d>. The degree morphology – including the comparative morpheme *–er* – then establishes some ordering between this degree and some standard of comparison.
(i) $[[tall]]_{<e, d>} = \lambda x . tall(x)$
(ii) $[[-er]]_{<ed, et>} . \lambda g_{<ed>} \lambda x . g(x) > d_{than}$

that position is defined with respect to a named degree (*five feet*) or a contextually provided one (*that tall*). When we say that someone is *taller than 5 feet,* what we mean is that the degree of height they are associated with exceeds five feet. When we say that someone is *tall* what we mean is that the degree of height they are associated with exceeds some contextual standard, presumably introduced by an unpronounced element in the adjectival projection (von Stechow 1984a, Kennedy 2007 and many others), and when someone is *very tall* they exceed that standard by a large amount (Kennedy & McNally 2005).

(19)   a. five feet tall

   b. that tall

   c. more tall

   d. ∅ tall

   e. very tall

The lexical entry in (18) projects the basic syntactic structure shown in (20), with an adjective selecting a DegP complement (in Bresnan 1973, the DegP is further embedded in a QP that hosts *much*):

(20)   AP

   DegP A
   {5'; that} *tall*
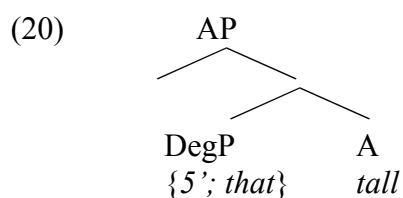
2.1   The meaning of *more/-er*

Because comparative constructions with *more/-er* are the focus of most of the work on the syntax and semantics of comparison, we begin by outlining some of the major threads in this body of work, and introduce basic assumptions about the structure and meaning of these constructions.

Comparative degree constructions such as (21)-(22) have several crucial ingredients. First, the degree morpheme itself, which is sometimes realized synthetically (*-er*) and sometimes analytically (*more*), contributes an ordering of two degrees. For example, (21) is true in case Mary's height is ordered higher than some contextually determined degree, while (22) is true in case her motivation is ordered higher than some contextually determined degree.

(21)   Mary is taller than that.

(22)   Mary is more motivated than that.

The degree morpheme, as mentioned, is often thought to fall within the same extended projection as a gradable property, which constrains the dimension of measurement and includes an order-preserving measure function (Kranz et al 1971); for a lexical entry like (18) to be coherent, where degrees are mapped to (the characteristic function of) a set of individuals, each set of individuals must have the same degree of some property; individuals who possess the property to a greater extent are related to higher degrees on a scale. In attributive comparatives like (23) the gradable property is provided by the adjective.

(23)   a. Mary is taller than that.   (dimension: height, *wealth, *intelligence)
        b. Mary is more intelligent than that.   (dimension: *height, *wealth, intelligence)
        c. Mary is more rich than that.   (dimension: *height, wealth, *intelligence)

In nominal comparatives, the degree morpheme is introduced by separate (and often covert) measure functions *much* and *many* (Bresnan 1973)*,* which constrain measurement to either cardinality (*many*) or non-cardinality (*much*) dimensions.

(24)   a. 2 liters (much) wine   (dimension: noncardinality quantity)
        b. –er much wine than that

(25) a. 2 (many) bottles   (dimension: cardinality quantity)

      b. –er many bottles than that

In such cases, measurements are derived by monotonically mapping larger sums of individuals to larger values on a scale. In order for *many* to express an order-preserving measure function mapping degrees to sets of individuals, Hackl (2001) points out that it cannot range over singular count nouns, since they denote a characteristic function of individuals which all map to the same degree, i.e. the cardinality one. *Many* is thus semantically required to range over pluralities if it is to provide a non-trivial mapping between degrees of cardinality and individuals; this predicts the unacceptability of (26)a-b.

(26) a. * many/much bottle

      b. * More student than professor was at the party.

These constraints hold of cardinality measurement across categories: within the verbal domain, cardinality comparison is also precluded with telic perfective predicates (Wellwood, Pancheva & Hacquard 2012).

(27)   * John killed the rabbit more than Mary.

Finally, the degree morpheme stands in a selectional relationship with a degree complement, such as *than that* in (28)-(30). Due to this selectional restriction the degree clause is commonly considered to be an argument of the degree morpheme (Chomsky 1965, Selkirk 1970, Bresnan 1973, Heim 2000), although it is necessarily extraposed in surface syntax (see Bhatt & Pancheva 2004 for discussion).

(28)   a. more … than/*as/*that/*to

        b. as ... as/*to/*than/*that

        c. so .. that/*as/*than/*to

(29)   a. John has **more** apples **than that.**

        b. * John has **more than that** apples.

(30)   a. John is **more** wealthy **than that.**

        b. * John is **more than that** wealthy.

Semantically, the degree clause provides a "reference point" on the scale, and is commonly called the *standard of comparison.* For example, the comparative in (30) provides information about John's wealth by indicating that it falls above the contextually indicated amount. Formally, the ordering of two amounts can be captured in terms of degree quantification: *more* denotes a *greater-than* relation between two sets of degrees, just as *every* denotes a relation between two sets of individuals (von Stechow 1984a, Heim 2000, Bhatt & Pancheva 2004 a.o.). A common approach is to compare the maximal elements of the two sets, as shown in (31), although other options have also been proposed (e.g., (32)).

(31)   [[more]] = λDλD'. max(D') > max(D)   (Heim 2000)

(32)   [[more]] = λDλD'. ∃d . D'(d) & ¬D(d)   (Seuren 1973, 1984; Schwarzschild 2008)

The first set of degrees corresponds to the standard of comparison, provided by the extraposed degree clause. Measure expressions like (*than*) *five / (than) five feet* / *(than) that* may themselves denote degrees or sets of degrees, making the internal composition of the degree complement potentially simple in such cases, perhaps phrasal. In *clausal comparatives* like (33), by contrast, the degree quantifier appears to merge with a more complex complement containing elided clausal material.

(33)   Shelly is taller than Bill is ~~tall~~.

Examples like (34) show that the clausal material in the *than*-clause is subject to island effects (Chomsky 1977), suggesting that its underlying structure is similar to that of a *wh*-question or relative clause with an unpronounced *wh*-operator and elided predicate. This makes clausal comparatives underlyingly analogous to instances of *comparative subdeletion* (Bresnan 1973) like (35), except that the gradable predicate in the *than*-clause is unpronounced. In clausal comparatives, this predicate is obligatorily elided under identity with the matrix, but otherwise the structure of the *than*-clause is unexceptional.

(34)   a. * Shelly is taller than I claimed that Bill is.
         b. * Shelly is taller than I wonder whether Bill is.

(35)   The table is longer than wh_1 the door is d_1-wide.

Crucially, *wh*-movement within the degree clause can produce a degree predicate of type $\langle d, t \rangle$, which will make it a suitable first argument for *more*:

(36) $[[\text{more}]](\lambda d.\ \text{Bill is d-tall}) = \lambda D.\ \max(D) > \max(\lambda d.\ \text{Bill is d-tall})$

Also importantly, this provides an explanation for the ill-formedness of comparatives like (37)a, which have the same infelicious clausal structure as (37)b (see Rullmann 1995).

(37)   a. * Shelly is taller than Bill isn't.
         b. * How tall isn't Bill?

        After the degree morpheme combines with the degree clause, the resulting degree phrase, a generalized degree quantifier of type $\langle \langle d, t \rangle, t \rangle$, cannot saturate the degree argument

of the adjective or measure function. This type mismatch can be resolved through clausal QR, leaving behind a variable of type *d* to merge with the adjective (Heim 2000):

(38)   Shelly is taller than Bill is.

```
                              TP
                 ┌────────────┴────────────┐
         DegP₂ <<d,t>, t>                   ·
        ┌──────┴──────┐              λd₂         TP
   Deg <dt,<dt,t>>  PP<d, t>                ┌──────┴──────┐
      -er         ┌────┴────┐           Shelly        ·
                than     CP <d, t>              is        AP <e, t>
                      ┌────┴────┐                      ┌────┴────┐
                    λd₁       TP                      d₂       Adj
                          ┌────┴────┐                       tall <d, <e, t>>
                        Bill      AP <e, t>
                              ┌────┴────┐
                            d₁        Adj
                                    tall <d, <e, t>>
```
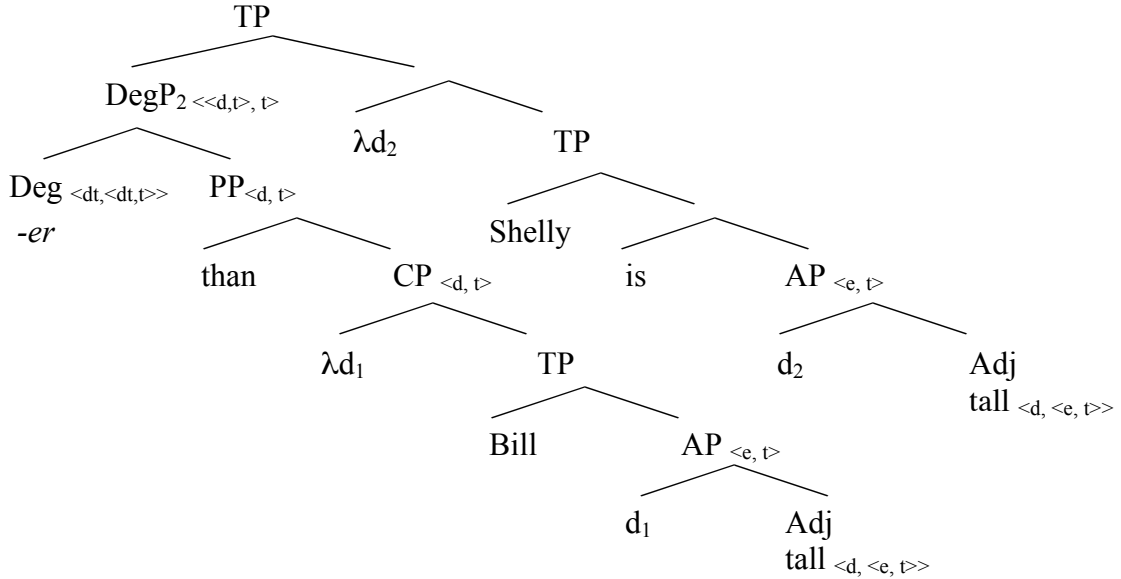
Heim (2000) motivates this QR with several considerations. First, the comparative operator participates in certain limited scope interactions. As expected on a quantificational analysis, scoping above or below an intensional verb yields two distinct readings for (39). When the degree quantifier takes narrow scope, every world compatible with the requirements is such that the paper is less than ten pages: possibly 5 or 9, but never 11 or more (the *exactly* reading). When the degree quantifier takes wide scope, as in (39)b, the sentence specifies that the *minimal* length of the paper in the acceptable worlds is less than ten pages (the *at least* reading).

(39)   (The draft is 10 pages.) The paper is required to be less long than that.
         a. required > -er: $\forall w \in$ Acc: max {d: long$_w$(p, d)} < 10 pages
             "It is required that the paper be shorter than 10 pages"
         b. –er > required: max {d: $\forall w \in$ Acc: long$_w$(p, d)} < 10 pages
             "The paper is required to be at least 10 pages"

Second, without QR, there will be infinite regress within the ellipsis site as the matrix predicate *is taller than ...* is itself interpreted in the ellipsis site of the *than*-clause, (40). The same problem occurs with other quantificational expressions – see (41) – and the movement of the quantificational expression at LF is the usual solution to such cases of antecedent-contained deletion (ACD): this movement leaves behind a variable, allowing the VP to be reconstructed unproblematically (Sag 1976; Williams 1977). In comparatives, the analogous movement of the degree operator will leave a variable of type *d*; this material can then be reconstructed in the ellipsis site unproblematically (Wold 1995).

(40)   a. Shelly <u>is taller</u> than Bill is [~~taller than Bill is~~ [~~taller than Bill is~~ […]]]
        b. [more than Bill is ~~x-tall~~] Shelly is x-tall

(41)   a. I met every girl you did [~~meet every girl you did~~ [~~meet every girl you did~~ […]]]
        b. [every girl you did ~~meet x~~] I met x

This type of account can be extended to explain the so-called Sag-Williams Ellipsis-Scope generalization, namely the observation that only the material within the scope of the DegP at LF may be reconstructed into the ellipsis site. The pattern of judgments in (42), for example, indicates that the degree operator *–er* and the elided material within the *than*-clause need to c-command the antecedent *tell her to work d-hard,* which itself conflicts with the TELL > ER scope configuration. This well-established correlation has been used to argue for a crucial relationship between the position of the *than*-clause and the way its contents are reconstructed.

(42)   Mary's father tells her$_i$ to work harder than her$_i$ boss does Δ.

       a. TELL > -ER, elided material = work d-hard

          ≈ Mary's father tells her: work harder than your boss works.

       b. *TELL > -ER, elided material = tell her to work d-hard

          ≈ Mary's father tells her: work harder than your boss tells you to work.

       c. -ER > TELL, elided material = work d-hard

          ≈ Mary's father tells her: work d1-hard; Mary's boss works d2-hard; d1 > d2.

       d. -ER > TELL, elided material = tell her to work d-hard

          ≈ Mary's father tells her: work d1-hard; Mary's boss tells Mary: work d2-hard; d1 > d2.

The quantificational analysis of comparatives is not universally accepted (see Kennedy 1999 for an alternative non-quantificational approach); for example, a well-known problem with it is that the comparative operator participates in only a limited range of scope interactions, and in particular is constrained in its interaction with individual quantifiers (Kennedy 1999, Heim 2000). Beck (2012) argues that what observable scope ambiguities exist can be reduced to movement of other elements in the DegP, such as the differential (e.g., **exactly two *more***), independent from the comparative morpheme. However, my discussion of comparative illusions will draw heavily on the quantificational analysis in part due to the empirical support for it observed in patterns of online processing described below.

In earlier work, Hackl et al. (2012) elicited processing patterns argued to be consistent with quantificational movement. Such movement is needed to resolve type mismatch in quantificational NPs in object position – which, as type $<<e, t>, t>$ expressions, are not suitable arguments for a type $<e, t>$ verbal predicate. As discussed above, this movement is also independently justified any time there is an ACD site, in order to avoid infinite regress. Combining these two facts, we might expect the upstream appearance of a quantificational NP to ease resolution of ACD downstream: the mechanism that resolves ACD has already been pre-emptively posited as soon as the comprehender reaches the quantificational noun phrase. Without this predictive processing, the comprehender will experience more difficulty at the ACD site. In a series of self-paced reading experiments Hackl et al (2012) found that ACD was easier to resolve when local QR was predictable due to the presence of the quantifier *every.*

Non-local QR – needed to resolve a larger ellipsis site – remained difficult in such cases, however, suggesting a garden-path type effect arising due to preference for local QR (though see Szabolcsi 2014, Jacobson & Gibson 2014 for competing accounts of the facts from this experiment).

(43)   The doctor was reluctant to treat …
       a. … the patient that the recently hired nurse admitted.   (no QP; no ACD)
       b. … the patient that the recently hired nurse did.   (no QP; local ACD)
       c. … the patient that the recently hired nurse was.   (no QP; nonlocal ACD)
       d. … every patient that the recently hired nurse admitted. (QP; non ACD)
       e. … every patient that the recently hired nurse did.   (QP; local ACD)
       f. … every patient that the recently hired nurse was.   (QP; nonlocal ACD)

In an extension of this work on quantificational NPs, Breakstone et al (2011) use a similar paradigm to investigate QR in comparatives. Using contexts strongly biased towards an *at least* interpretation, such as (44), they compared reaction times to comparatives containing two types of differential phrases: those that need to take wide scope to obtain the *at least* reading, and those for which the narrow scope reading can yield an *at least* reading (e.g., *a few*).

(44)   In order to become the all-time champion, John was required to win exactly 3 more races than Bill.
       EXACTLY 3 MORE > REQUIRED: *at least* 3 races
       # REQUIRED > EXACTLY 3 MORE: *exactly* 3 races

Pre-emptive QR of *exactly n more* above the modal verb *required* will create a configuration that resolves long distance ACD. This predicts that the processing advantage for local ACD resolution found by Hackl et al (2012) should be mitigated in (45)b, but not in (45)d, results that are borne out in the reading times several words downstream from the ellipsis site.

(45)   In order to become the all-time champion, the American athlete was required to

        win…

        a. MORE > REQUIRED only; local ACD:

            … **exactly three** more matches than the British athlete **did**

        b. MORE > REQUIRED only; nonlocal ACD:

            … **exactly three** more matches than the British athlete **was**

        c. REQUIRED > MORE OK; local ACD:

            … **a few** more matches than the British athlete **did**

        d. REQUIRED > MORE OK; nonlocal ACD:

            … **a few** more matches than the British athlete **was**

Thus, although one can derive scope ambiguities by moving the differential phrase alone, the online processing patterns associated with comparatives support a quantificational treatment of *more* with quantifier raising that is independent of the quantificational properties of the differential (contra Beck 2012), and that this movement happens predictively. For the purposes of this dissertation, it is also interesting to note the depth of processing that these results imply, especially in light of the difficulty many people report in trying to understand complex scope configurations involving comparative operators and modal quantifiers. Comprehenders appear to be able to use the context of the sentence to immediately generate predictions about possible scope configurations, and then check these predictions again at the ellipsis site.

The question of whether the QR approach should be extended to phrasal comparatives such as (46), where there is no overt evidence of ellipsis, is an area of active research (Hankamer 1973; Heim 1985; Kennedy 1999; Lechner 2001, 2004; Pancheva 2009; Bhatt & Takahashi 2007, 2011 a.o.). If we wish to simplify the syntax of (46) so that *than* merges directly with a DP *Canadians* (see Heim 1985 for discussion of the so-called "Direct analysis" of comparatives, and Bhatt & Takahashi 2007 for a more recent implementation), then we must also posit another meaning for *more* that is compatible with this configuration. For example, one contender in (47) requires movement of *more than Canadians* to a type <d, <e, t>> node that can be applied to the two DPs under comparison.

(46)   I ate more doughnuts than John.

(47)   [[more$_{PHRASAL}$]] = λxλP$_{<d,et>}$. λy . ∃d[P(y,d) & ¬P(x, d)]

Since we focus primarily on the processing of clausal comparatives in this dissertation, we will largely set aside questions about the semantics of phrasal comparison (though see discussion in section 2.3).

2.2   The meaning of *too & enough*

Expressions with *too* and *enough* are very similar to comparatives with *more* in many respects, and in fact are found in largely the same environments, expressing, for example, an ordering in the verbal, adjectival or nominal domains, as illustrated in (48)-(50):

(48)   a. too old   (adjectival)
        b. old enough
        c. older

(49)   a. too many apples   (nominal)
        b. enough apples
        c. more apples

(50)   a. ran too much/too often   (verbal)
        b. ran enough/often enough
        c. ran more/more often

Although there are some syntactic differences in the internal composition of the extended adjectival projections containing *too, enough,* and *more,* especially with respect to the (non-) inclusion of m-words like *much* (see Corver 1997 for discussion), semantically they

behave very similarly. In fact, it is often noted that sentences with *too* and *enough* can be paraphrased using *–er* (Meier 2003, von Stechow, Krasikova & Penka 2004):

(51)   a. John is too tall.

b. John is taller than it is acceptable for him to be tall.

(52)   a. John is tall enough.

b. John is as tall as it is required for him to be tall.

Like comparatives, expressions with *too/enough* also denote an ordering among two degrees or two sets of degrees, with the gradable predicate providing the property to be measured. Also like comparatives, they select a degree clause that appears overtly extraposed in the surface syntax; the degree clause again provides a standard for comparison. And finally, similar to comparatives, some aspect of the matrix has to be reconstructed into the degree clause at LF, albeit not through the mechanism of syntactic ellipsis. In the examples in (53) *John's height* is represented twice: once in the matrix clause, and once in a modalized context introduced by *too* and its complement clause. Although the complement clauses do not contain an ACD site to motivate QR, such movement may be independently motivated given that *too* and *enough* expressions participate in similar scope ambiguities as comparatives, with intensional verbs scoping either over or under the degree operator (Heim 2000; Meier 2003):

(53)   John needs to have too much money.   (from Heim 2000)

a. NEEDS > TOO: $\forall w \in Acc_{need}$: max {d: John has d-much money in w} >

max {d: $\exists w' \in Acc_{too}(w)$: John has d-much money in w'}

"What John needs is to have too much money"

b. TOO > NEEDS: max {d: $\forall w \in Acc_{need}$: John has d-much money in w} >

max {d: $\exists w' \in Acc_{too}(w)$: John has d-much money in w'}

"John's financial needs are too high"

As a result, it is often assumed, implicitly or explicitly, that the same basic analysis used for *more* can be extended to *too* and *enough* (Heim 2000; von Stechow, Krasikova & Penka 2004; Meier 2003). As the comparative equivalents in (51)-(52) make clear, however, expressions with *too* and *enough* have an added layer of complexity in their semantics, since they introduce implicit modal force: John's height is greater than it *should* be, or greater than it *is allowed to* be, or greater than it is *required* to be. Meier (2003) observes that this modality may be fixed explicitly without significantly affecting the interpretation of the sentence, (54)-(55):

(54)   Bertha is old enough to drive. ~ Bertha is old enough to be able to drive.

(55)   Bertha is too young to drive. ~ Bertha is too young to be able to drive.

The internal composition of the complement clause is also somewhat different from comparatives. First, most analyses assume the *to*-clause is a sentential complement rather than a degree clause. This is because the complement does not appear to contain an elided gradable predicate, as with *than*-clauses. Instead, the lexical entry of the degree quantifier stipulates that the predicate from the matrix should be interpreted twice: John's age should be assessed in the actual world and compared with his age in worlds where the sentential complement is true; this treatment is in some ways analogous to the direct analysis of comparatives, where the gradable predicate is not supplied through syntactic ellipsis but rather is used twice in the lexical entry of – *er*. Although the DegP (*too/enough* and its complement) are taken to denote a type <<d, t>, t> generalized degree quantifier that moves for reasons of interpretability, the degree quantifier itself only takes one set of degrees as an argument, not two; the second is generated lexically using the material from the sentential complement. The nonfinite complement clause typically includes a null proform, and may additionally contain a gap (which, for reasons that are somewhat unclear, appears to affect the possible landing sites of the degree quantifier; see Nissenbaum & Schwarz 2010).

(56)   John is too young [PRO to invite _]

(57)   John is too young [PRO to invite him]

Too and enough generally can be regarded as logical duals (Meier 2003, von Stechow et al 2004). Duality relations are typically defined using the notions *external negation* (broadly corresponding to sentential negation) and *internal negation* (broadly corresponding to constituent negation). External negation is applied to the entire statement with the quantifier, whereas internal negation is applied to the quantifier's scope. Two quantifiers may be considered duals when the types of equivalences shown in (58) hold. As (59)-(60) illustrate, logical duality is a common property of existential versus universal individual (*some/every*) and modal (*can/must*) quantifiers.

(58)　a.　$\exists x: P(x)$　　$\equiv$　　$\neg \forall x: \neg P(x)$

　　　b.　$\forall x: P(x)$　　$\equiv$　　$\neg \exists x: \neg P(x)$

　　　c.　$\neg \exists x: P(x)$　　$\equiv$　　$\forall x: \neg P(x)$

　　　b.　$\neg \forall x: P(x)$　　$\equiv$　　$\exists x: \neg P(x)$

(59)　　a. there is some suspect who is guilty $\equiv$ not every suspect is not guilty/innocent

　　　　b. every suspect is guilty $\equiv$ there is no suspect who is not guilty/innocent

　　　　c. there is no suspect who is guilty $\equiv$ every suspect is not guilty/innocent

　　　　d. not every suspect is guilty $\equiv$ there is some suspect who is not guilty/innocent

(60)　　a. John can sleep $\equiv$ it's not the case that John must stay awake

　　　　b. John must sleep $\equiv$ it's not the case that John can stay awake

　　　　c. It's not the case that John can sleep $\equiv$ John must stay awake

　　　　d. It's not the case that John must sleep $\equiv$ John can stay awake

(61)   a. John is too old ≡ it's not the case that John is young enough.

        b. John is old enough ≡ it's not the case that John is too young

        c. John is not too old ≡ John is young enough

        d. John is not old enough ≡ John is too young

An intuitive way to characterize the difference in meaning between *too* and *enough* is in terms of the nature of the standard of comparison: expressions with *too,* such as (62), invoke maximal allowable degrees of age, whereas those with *enough,* such as (63) invoke minimal allowable degrees of age. Meier (2003), proposing the semantic treatment outlined in (62)-(63), shows that this meaning – together with the assumption that polar adjective pairs such as *old/young* project complementary extents on the same scale (von Stechow 1984b) – is sufficient to generate the equivalences above.

(62)   The food is too good to (be able to) throw it away.
The value d such that the food is d-good is <u>greater than or equal to</u> the <u>maximum</u> of all values d', such that if the food is d'-good, one is <u>able to</u> throw it away.

(63)   Bertha is old enough to (be able to) drive a car.
The value d such that Bertha is d-old <u>is greater than or equal to</u> the <u>minimum</u> of all values d', such that if Bertha is d'-old, she is <u>able to</u> drive a car.

Minimal and maximal degree thresholds can be obtained two ways. The first is to explicitly propose a minimality operator alongside the usual maximality operator, the approach used in Meier (2003). The minimality operator is then invoked inside of the lexical entry of *enough,* while the maximality operator is invoked in the lexical entry of *too.* The motivation for stipulating a minimality operator, however, is unclear as it would have seemingly limited use, in contrast to the widespread use of maximality across a range of grammatical construction (including definite descriptions).

(64)   min(D) = ιd[d ∈ D & ∀d' ∈ D ⇒ d' ≥ d]

(65)   max(D) = ιd[d ∈ D & ∀d' ∈ D ⇒ d ≥ d']

(66)   a.   [[too]] = f : D$_{<s, <<s, pt>, <dp, t>>, t>}$ (Meier, 2003)
              For all w ∈ W, Q ∈ D$_{<s, pt>}$ and P ∈ D$_{<d, p>}$:
              f(w)(Q)(P) = 1 iff max(λe . P(e)(w)) > **max**(λe*. Q(w)(P(e*)))
        b.   [[enough]] = f(w)(Q)(P) = 1 iff max(λe . P(e)(w)) > **min**(λe*. Q(w)(P(e*)))

Minimal thresholds can also be obtained naturally by embedding a universal quantifier under the maximality operator. In that case, the different standards of comparison for *too/enough* are given by the strength of the modal quantifier (von Stechow et al 2004). In (67) the only age degree associated with John in every world where C holds is d_1, the minimal value of the set. By contrast, all of {d_1, d_2, d_3} are degrees such that John is that tall in some world where C holds. This approach further strengthens parallels between the meaning of *more* and *too/enough,* the latter essentially reducing to the meaning of *more* with modal quantification.

(67)   John is too old/old enough (to C).
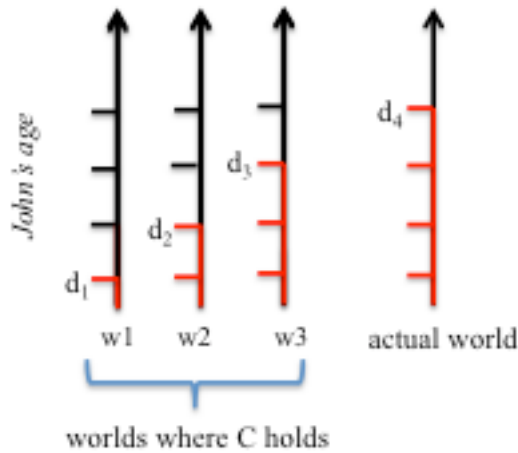
(68) a. $[[too]]^w = \lambda C_{<s, <s, t>}$. $\lambda D_{<s, <d, t>}$. $\max\{d: D(w)(d)\} >$

      $\max\{d: \exists w': w' \in Acc(w): D(w')(d)\ \&\ C(w')\}$

    b. $[[John\ is\ too\ old\ (to\ C)]] = 1$ iff $\max\{d: old(John) \geq d)\} >$

      $\max\{d: \exists w' \in Acc(w): old(John) \geq d)$ in $w'\ \&\ C(w')\}$

      $(= 1$ in (67) iff $d_4 > d_3)$

(69) a. $[[enough]]^w = \lambda C_{<s, <s, t>}$. $\lambda D_{<s, <d, t>}$. $\max\{d: D(w)(d)\} >$

      $\max\{d: \forall w': w' \in Acc(w)\ \&\ C(w') \rightarrow D(w')(d)\}$

    b. $[[John\ is\ old\ enough\ (to\ C)]] = 1$ iff $\max\{d: old(John) \geq d\} >$

      $\max\{d: \forall w \in Acc(w)\ \&\ C(w') \rightarrow old(John) \geq d\}$

      $(= 1$ in (67) iff $d_4 > d_1)$

Note that it is necessary on this approach that the modal quantifier *not* be a scope-bearing element that is syntactically independent from the degree quantifier; von Stechow et al (2004) introduce the quantificational force within the lexical entry of *too,* for example. This is because it might otherwise be expected to participate in scopal ambiguities with the degree operator, causing *enough* readings to become *too* readings when the universal modal quantifier scopes over the degree operator, as shown in (70):

(70) $\forall w': w' \in C(w) \rightarrow \max\{d: John\ is\ d\text{-}old\ in\ w\} > \max\{d: John\ is\ d\text{-}old\ in\ w'\}$

    $(= 1$ iff $d_4 > d_3)$

Although no formal analysis has explicitly addressed this point, two intuitions about the meaning of *too* and *enough* will be relevant to our discussion of inversion sentences and therefore should be clarified and discussed. First, the choice of degree quantifier affects the relationship between the sentential complement and the gradable predicate in a particular way, setting up what we will term the "internal meaning" of *too* and *enough* expressions. As Wason & Reich (1979) originally noted, there is an intuition that when we say e.g. *it is too sunny to rain,* we presuppose that as the degree of sunniness increases, the likelihood of raining

decreases; and likewise, when we say that *it is cloudy enough to rain,* we presuppose that as the degree of cloudiness increases, the likelihood of raining increases. Wason & Reich term this a presupposition, probably because it survives in the usual environments – under negation (72), in questions (73), and in conditionals, (74). In other words, it does not affect the overall conclusion about the possibility of rain.

(71)   It's cloudy enough to rain.

(72)   It isn't cloudy enough to rain.

(73)   Is it cloudy enough to rain?

(74)   If it's cloudy enough to rain, I will bring an umbrella.

In actuality, it is unclear whether there needs to be a continuous, order-preserving mapping between e.g. degrees of cloudiness and probability of rain to capture this intuition. For example, in the context of an eye exam, suppose there are letters on a screen that one can only accurately name when they have 20/20 vision. In this context, someone with 20/100 vision would be no more likely to be able to name the letters than someone with 20/200 vision; similarly, someone with exceptional 20/10 vision is no more likely to be able to name the letters than someone with normal 20/20 vision. Yet, (75)-(76) can be uttered in this context, suggesting that the intuition about the "internal" meaning of *too* and *enough* is related to the existence of the *minimum* and *maximum* thresholds posited in the analyses above, rather than a continuous relationship between degrees and probabilities of naming letters accurately. This threshold is made available in both of the analyses of Meier (2003) and von Stechow et al (2004) – either by explicitly positing a minimality operator or through the force of the modal quantifier.

(75)   Mary's vision is good enough to name the smallest letters on the screen.

(76)   Mary's vision is too poor to name the smallest letters on the screen.

A second component of *too* and *enough* sentences concerns its *injunction.* This portion of its meaning relates to the conclusion about what can or should be the case: in (75) we conclude that Mary *can* name the letters; in (76) we conclude that Mary *cannot* name the letters (whether she actually *does* name the letters is another story related to the implicatures generated by *too* and *enough* with perspective aspect; see e.g. Hacquard 2005). This portion of the meaning arises because of the way the minimum and maximum thresholds carve up the relevant scale: if Mary's vision in the actual world falls within the interval containing her vision in all of the letter-naming worlds, then she *can* name the letters; if it does not fall within that interval, then she *cannot* name them. Thus, although the *too* and *enough* sentences simply assert an ordering of degrees, this ordering generates an entailment about what can or should be the case in the actual world.

To summarize, in these sections we saw that constructions with *more, too* and *enough* involve the comparison of two extents associated with some gradable property. Formally this comparison can be modeled as degree quantification, with hypothesized syntactic and semantic parallels to individual quantification. The first set of degrees in provided by the clausal complement of the degree operator, which is obligatorily extraposed; the second set of degrees is derived through covert movement of the degree quantifier, creating a predicate of degrees in the matrix clause. Comparative morphemes express an ordering on these two sets of degrees. Comparatives expressing sufficiency (*enough*) or excess (*too*) are usually thought to contain the same basic ingredients as those with *more,* except with additional, implicit modality and a slightly different internal structure of the degree clause.

2.3 Processing the LF of comparison

One of the reasons why there has historically been so much interest in the semantics of comparison is that these constructions pose interesting challenges for the syntax-semantics interface: in order to derive correct interpretations and account for the facts involving scope ambiguity, extraction, and ellipsis resolution, it becomes necessary to posit a somewhat complex mapping between the overt syntax and the interpreted logical form, including multiple sequences of movement and a range of covert semantic material. The prior sections have

highlighted the fact that the properties of this mapping are empirically motivated by structural and semantic considerations, although they make the proposed grammar of degree quantification admittedly complex, particularly in light of the complications posed for a left-to-right parser. In this section I briefly discuss the complications inherent in implementing the LF of comparison compositionally during incremental processing (see also Grant 2013 for work that begins to look at this question), which may turn out to be relevant to the existence of comparative illusions.

First, consider the sequence of actions that must be taken to interpret a *more* comparative like (77) given relatively standard assumptions about its veridical meaning:

(77) More Americans have been to Berlin than …

At *more Americans,* the parser has encountered a degree operator that must be covertly moved. The results of Breakstone et al. (2011) suggest that local QR is predictively initiated: a gap site is posited before *Americans* and a local QR site for *more* located. However, given that the first argument of *more* – provided by the extraposed *than*-clause at the right edge of the sentence – cannot yet be integrated, it is unclear how much interpretation can take place at this point. By the end of the matrix clause *More people have been to Berlin,* the second argument of *more* can be fully constructed as λd . *d-many people have been to Berlin*. The contents of the *than*-clause at this point are unknown, other than the fact that it should denote a cardinality of a set of individuals – though this may be realized with any range of possible continuations, as shown in (78). The details of the filler-gap dependency between the moved quantifier and the gap site *d-many people* accordingly cannot be fully resolved until the *than*-clause is reconstructed.

(78) a. … than Canadians (have)
     b. … than I expected
     c. … than to Toyko

Recall that a major topic of ongoing research concerns whether the semantics of *more* are uniform across the diverse range of constructions in which it appears. Positing a single,

uniform meaning for *more* also necessitates invisible clausal structure in (79), shown in italics, for the *than*-clause to be interpretable. Otherwise, dual analyses of *more* would be required: one that combines with a clausal complement and another "direct analysis" that merges *more* with an individual-denoting noun phrase.

(79)   More Americans have been to Berlin than *wh d-many* Canadians *have been to Berlin*.

The possibility of a direct analysis of –er would put the parser in an even worse position due to the additional lexical ambiguity of the comparative morpheme. In this case, at *more Americans* we can no longer predict almost anything about the LF of the sentence – including the meaning of *more* itself – until the filler site, the content of the *than*-phrase, is processed. The parser does not know which meaning of *more* to retrieve, or what node movement should target (if it should be moved at all, given that phrasal *more* posited by Kennedy 1999 is interpreted in situ). Because the meaning of the matrix clause *More Americans have been to Berlin* is dependent on the meaning of the *than*-phrase in a critical way, composition would need to be delayed until the content of the *than*-phrase is encountered – at which point a meaning for *more* can be selected and a derivation initiated. In other words, at minimum, the meaning hypothesized by the end of the matrix clause can only be, at best, extremely tentative.

Note that the meaning of the ellipsis site in the *than*-clause is also dependent on the meaning of the matrix in an equally critical way. In order to reconstruct the *than*-clause, particularly when potentially ambiguous, details of the matrix clause must be inspected, including its focus structure and any relevant parallelisms between remnant and associate (Carlson 2001).

(80)   Tasha called Bella more often than the doctor.   (Carlson 2001, ex. (10))
        a. … than the doctor called Bella
        b. … than Tasha called the doctor

A sufficiently well-articulated matrix syntax must also be accessed in order to locate a suitable host for the degree variable, thus properly reconstructing the silent material within the *than*-clause of e.g. (81)a as shown in (81)b and not (81)c.

(81)   a. More Americans like cookies than Canadians like donuts.
       b. more than wh$_1$ d_1-*many* Canadians like donuts
       c. more than wh$_1$ Canadians like d_1-*many* donuts

Putting all of these facts together we find that comparatives raise the following problem for the parser: throughout the content of the matrix clause the parser knows relatively little about the LF of the sentence, given the discontinuous relationship between *more* and its first argument. The details of sentence meaning are determined in a crucial way by the element encountered last in the sentence, and this in turn poses a potential problem for a parser that composes meaning incrementally from left to right.

With respect to *too* and *enough,* the situation is in some ways simpler, and in some ways more complex. Like *more, too* and *enough* have an extraposed first argument and potential scopal ambiguity; however, the complexities raised above should be mitigated somewhat by the lack of syntactic ellipsis. However, it is easy to see from section 2.2 that the semantics of *too* and *enough* are also quite complex. Because the standard of comparison is generated without ellipsis, its meaning is correspondingly more dependent on the correct reconstruction of a number of implicit interpretational components which are not represented in the overt syntax. The gradable predicate needs to be represented twice in the semantics, and a maximal or minimal value needs to be obtained through a silent modal quantifier. Like *more*-comparatives, it is reasonable to assume further complications associated with scope interactions with other overt quantifiers in the clause.

The complexity of the syntax-semantics mapping is relevant to several of the accounts of comparative illusions. One possibility is that interpretational problems arise when the LF is generated out of the wrong ingredients: perhaps the representational instability of the matrix in *more*-comparatives, or the implicit semantic components in *too/enough*-comparatives, yields special susceptibility to certain types of errors, thus generating the wrong truth conditions. This account predicts a relatively constrained distribution of errors that are intimately related to the

nature of the grammatical derivations themselves, and, potentially, more or less independent of top-down considerations. Another possibility is that there is deliberate repair initiated by the parser in response to the illusion, but because the parser is likely to be more uncertain about the original parse to begin with, the processing difficulty associated with reinterpretation is relatively subtle, at least enough so that the comprehender is consciously unaware that it has occurred.

3   EVIDENCE FOR ONLINE REPAIR OF ESCHER SENTENCES

3.1   Introduction

Escher sentences were first documented by Montalbetti (1984) – who in turn attributes the famous sentence in (1) to Hermann Schulze – but they have received little attention in the formal or experimental literature until recently. The grammatical problem with the famous sentence in (1) is often initially quite difficult for English speakers to detect. The logical form of the main clause, constructed by the rules of syntax and logical semantics, should require comparison of cardinalities of sets of individuals. The *than*-clause therefore needs to contain a bare plural noun phrase, from which degree abstraction will be possible (Chomsky 1977, Heim 2000, and others). (1) fails to 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.

(1)   a.   More people have been to Berlin than I have.

   b.   More **λd . d-many people** have been to Berlin

   than **wh λd . d-many (*I)** have been to Berlin

A grammatical continuation of the first clause – as shown in (2) – will contain a bare plural noun phrase, either overt or covert, such as *people* in (2). This is required by the determiner *more* and its incorporated *many,* a gradable determiner incorporating a measure function, whose semantics in (3) requires a semantically plural NP for an orderly, non-trivial mapping of individual sums to degrees of cardinality (Hackl 2001; see section 2.1). The counterpart to *more* in the *than*-clause – the unpronounced wh-operator *wh-many* – imposes similar restrictions on the *than*-clause subject, so that across both clauses, the first argument to *many* is required to be a bare plural noun phrase.

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

Meanwhile, the second clause of the illusion, *than I have,* is an acceptable continuation for a comparative of another type, such as (4). The degree variable is hosted by an event measure function, and therefore the singular noun phrase poses no problem. 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**

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, making this phenomenon a potentially good example of shallow processing. Along these lines, Townsend & Bever (2001) have suggested that the impression of acceptability arises because the sentence 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 later asked to explain what it means.

However, in a series of offline studies, Wellwood, Pancheva, Hacquard, & 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*) are consistently rated higher than those containing non-repeatable predicates (e.g. *graduated from high school*), see (5).

Pragmatically a comparison of events is felicitous only in the former case; in the latter case, the semantics itself can support a comparison of events, for example by constructing a complex scenario involving multiple high school degrees, but this meaning is not very plausible. Because comprehenders show sensitivity to predicates that are pragmatically repeatable versus those that are not, they infer that speakers extract a comparison of events from the illusion, and that the availability of this interpretation is affected by grammatical constraints requiring the cardinality measure function to combine with a semantic plurality, in this case a plural VP (see e.g., Kratzer 2005 for discussion about verbal plurality).

(5)   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 (6) (Krifka 1990, Moore 1994, Doetjes and Honcoop 1997, Barker 1999). Although the determiner *4000* in (6) most saliently counts individuals – yielding a one-to-one pairing of ships and lock-passings – an alternative reading involves 4000 events of lock-passing involving possibly fewer than 4000 ship entities. To obtain these readings, Krifka (1990) posits the null determiner in (7), 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 (8).

(6)   4000 ships passed through the lock.

(7)   a. $[[D_\varnothing]] = \lambda P_{<e,t>} . \lambda R_{<e,<v,t>>} . \lambda e_v . \text{OEMR} (R)(e)(P)$

      b.   OEMR (R) is the smallest relation between events & quantity predicates, such that for any event *e* and quantity predicates P and Q:

         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 $\exists x . P(x)$ & $R(e)(x)$.

         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 $(R)(e_1 + e_2)(P+Q)$

(8)   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 *n*=4000

Krifka's analysis can be extended to comparatives, as he himself notes, and thus to the main clause in the illusion sentence. The *than*-clause will still be ill-formed, however, since the quantity *wh*-determiner, needed to create a degree predicate, 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 in the *than*-clause. The illusion sentence remains ungrammatical even under a Krifka-style event-related interpretation, as illustrated in (9) (with the problem highlighted in bold).

(9)   more $\lambda d . [\exists e . \text{OEMR} ([[\text{went to Russia}]])(e)(\lambda x . \text{people}(x)=d)]$

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

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[5]), may be combined with an event-quantification reanalysis of the embedded clause through positing an event measure function *much,* along the lines of (10) 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* – a potential reason why the illusion does not have a stable interpretation.

(10)　　more λd . [$\exists e$. OEMR ([[have been to Russia]])(e)(λx . people(x)=d)]

　　　　than wh λd . I have been to Russia d-much

The two proposed accounts differ crucially in terms of whether or not the anomaly in the illusion sentence is implicitly detected by the parser. The event comparison approach is fundamentally a *repair* hypothesis, meaning that the illusion is accepted in spite of its known ungrammaticality, because a suitable alternative interpretation can be derived relatively easily. The "illusion" of acceptability arises because we are not consciously aware of the changes we have made to accommodate the anomaly in the illusion sentence. By contrast, a template blending account such as that proposed by Townsend & Bever (2001) is fundamentally a *shallow processing* hypothesis, since the illusion arises due to the incompleteness of the grammatical representation – and as a result, the anomaly can pass undetected.

The goal of this chapter is to compare these two accounts by using sensitive online measures to assess whether there is evidence for implicit anomaly detection. Using self-paced reading to elicit reading times for illusion and non-illusion control comparatives, we examined reactions to the auxiliary of the *than*-clause, the point at which all remaining grammatical continuations (such as *than I expected*) are ruled out. We reasoned that, under a shallow processing 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 to be acceptable, and the anomaly has not been detected. By contrast

---

[5] The proper analysis of event-related readings of adnominal quantifiers remains a topic of continued debate (see Doetjes and Honcoop 1997, Barker 1999). I am 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.

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, which should result in observable increases in reading times in the illusion conditions.

Note that a slowdown at the critical region of illusion sentences can only speak to the feasibility of a shallow processing account if it can be shown to occur even in cases where people seem to be "fooled" by the illusion. An overall increase in reading times for Escher sentences, for example, may simply reflect the fact that certain participants on certain trials were able to consciously detect the problem in the sentence, slowing down in response to it, whereas others were unaware and read illusions and controls at similar rates. This scenario would be fully compatible with both the blending account of Townsend & Bever (2001) as well as the event comparison approach. Thus, a secondary aim was to compare reading times in scenarios where comprehenders are likely to be "fooled" by the illusion with scenarios where they are not. One strategy for doing so is to explicitly manipulate factors that are known to affect offline acceptability, since acceptability ratings are likely to provide an index of the frequency of detection. If there is slowdown that is associated exclusively with conscious anomaly detection, then both the repeatability of the predicate and the plurality of the *than*-clause subject should affect its magnitude, given the findings of Wellwood et al. (2009, 2012). Comprehenders are also more likely to detect the problem in tasks that facilitate deeper semantic processing and with increasing exposure to the problematic sentences, predicting a positive relationship with both of these factors. This chapter will examine each of these factors in turn to assess whether there are differences in reactions times to Escher sentences and non-illusion controls, and if so, whether these differences are fully reducible to conscious anomaly detection.

3.2 Experiment 1: Reading times, *More* NPs… *than the* NP *did*

Experiment 1 elicits reading times using a combined reading and acceptability judgment task. Participants read Escher sentences and their non-illusion comparative counterparts one word at a time and then indicated their acceptability; reactions to the illusion sentences – both online and off – were collected for analysis. To determine whether reading times differed as a function of offline acceptability, offline ratings were explicitly modulated

by relying on factors that are known to affect acceptability, namely the repeatability of the predicate (Wellwood et al. 2009, 2012). On the event comparison approach, illusions with a singular VP cannot be shifted to a comparison of events, which is why these sentences are rated lower; we expected to replicate the pattern of ratings reported by Wellwood et al. (2009, 2012). The primary question of interest, however, lies in whether the mechanisms that lead to these ratings can be detected as disruptions in reading times. 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 anomaly can pass undetected by the parser, then we expect to find no differences in processing difficulty for illusions and controls.

As noted above, a major goal if this experiment was to investigate reaction times to illusion sentences *that are perceived to be highly illusory.* Any account of this phenomenon would predict that conscious detection of a grammatical anomaly would yield measurable processing difficulty; however, a repair account makes the more specific prediction that *all* illusions, whether detected consciously or not, disrupt processing times. We predicted that comprehenders would be more likely to detect the illusion when it contained a non-repeatable predicate, or when they had been exposed to many tokens of it, leading a shallow processing account to predict longer reading times in both cases.

Finally, 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, one might reason that repeatable predicates are easier to process, and thus the *than*-clause is not deeply parsed or thoroughly integrated into the matrix comparative – thus making the repeatability facts potentially consistent with such an account. 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 would lead 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 of the relative ease of reanalysis. This type of result could possibly explain the effects of repeatability as a result of shallow processing.

3.2.1   Methods

3.2.1.1   Materials & Design

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) and PREDICATE TYPE (repeatable vs. non-repeatable):

	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. Experiment 1 design.

The stimuli consisted of 48 target sentences; each participant saw one condition of each item. In the illusion conditions, the sentences were syntactically and semantically anomalous due to the presence of a singular definite noun phrase in place of the usual bare plural *than*-clause subject: *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.* Non-illusion control comparatives were identical to the illusion sentences but contained a bare plural *than*-clause subject, so that the common perception of the sentence's meaning matched that conveyed by its syntax.

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 first 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 in the simple past tense. The critical region included the auxiliary, *did/were,* where grammatical continuations of the illusion are no longer possible, and the following spillover region, which was always eight words long and was the same for all conditions (e.g., *because of the beautiful beaches and warm weather*).

As discussed in Section 2.1, comparatives may also contain some element of negation associated with *more;* to test whether this was of import, the degree quantifier type was also counterbalanced: half of the items contained comparative *more* and half the equative *as many.*[6] This factor did not modulate reactions to illusion sentences by any measure, however, and so will not be discussed further.

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. The filler items included degree constructions and VP ellipsis sentences with various other anomalies (including island effects and antecedent mismatch in ellipsis) and focus ambiguities.

3.2.1.2   Participants

24 undergraduates from the University of Southern California (Los Angeles, USA) completed the experiment and were each paid $10 for their participation. All participants were native English speakers with no known language disorders.

---

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

3.2.1.3   Procedure

The experiment was administered 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 participants spent 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."

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.

3.2.2   Results

3.2.2.1   Ratings

Prior to analysis all ratings were standardized based on each participant's mean ratings for all experimental items, including fillers. The raw and transformed ratings are shown in Table 5 below:

54

	ILLUSION: *…than the lawyer did*	CONTROL: *…than lawyers did*
NON-REPEATABLE: *More judges retired to Florida …*	raw: 4.30 (1.79) z-score: -.26 (.85)	raw: 5.82 (1.33) z-score: .61 (.83)
REPEATABLE: *More judges vacationed in Florida …*	raw: 4.63 (1.72) z-score: -.07 (.62)	raw: 5.70 (1.52) z-score: .52 (.70)

Table 2. Experiment 1: Mean (standard deviation) of acceptability judgments by condition.

The ratings were modeled with mixed effects linear regression models using the lme4 package in R (Bates et al 2013) with fixed effects for *illusion, predicate type* and their interaction. Fixed effects for *trial order* and its interaction with *illusion* were also included in order to assess whether perceptions of illusion sentences changed over time, particularly given the large number of target stimuli. *Trial order* was represented as a continuous variable corresponding to the order among the target items, and it was centered and scaled prior to analysis in order to ease model convergence and interpretation. Because of the complexity of the model, it was not feasible to add a maximal random effects structure (per recommendations by Barr et al. 2013); instead, a data-driven approach was taken to selecting random slopes, by including only slopes revealed to create a significantly better model fit (without yielding convergence errors), using forwards model comparison with the anova() function and a liberal threshold of $\alpha = .15$ (see 1.3.2 for more details). This procedure justified the inclusion of random slopes for the effect of *illusion* by subject, and the effect of *order* and *repeatability* by item.

The presence of the anomalous illusion was found to seriously and significantly affect acceptability ratings, irrespective of predicate type ($\beta = -.86$, t = -10.36; $\chi^2(1) = 39.76$, $p < .001$). This main effect was qualified by a significant interaction with predicate type, such that illusion sentences with repeatable predicates were judged as more acceptable than those with non-repeatable predicates ($\beta = .28$, t = 3.47; $\chi^2(1) = 12.02$, $p < .001$), while there was no general effect of repeatability. In addition, although there was no main effect of trial order, the acceptability of illusion sentences decreased significantly over the course of the experiment ($\beta = -.08$, SE = .003, t = -2.0; $\chi^2(1) = 3.98$, $p < .05$).

The acceptability of the target sentences with respect to different filler types is plotted in Figure 3. Whereas ratings for control comparatives are clearly within the range of other grammatical comparative and ellipsis filler sentences, those for illusion comparatives fall in between the two extremes: they are not as unacceptable as comparatives with e.g. island violations, yet they are not perceived as normal sentences, either.
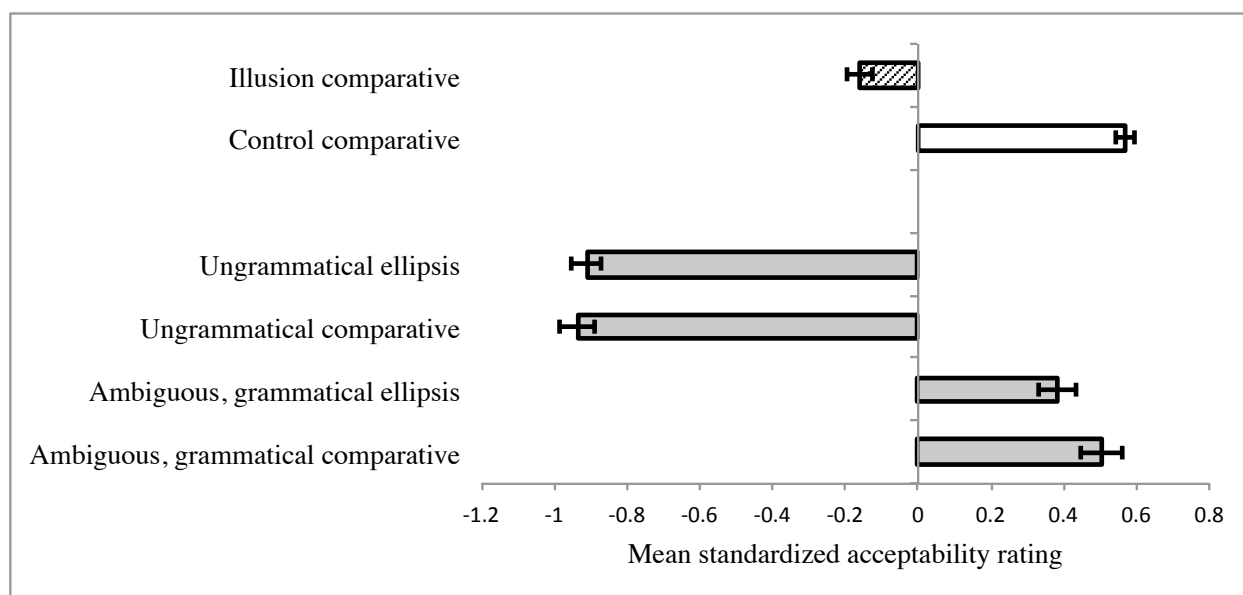
Figure 3. Ratings for illusion sentences relative to filler sentences.

3.2.2.2 Reading times

The reading time patterns are shown in Figure 1 below. Prior to analysis, extreme outliers (<150 ms or >3000 ms) were removed from the dataset. Any values over three standard deviations from the mean reading time per word position were then adjusted to the cutoff value, calculated separately for each condition. Overall, 3% of the data were affected. The reading times were modeled as described above, with separate models for each region, starting at the determiner in the *than*-clause (which was only present in illusion sentences), or the auxiliary (in controls). The final word of the sentence was omitted from analysis due to sentence wrap-up effects at that region.

All models again included fixed effects for *illusion, predicate type* and their interaction. Fixed effects for *trial order* and its interaction with *illusion* were also included in order to assess whether and how illusion-specific reading patterns changed over the course of the experiment. Random intercepts for subject and item were included in all models, and the procedure for determining the inclusion of random slopes involved model comparison using the anova() function, as described above. Reading times are plotted in Figure 4.
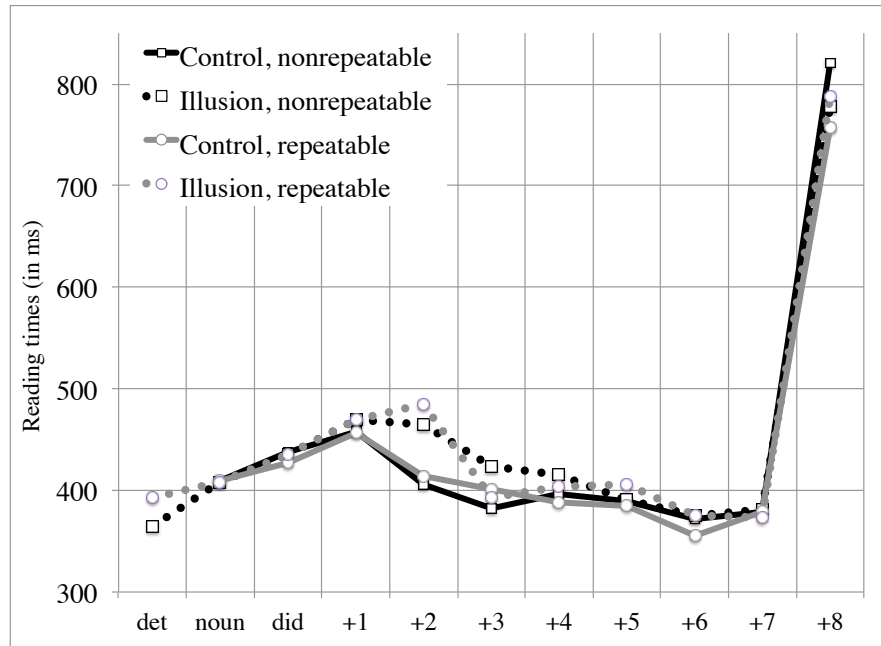
Figure 4. Experiment 1 reading times in milliseconds by condition:
… (than) my lawyer did because$_{+1}$ of$_{+2}$ the$_{+3}$ warm$_{+4}$ weather$_{+5}$ and$_{+6}$ beautiful$_{+7}$ beaches$_{+8}$

Main effects of trial order: A main effect of trial order was detected in nearly every region, starting at the determiner: as the experiment progressed, participants read through the items more rapidly, indexing some amount of fatigue throughout the course of the experiment.

Main effects of predicate type: Only the illusion conditions had determiners *my/our/the,* which preceded the point of the anomaly. These determiners were read significantly slower when the predicate was repeatable ($\beta = 27.25$, t $= 2.14$; $\chi^2(1) = 4.58$, $p < .05$). There were no other main effects associated with predicate repeatability at any other region.

Illusion-specific effects: We now examine the effects elicited specifically in illusion sentences. As mentioned, illusion sentences always differed from controls in two respects: first, they contained an additional determiner preceding the *than*-clause subject, and second, the *than*-clause subject was a singular noun phrase instead of a plural one; however, because of the theoretical possibility of a grammatically licit subset comparative continuation (e.g., *more lawyers than (just) the judge*) or an elided CP (*than the judge claimed*), the sentence did not become formally ungrammatical until the auxiliary *did/was*. Nevertheless, at the noun position (above: *judge(s)*) there was a significant interaction between presence of illusion and trial order, indicating that participants read singular noun phrases following determiners for a longer time as the experiment progressed ($\beta = 21.46$, t = 2.14; $\chi^2(1) = 4.55$, $p < .05$). There was, however, no main effect of presence of illusion, and in fact, numerically, illusion sentences were read slightly faster than controls ($\beta = -2.67$), suggesting an initial processing advantage for singular versus plural nouns that disappeared over time.

At the critical word itself – the auxiliary *did/was* – there were no effects associated with the anomaly. In the following word of the spillover region, there was a significant interaction between trial order and presence of anomaly, with illusion sentences susceptible to greater speedup effects as the experiment progressed ($\beta = -46.53$, t = -3.13; $\chi^2(1) = 5.25$, $p < .05$). Recall that the list of items was presented in one order to half of the participants, and in the reverse order to the other half of participants; as such, each item was presented once in both experiment halves. A plot of the effects of the illusion split by experiment half, as shown in Figure 5, illustrates a general pattern whereby illusion sentences were read more slowly than controls at the experiment onset, with this effect disappearing by the middle of the experiment (hence the lack of main effect associated with the illusion, given that trial order is centered here), and trending towards a reversal by the end of the experiment.
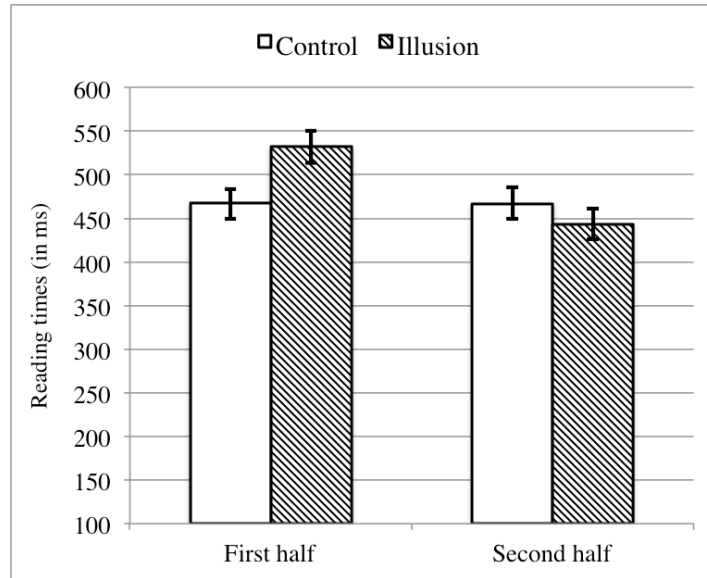
Figure 5. Illusion by trial order interaction at the first word in the spillover region (error bars represent standard error of the mean).

At the second word in the spillover region, illusion sentences were read significantly more slowly than controls ($\beta = 63.47$, t = 2.81; $\chi^2(1) = 11.10$, $p < .001$); however, unlike the previous region, there was no interaction with trial order – illusion-related slowdown remained constant throughout the experiment. This pattern is illustrated in Figure 6.
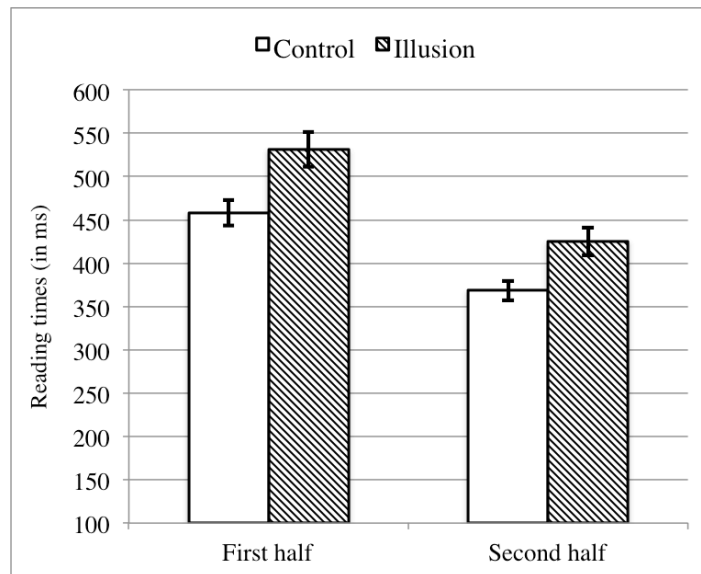
Figure 6. Main effects of illusion and trial order at the second word in the spillover region (error bars represent standard error of the mean).

Finally, in the third word of the spillover region, there was a significant *illusion* x *repeatability* interaction, such that illusion sentences with repeatable predicates were read more quickly than those with non-repeatable predicates (β = -40.02, t = -2.06; $\chi^2(1)$ = 4.25, *p* < .05), but there was no main effect associated with the illusion (β = 38.13, *p* = .14). Two words later (at *did*+5) this effect reversed, with slightly longer reading times predicted for illusions with repeatable predicates – this was significant via likelihood ratio tests only (β = 18.46, t = 1.01; $\chi^2(1)$ = 4.15, *p* < .05). There were no other significant effects associated with the illusion.

Given that we have observed significant illusion-related slowdown, a shallow processing account could argue that such slowdown is isolated to specific trials where participants happened to detect the anomaly. This explanation is not very likely given the existence of illusion-related slowdown that is insensitive to predicate repeatability, even though non-repeatable illusions were clearly perceived to be less acceptable; however, one could point to a lack of power in the reading times analysis, for example. To investigate this question in a slightly different way, we collapsed reading times for the two regions where the illusion yielded a measurable slowdown – the first and second words of the spillover region – and modeled these data with fixed effects for *rating* (a continuous variable corresponding to the standardized and centered acceptability rating for the trial in question), *illusion,* and the interaction between *rating* and *illusion,* using a maximal random effects structure.

Numerically there was a negative relationship between ratings and reading times, with higher ratings associated with faster reading times (β = -35.32, t = -2.7) although this effect was not significant via likelihood ratio tests ($\chi^2(1)$ = 1.42, p = .24). The presence of the illusion was associated with a significant slowdown in reading times (β = 31.53, t = 2.36; $\chi^2(1)$ = 3.99, p < .05); in addition, there was a significant interaction between the effects associated with the illusion and acceptability, such that high ratings predicted faster reading times for control sentences, but not illusion sentences (β = 43.49, t = 2.56; $\chi^2(1)$ = 6.35, p < .05); see Figure 7.
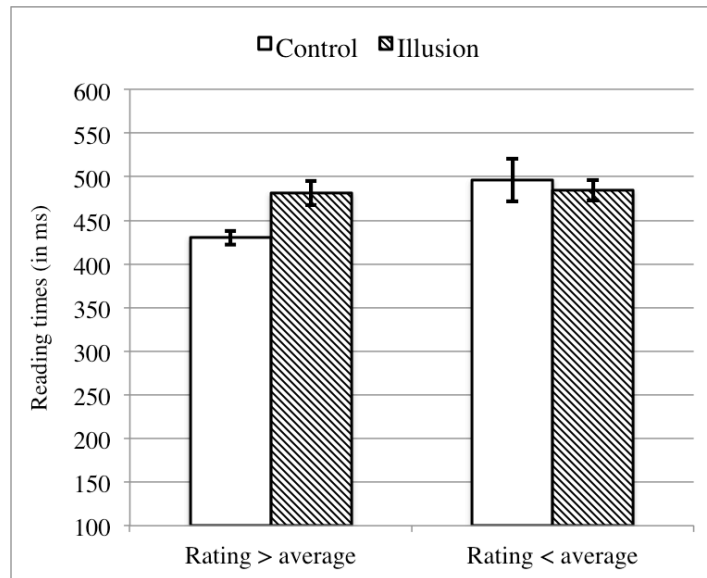
Figure 7. Collapsed reaction times at the first two words of the spillover region, plotted by offline acceptability (error bars represent standard error).

3.2.3 Discussion

The experiment outlined here investigated how deeply participants process Escher sentences. Two hypotheses were explored: one hypothesis was that the illusion arises 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, before deriving any meaning (Townsend & Bever 2000). The second hypothesis was that illusions are rated highly because they receive an appropriate 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, once again suggesting that the availability of event comparison is an important contributor to the acceptability of the illusion.

The results here indicate, however, that regardless of offline acceptability, the processor does distinguish between anomalous illusion sentences and non-anomalous controls. At the

onset of the spillover region, illusion sentences are read significantly slower than controls, suggesting that there is at least implicit detection of the anomaly that comprehenders do not consciously report. Although ratings for illusion sentences tended to decline over the course of the experiment – suggesting that participants may have started to consciously detect the illusion with increasing exposure – this pattern did not correspond to an increase in reading times over time. In fact, although low ratings were associated with longer reading times in control sentences, precisely the opposite pattern emerged with illusion sentences: as participants progressed through the experiment and became more familiar with the illusion, they slowed down *less*. The finding of general processing difficulty associated with the illusion – difficulty that does not seem at all related to conscious detection – speaks strongly against a shallow processing explanation of the phenomena. If participants were determining illusion sentences to be acceptable prior to any grammatical analysis, then slower reading times should go hand in hand with conscious detection, while trials where the participant is "fooled" should yield no slowdown at all. Clearly, the parser can and does differentiate between the ungrammatical illusions and grammatical controls absent conscious detection, and for this to happen, the illusion must be processed at a relatively deep level.

The pattern of results in this experiment speaks in favor of a repair analysis of Escher sentences: the parser detects the anomaly, but the perception of grammaticality arises because of a relatively easy switch to an alternative interpretation, potentially a comparison of events. Reading times suggest tentatively that event comparison could be an available analysis as soon as a repeatable predicate is encountered: slower reading times for sentences with repeatable predicates at the determiner – where it is not yet clear that the illusion sentence is ungrammatical – could be interpreted as a competition effect. If comparatives with plural VPs are systematically ambiguous between individual and event quantification readings, then maintaining both analyses may take processing effort, leading to slowed reading times. Such competition effects are relatively common in cases of lexical ambiguity, where increased fixation times are common in pragmatically neutral contexts (Rayner and Duffy 1986; Duffy, Morris, and Rayner 1988); when context provides suitable information for selecting one meaning over another, such competition effects disappear. If the source of event ambiguity is

inherently lexical (a possibility explored in Krifka 1990)[7], then the main effect of repeatability observed here is more or less in line with prior literature: non-repeatable predicates make event readings unavailable, and therefore no competition effects arise, while repeatable predicates make event-measuring readings salient enough that they generate observable ambiguity effects. The timing of the effect, slightly after *than*, suggests that the parser considers this ambiguity important either in preparation for reconstruction of the ellipsis site, or for fixing the units of measurement for comparison. Alternatively, it may be possible that repeatable predicates are simply representationally more complex in some way, especially given that plurality itself tends to increase reading times (e.g. Wagers, Lau & Phillips 2009).

When individual quantification is determined to be impossible there will be a shift to event comparison if such an option is available; otherwise the sentence will be perceived as unacceptable. These effects surface in different places in the sentence. The termination of a parse as ungrammatical seems to incur special computational difficulty – over and above whatever cost is incurred by the operations associated with anomaly detection – since illusions with non-repeatable predicates were associated with a slightly more prolonged slowdown than illusions with repeatable predicates, three words into the spillover region.

Below a second experiment is presented, designed to address several potential concerns related to the 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 when sentences are only judged for acceptability. To address this concern we changed the task such that participants repeated – rather than paraphrased – items, since this task is less likely to trigger deep processing of the meaning.

Second, Experiment 2 will test whether the slowdown was in fact caused by superficial properties of the illusion condition, which differed only in the presence of an extra determiner

---

[7] Krifka (1990) develops dual analyses of event ambiguities, one of which is inherently structural (an additional determiner is projected that combines with the NP to yield event measurement), and one of which is lexical (positing dual meanings for numerals). If slowdown associated with repeatable predicates is tied to this ambiguity, then it may be more plausible to adopt the lexical ambiguity approach, given that the evidence for slowdown associated with syntactic ambiguity is less clear (see Clifton, Staub & Rayner 2007; Clifton & Staub 2008).

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 semantic and syntactic parallelism between noun phrases to resolve ellipsis, and is thus sensitive to differences between e.g. names and definite descriptions (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.

3.3   Experiment 2: Reading times, *More* NPs…*than my* NPs *did*

3.3.1   Methods

3.3.1.1   Materials & Design

As before, a self-paced reading task was 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 not be essential to the phenomenon, since the illusion persists even with pronouns and singular NPs and therefore cannot be exclusively caused by dropping

or ignoring determiners (although we will address this question in greater detail in Chapter 4). To avoid this possibility only the items from the first experiment that had possessive determiners were used, since those determiners have more semantic import and are thus less likely to be ignored. This change should reduce the likelihood that reactions to illusions would be affected by complications associated with the discourse requirements of the definite article, since possessive pronouns are more easily accommodated without additional context.

3.3.1.2   Participants

50 participants were recruited from Amazon Mechanical Turk, and were paid $2.25 for their participation. All participants were self-described native monolingual English speakers with a U.S. IP-address and with a task approval rating of 97% or higher. Due to the length of the experiment, the sensitivity of reading times to outside distractions, and the fact that participants were not monitored in a quiet lab environment, responses were screened prior to analysis to ensure that participants had indeed paid careful attention to the task at hand; only participants who correctly answered items that tested that they were using the appropriate end of the rating scale (for nonsensical filler sentences such as *The salad build a fork ten times* or normal filler sentences such as *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. Eight participants did not meet these criteria and their data were not included in the analysis. In addition, one participant was excluded for providing paraphrases that were not consistent with native proficiency of English; and one for excessively short reaction times (*M* = 180 ms), suggesting a strategy of holding the space bar down to proceed through the experiment as quickly as possible.

3.3.1.3   Procedure

The experiment was conducted online using Ibex (designed by Alex Drummond, University of Maryland, http://spellout.net/ibexfarm). The procedure was similar to that of Experiment 1, except that participants were asked to *repeat* items when prompted, rather than

paraphrasing them. Participants were told that the repetition task might 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.

3.3.2   Results

3.3.2.1   Ratings

Prior to analysis the ratings were again standardized to z-scores based on mean ratings across all items for each participant. The normalized ratings were modeled as before, with fixed effects for *illusion, repeatability,* and their interaction, as well as *trial order* and its interaction with *illusion*. Random intercepts were included to model variation across subjects and items. In addition, model comparison justified the inclusion of random slopes for trial order by subject, as well as repeatability, illusion and trial order, as well as the interaction between illusion and trial order, by item. The model revealed a significant main effect of presence of illusion, with illusion sentences rated significantly lower than controls (β = -.22, t = -2.97; χ^2 = 7.95, $p < .01$), but no interaction with repeatability: illusions with repeatable predicates were no more acceptable than illusions with non-repeatable predicates ($p = .65$). There was no effect of trial order and no interaction between the presence of illusion and trial order. The raw and transformed mean ratings are given in Table 3.

	ILLUSION: *…than my lawyers did*	CONTROL: *…than lawyers did*
NON-REPEATABLE: *More judges retired to Florida*	raw: 5.09 (1.61) z-score: .25 (.68)	raw: 5.54 (1.44) z-score: .46 (.63)
REPEATABLE: *More judges vacationed in Florida*	raw: 5.09 (1.71) z-score: .25 (.73)	raw: 5.40 (1.64) z-score: .40 (.71)

Table 3. Experiment 2: Mean (standard deviation) of acceptability ratings by condition.

3.3.2.2   Reading times

Extreme outliers falling below 150 ms or above 3000 ms were trimmed from the data, and any values falling over three standard deviations from the mean (calculated separately for each word position and each condition) were adjusted to the cutoff value, affecting 3.3% of the reaction times. Analysis proceeded as before: separate models were created at each of the regions starting at the determiner and ending at the second to last word of the sentence (*did* + 7). All models included predictors for *illusion, repeatability,* and their interaction; as well as *trial order* and its interaction with the illusion. Random intercepts to model variation across subjects and items were always included, with the random slopes determined through model comparison using the procedure described for Experiment 1. The resulting reading times are plotted in Figure 8.
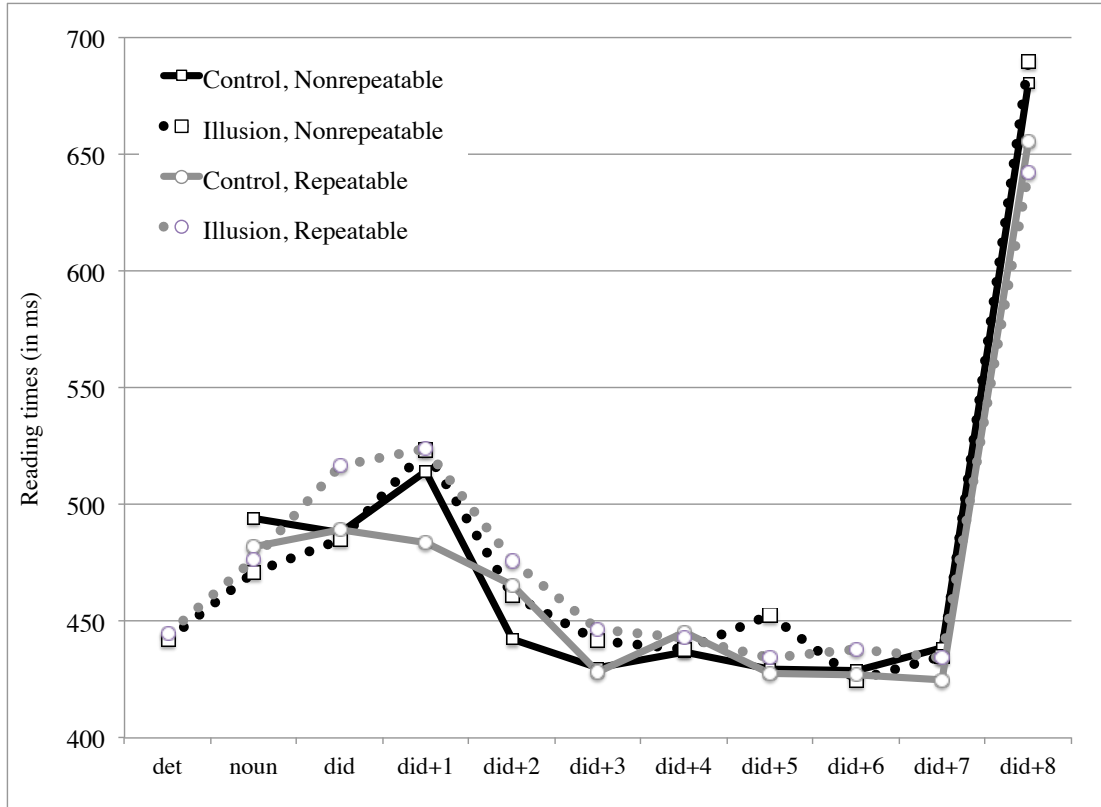
Figure 8. Reading times by word position and condition for Experiment 2.

As in Experiment 1, there was a main effect of trial order in all regions (except as noted below), such that participants read more quickly as the experiment progressed; see the Appendix for model details. There were no main effects associated with predicate type detected at any region.

The presence of the illusion began to affect reading times at the critical word *did.* As in Experiment 1, there was a significant interaction between presence of illusion and trial order, with illusion-related slowdown decreasing significantly as the experiment progressed (β = -29.46, t = -2.52; $\chi^2(1) = 6.36$, $p < .05$). The main effect of trial order consistently observed in other regions was less clear at this region, with the model indicating that the parameter estimate was not significantly different from zero (t = -1.72) but with likelihood ratio tests suggesting a better fit for the model including a fixed effect for trial order than those that did not ($\chi^2(1) = -9.75$, $p < .01$), likely because the former explained some of the variation that would otherwise be attributed to the omitted interaction term. The same general pattern

persisted into the following region, both in terms of the main effect of trial order (β = -19.02, t = -1.37; $\chi^2(1)$ = 6.92, p < .01) as well as its interaction with the illusion (β = -29.31, t = -2.32; $\chi^2(1)$ = 5.39, p < .05). Finally, at the second word in the spillover region there was a numerical difference between illusions and controls that was smaller in magnitude and non-significant (β = 17.60, t = .99) and did not change over time; there were no other effects through the end of the sentence.
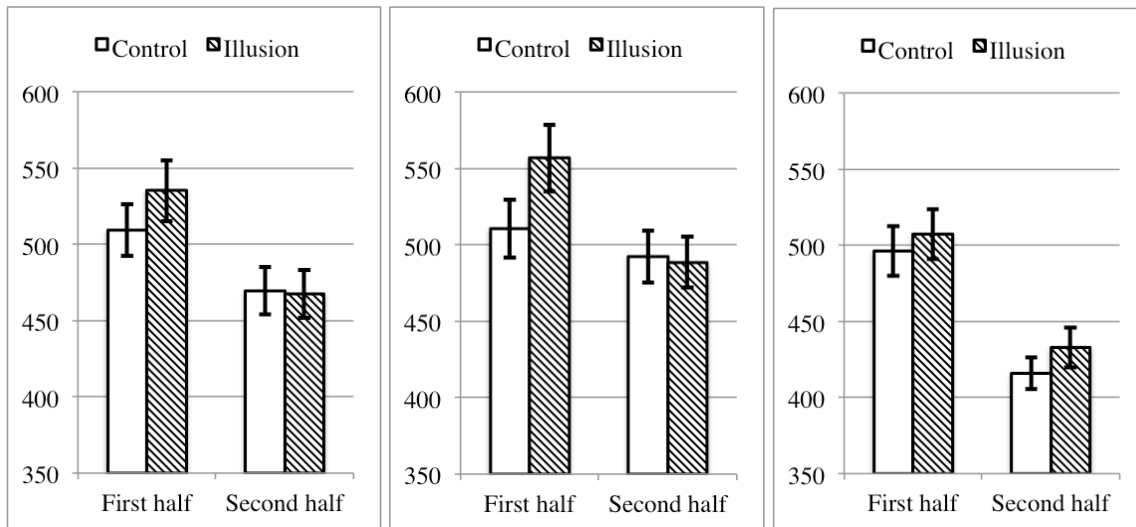
Figure 9. Illusion-associated ordering effects at the critical region, with mean reading times (in milliseconds) plotted by experiment half for *did* (left), *did+1* (middle), *did+2* (right). Error bars represent standard errors.

The visual plots of reading times in Figure 9, combined with the results of Experiment 1, suggest important qualitative differences between regions. In the early regions, the illusion-related processing difficulty was attenuated over time, while in the later region, the illusion-related slowdown was small in magnitude (too much so to emerge as significant), but as in Experiment 1, it remained constant over time. To assess whether these properties emerged as stable when generalizing across experiments, the reading times from both experiments were collapsed and a model for each region was constructed with fixed effects for *illusion, trial order* and their interaction. An additional goal of this analysis was to disentangle the effects of trial order from the effects of acceptability, to assess whether these properties are associated with distinct temporal profiles. As in Experiment 1, predictors for *rating* (a centered, continuous variable corresponding to the acceptability rating for the trial in question) and its

interaction with *illusion* were included. Random effects were included to regress out variation across *items, subjects* and *experiment,* with model comparison used to determine random slopes (details about random slopes can be found in the Appendix).

These models confirm that there is a biphasic pattern of online processing associated with illusion sentences. The first two words – *did* and *did+1* – are associated with a slowdown for illusion sentences that decreases significantly over time (*did*: β = -18.25, t = 2.09; $\chi^2(1)$ = 3.95, *p* < .05; *did+1*: β = -34.68, t = -3.39; $\chi^2(1)$ = 11.5, *p* < .001), but the association between ratings and reaction times is weak or non-existent. In the following region, however, the opposite pattern can be observed. Although the model points to a significant effect of trial order (β = -54.08, t = -6.76; $\chi^2(1)$ = 36.21, *p* < .001), there was no general slowdown associated with the illusion and no change in illusion-related reading times over the course of the experiment. Instead, a main effect of acceptability emerged as marginally significant, with higher acceptability ratings generally predicting faster reading times (β = -30.97, t = -3.10; $\chi^2(1)$ = 2.87, *p* = .09), *except* in the case of illusion sentences, where higher ratings are robustly associated with *longer* reading times (β = 32.87, t = 2.65; $\chi^2(1)$ = 6.76, *p* < .01).

To summarize, an analysis over the reading times of both experiments reveals that the magnitude of the first phase of slowdown is predicted by trial order but not ratings, while the second phase of slowdown is predicted by ratings but not trial order. Somewhat surprisingly, the greatest amount of slowdown is associated with early exposure to illusion sentences and illusion sentences that are perceived to be highly acceptable as compared to those perceived to be less acceptable.

3.3.3   Discussion

This experiment 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, they were asked 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, illusions of a different sort were tested, namely those with plural *than*-clause subjects. Plural illusions differ first in that they are generally found to be more acceptable than those with singular *than*-clause subjects (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. Despite the changes in design to facilitate shallow processing and maximize the illusion, illusions were again rated less acceptable than controls and read 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 basic shallow processing account.

The key patterns in reading times in Experiment 2 were largely the same as those in Experiment 1, with some differences in timing. At the onset of the critical region illusion sentences were read more slowly than controls; however, this slowdown was heavily mediated by a change over the course of the experiment, such that comprehenders gradually slowed down less in the illusion conditions and eventually read illusions and controls at roughly the same rate. And, as in Experiment 1, the illusion-associated slowdown was not driven by trials where comprehenders had consciously detected the anomaly, with the results suggesting in fact the opposite pattern. Finally, in the second spillover region, a descriptive pattern emerged whereby illusion sentences were read more slowly than controls, but this slowdown did *not* decrease over time. This effect was small in magnitude and nonsignificant, but parallels the pattern found in Experiment 1, and emerges as stable in models collapsing results across both experiments.

In Experiment 1, illusion sentences were read increasingly more slowly throughout the experiment, an early effect that emerged at the noun – and thus before the anomaly was even apparent. The fact that this effect did not surface in Experiment 2 suggests that it was likely related to differences in the morphological number of the *than*-clause subject across the two conditions. In Experiment 1, control conditions always had plural subjects, while illusion conditions had singular ones; the representational complexity associated with the plural-marking in control conditions may have therefore caused some added difficulty at the noun,

relative to illusion sentences with singular noun phrases; in Experiment 2 both illusions and controls had plural nouns phrases and no similar processing asymmetry was observed. Alternatively, this effect may be related to the processing cost associated with positing a null measure function *d-many* before the plural noun phrase, an operation that cannot be entertained for a semantically singular noun phrase, perhaps leading the parser to expect an appropriate plurality later in the sentence.

As expected, plural illusions in Experiment 1 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 the results suggest, since a robust interaction was observed between the presence of the illusion and the repeatability of the predicate with singular *than*-clause subjects (Experiment 1), but not plural ones (Experiment 2). 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 *than*-clause subjects. 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.

Note that, in addition to predicate repeatability not exerting an effect on acceptability, there was no interaction in the reading times, either – unlike in Experiment 1, where illusions with non-repeatable predicates sustained a more prolonged slowdown extending into the third word of the spillover region. This suggests that the prolonged slowdown in Experiment 1 is in fact associated with the termination of a parse as ungrammatical; and because illusions with non-repeatable predicates were perceived as acceptable in Experiment 2, no such effect arose. We address this in greater detail in the next section.

3.4   General Discussion

The two self-paced reading experiments presented here take a closer look at Escher sentences, a type of semantic illusion that yields a felicitous percept in spite of having no coherent meaning. It is a puzzle to many (if not most) sentence processing theories how Escher sentences arise: acceptability judgments should logically require 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 (2001), 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 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 here 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 (*especially* the most acceptable sounding ones), and even with simple tasks – such as rote repetition – that should clearly allow for shallow processing. The evidence presented here thus weighs heavily against the account proposed by Townsend & Bever (2001), who predict the level of analysis at the time the sentence is rated to be extremely shallow, lacking any grammatical or semantic analysis. The illusion is clearly 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. This would explain the positive relationship between

ratings and slowdown in a natural way: successful repair increases acceptability, but also presumably incurs a processing cost.

Note that the event comparison and shallow processing accounts stand at different ends of the spectrum in their explanation of the acceptability of the illusion: whereas the event comparison hypothesis posits immediate, exhaustive and grammar-based interpretation with localized changes to fix targeted problems, the uninterpreted heuristics approach of Townsend & Bever (2001) posits no grammatical analysis at all at the time the illusion is first judged acceptable. One might wish to appeal to a middle ground, such that the syntax of the illusion is interpreted, but at least some aspect of its interpretation is associated with shallow or heuristic processing. For example, a dual-route model such as Ferreira et al. (2002) may be able to account for processing difficulty at the illusion by arguing that syntactic parsing happens incrementally up to the point where reanalysis is required, at which time grammatical analysis fails and a heuristic takes over. This is similar to the widespread misinterpretation of garden path sentences: Christianson et al. (2001) found that sentences like *While Anna dressed the baby played* yield a measureable increase in reading times, but that in comprehension questions participants continued to indicate that the sentence meant that *Anna dressed the baby* as well as that *the baby played.* They interpret this as support that syntactic reanalysis is initiated but that interpretations are still influenced by a lingering "NVN"-heuristic interpretation.

In the case of Escher sentences, an NVN heuristic plays no obvious role, so one would need to define an alternative that is sensitive to the semantics of comparison, in particular one that can detect event plurality, without taking into account the details of the grammar. This quickly becomes problematic, however: such models quickly become ad hoc and add only minimal explanatory force while requiring the addition of problematic assumptions. To see why this is so, we will consider two questions in sequence: first, the nature of the meaning provided by the heuristic (and by extension, the nature of the heuristic); and second, the motivation for the existence of this heuristic.

With respect to garden-path sentences, the NVN heuristic posited by Bever (1970) generates a relatively simple semantic output: it takes a sequence of words consisting of a noun-verb-noun – e.g. *the boy kicked the ball* – and yields an interpretation where there is an event involving the verb whose agent is the first noun and patient is the second noun. Ferreira (2003) has demonstrated the effects of this heuristic by focusing on interpretations people

obtain from passive sentences. For example, a sentence like *The boy was kicked by the ball* is sometimes construed with *boy* as agent and *ball* as patient, in spite of its unambiguous syntax. Ferreira suggests that this meaning is made available by a rapidly applied NVN heuristic, which usually – though crucially, not always – is overridden by the details of the grammatical analysis. The semantics of comparatives is more complex than that of a simple active sentence, as we saw in Chapter 2, and thus the nature of the heuristic would be correspondingly trickier to define. The results of Wellwood et al. (2009) suggest that it must approximate a comparison of events, or is at minimum is highly underdetermined with respect to what is being counted (and thus at least *allows* for a comparison of events even if it does not require it). This might be accomplished, for example, if the two clauses of the comparative could be independently interpreted, and then integrated heuristically, as shown below:

(11)   a. Lawyers have vacationed in Florida.

          b. You vacation in Florida.

          c. The first amount is greater than the second amount.

To arrive at these meanings, we must then specify the extent to which the interpretation of the two independent clauses is derived grammatically or heuristically. As far as local grammatical analysis is concerned, the way each clause is independently interpreted is not obvious. First, how does the parser target the sequence *People have been to Berlin* without including *more?* The matrix contains no local constituent corresponding to *people have been to Berlin* that excludes *more* or its degree variable (*d*-(many) *people have been to Berlin*), and the presence of *more* in nominal position seems to strongly signal a comparison of individuals. The grammar would need to "turn off" enough to completely ignore the nominal position of *more,* then "turn on" to analyze a non-constituent *people have been to Berlin,* and then "turn off" again to project some contextually determined scale onto that proposition. In other words, an explanation is needed as to how and why some parts of the clause are parsed and interpreted grammatically to the exclusion of others.

     With respect to the *than*-clause, local analysis of the string *than I have* should yield the meaning (*than*) *I have been to Berlin* only via global grammatical integration. In particular, the content of the ellipsis site depends on details about how the *than*-clause is integrated into the

matrix, as evidenced by known correlations between the size of the ellipsis and the merge position of the *than*-clause (Sag 1976, Williams 1974). The pattern of judgments in (12), for example, indicates that the degree operator *–er* and its associated elided material within the *than*-clause need to c-command the antecedent *tell her to work d-hard,* which itself conflicts with the TELL > ER scope configuration. This well-established correlation has been used to argue for a crucial relationship between the position of the *than*-clause and the way its contents are reconstructed.

(12)   Mary's father tells her$_i$ to work harder than her$_i$ boss does Δ.

    a.   TELL > -ER, elided material = work d-hard

        $\approx$ Mary's father tells her: work harder than your boss works.

    b.   *TELL > -ER, elided material = tell her to work d-hard

        $\approx$ Mary's father tells her: work harder than your boss tells you to work.

    c.   -ER > TELL, elided material = work d-hard

        $\approx$ Mary's father tells her: work d_1-hard; Mary's boss works d_2-hard; $d_1 > d_2$.

    d.   -ER > TELL, elided material = tell her to work d-hard

        $\approx$ Mary's father tells her: work d_1-hard; Mary's boss tells Mary: work d_2-hard; $d_1 > d_2$.

Because there is considerable evidence that the content of the ellipsis site is reconstructed when there is global integration, it is not immediately clear what it would mean to have a strictly local grammatical analysis of *than I have,* with ellipsis resolution, *before* global integration is initiated.

Perhaps there are ways around these problems, but at minimum, something more will need to be said about how and why we are allowed to interpret the two independent clauses using the grammar, yet fail to integrate them grammatically. Otherwise, if the grammar cannot straightforwardly build these two clauses independent of one another, we are left to assume full heuristic semantic analysis, both locally and globally. One possibility is that the *NVN* heuristic can be recruited to generate the internal content of the matrix, and then an additional heuristic integrates the clauses in the way described above (the way ellipsis is resolved still remains a mystery, however). The nature of comparison is then determined wholly through context.

(13)   Surface: More people bought toys than I did (buy toys)

        NVN heuristic: More people$_{AGENT}$ bought$_{VERB}$ toys$_{THEME}$

                                    than I$_{AGENT}$ did buy$_{VERB}$ toys$_{THEME}$

        Heuristic integration: the quantity associated with

                                    [people$_{AGENT}$ bought$_{VERB}$ toys$_{THEME}$] >

                                    the quantity associated with

                                    [I$_{AGENT}$ did buy$_{VERB}$ toys$_{THEME}$]

However, if the final interpretation is strictly associated with heuristics and not grammatical analysis, then it follows that any person "fooled" by an illusion sentence like (14) will also perceive there to be *no interpretational difference at all* between (14)-(15) which (at least on an intuitive level) does not seem right.

(14)   More customers$_{AGENT}$ emailed$_{VERB}$ the salesmen$_{THEME}$ than I did.

(15)   More customers$_{AGENT}$ were emailed$_{VERB}$ by the salesmen$_{THEME}$ than I was.

        Note further that, from the perspective of extra-linguistic cognition, there is no reason a plurality could not be compared with a singularity – this type of comparison is not actually incoherent at all. If shown a picture with three squares and another picture with only one square, and asked to pick which is "more," people will probably not respond that comparison is impossible. The plurality constraint is specifically one that is contributed by the grammar. As a consequence, by choosing to bypass the grammar, it is unclear that it is possible to explain why a non-repeatable predicate with a singular subject, like (16), would sound bad in the first place.

(16)   More lawyers have retired to Florida than I have.

        a. Lawyers have retired to Florida.

        b. I have retired to Florida.

        c. The cardinality of (a) exceeds the cardinality of (b).

Assuming that we can find solutions to these problems, the next question concerns the motivation that would underlie a heuristic that derives event comparison interpretations of illusion sentences. This question is an important one to bear in mind, particularly given that heuristics in reasoning and decision-making have been occasionally criticized for lacking explanatory power and simply restating observations under different terms (e.g., Gigerenzer 1991). If an entirely new processing stream is postulated, then at minimum the relevant heuristics should have an independently motivated reason for existing other than to predict problematic data.

The grammar-independent processing stream is usually motivated by top-down considerations, in particular the fact that it could be advantageous for a parser to track which meanings are the most likely, on the basis of information about the probability of various syntactic frames or the probability of certain events occurring in the world (and thus the probability that speaker would utter the sentence). For example, the NVN heuristic is motivated by the assumption that such linear sequences are most frequently assigned an agent-verb-theme meaning. Abstracting this pattern from the input and employing it during comprehension would lead the parser to strategically adopt the agent-verb-patient meaning as a hypothesis any time it encounters a noun-verb-noun sequence, which has a tangible benefit: it would speed comprehension of many sentences.

From what we know at this point, Escher sentences could have one of two plausible percepts. The first is a narrow comparison of events, similar to a *more-often* reading. Sentence-initial *more* can measure the noun phrase in terms of event participation when the matrix predicate denotes an action that can be repeated more than once per individual (Krifka 1990; Wellwood et al. 2009), but this is certainly not the most frequent interpretation a sequence beginning with *more NP*. While it seems plausible that a parser might keep this analysis under consideration, having it function as a heuristic is too extreme: a heuristic that takes a nominal comparative and posits an event reading as its leading hypothesis would be a very poor one, since it would turn out to be wrong a majority of the time. The second type of interpretation potentially associated with Escher sentences is an underdetermined comparison with the details about the scale determined e.g. by context. A heuristic generating this meaning fares even worse, because it is not at all grounded in probabilistic information: this use of *more* is

completely unattested in linguistic input beginning with *more NP*. In positing this heuristic, we claim that *every single time* a person encounters a nominal comparative, they contend with conflict between the frequent and grammatically-licensed comparison of individuals, and an environmentally unattested, heuristically-generated comparison with no fixed scale. In summary, there is a strong *a priori* reason to believe that the parser is biased towards the agent-verb-patient meaning from the onset, but no clear reason to think that the parser would be equivalently confident about either of the meanings above, which are either infrequent or completely unattested for a sequence beginning with *more NP*.

Finally, from the perspective of plausibility heuristics, comparison of events is not a more "plausible" percept for the illusion than the grammatically sanctioned comparison of individuals, at least not until *after* the illusion's anomaly is taken into account. But if plausibility becomes important only after anomaly detection, then by definition what we are dealing with is repair (i.e., fixing a problem) instead of heuristics.

To summarize, a major contribution of the shallow/Good Enough processing literature is the recognition that more needs to be said about the online syntax-semantics interface. Interpretation is clearly not always as seamless a process as is commonly assumed, and comprehenders appear to occasionally arrive at nonveridical interpretations of the linguistic input. Although this interesting problem deserves more attention than it has received in the past, the heuristics approach seems to be a more-extreme-than-necessary solution to the problem of Escher sentences, particularly since it lacks the precision that the grammar provides, as well as (in some cases) strong enough motivation to justify dual processing streams.

A repair hypothesis of Escher sentences, by contrast, has no difficulty taking into account the complex details of the grammar of comparison. On this account, it is not clear that there is an "illusion" at all: the parser is *never* "fooled" by the ungrammaticality of the illusion sentence – it is simply so good at repairing the problem that we fail to consciously report it. This claim may seem unintuitive given the way comprehenders usually respond to grammatical anomalies, and it remains a mystery what makes this case so different. But given that language data is often quite noisy – containing minor irregularities, incomplete, or otherwise ill-formed – and yet even child and non-native speech can be comprehended with relative ease, it would be

independently advantageous for a parser to have precisely the sort of flexibility entailed by a repair hypothesis.

In order to "repair" an Escher sentence, a complex sequence of steps must be initiated: first the anomaly itself must be detected; it must be determined whether the sentence is reparable; if so, the specifics of the repair must be selected and then executed; if not, parsing of the illusion terminates as ungrammatical (and presumably, in the context of an acceptability judgment task, the input will be compared with the closest near-neighbor grammatical parse to determine the extent of ungrammaticality). It is not surprising, then, that we have observed a complex, multi-phase response the illusion sentences in Experiments 1 and 2. In particular, we observed at least two qualitatively different points of slowdown associated with the anomalous illusion sentences, only one of which is affected by language experience.

We saw that the initial phase of the slowdown – beginning at *did* (Experiment 2) or *did*+1 (Experiment 1) – depends heavily on the position of the item within the experiment: it is strongest for the items that participants first encounter, and disappears or reverses by the end of the experiment. Intriguingly, these effects parallel the adaptation effects reported by Fine et al. (2013) for garden-path sentences like (17), who find the cost of resolving garden-path ambiguities to be attenuated over the course of an experiment as comprehenders assign increasingly greater probability to the *a priori* less frequent relative-clause analysis. Fine et al (2013) tie these effects to implicit learning mechanisms that adjust the probabilities associated with various possible syntactic analyses, and suggest that these learning mechanisms underlie syntactic priming effects. Importantly, they find that as low-probability analyses become easier to process, high-probability analysis also become more difficult to process (see Figure 10); this suggests that adaptation is crucially tied to the weighting between the two competing analyses, which presumably shifts with exposure.

(17)   The experienced soldiers *warned* about the dangers conducted the midnight raid.
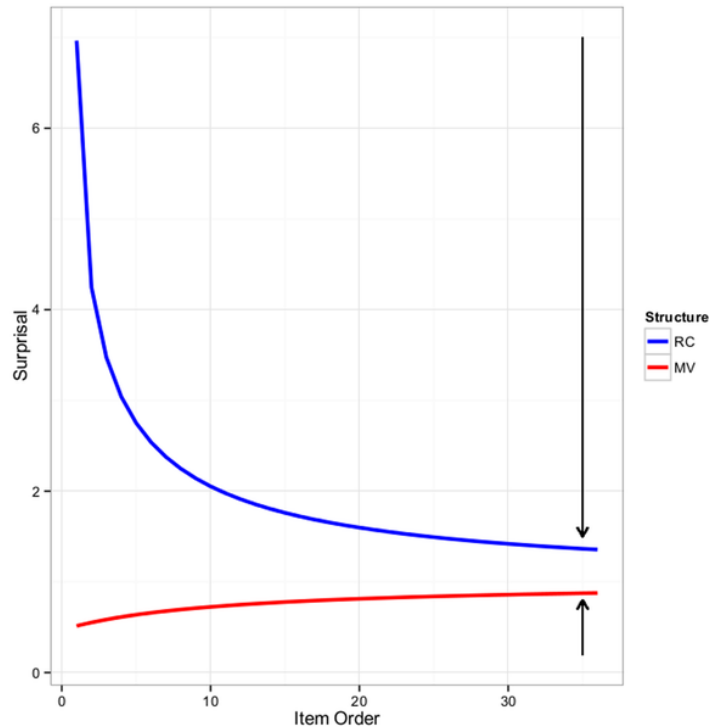
Figure 10. Fine et al. (2013)'s predictions for how the distribution of items within an experiment can change surprisal for low-frequency RC analyses as well as the high-frequency MV analyses.

Although we did not necessarily observe that control sentences became more difficult over the course of the experiment, the faster reading times associated with increasing trial order – observed at every other region in both experiments – were inconsistent in precisely the regions exhibiting illusion-specific adaptation effects. Although the inclusion of a predictor for trial order almost always yielded a better fitting model, whenever there was illusion-specific adaptation effects, the models did not find the effect of trial order to differ significantly from zero, indicating that its impact on model fit may have been related primarily to the way it interacts with the illusion. One way to interpret this is that the control sentences *did* become more difficult over time due to their decreased probability, but that this effect was cancelled out by an additional separate effect of fatigue, causing comprehenders to speed up to all stimuli overall. Note that it is not clear whether the account proposed by Fine et al (2013), which ties adaptation effects to syntactic priming, can extend naturally to inherently ungrammatical illusion sentences under consideration here – such operations are typically predicated on the existence of a licit representation, which illusion sentences notably lack (at least at a global level).

Experiments 1 and 2 are not the first to find adaptation effects for ungrammatical stimuli. Specifically, Kaschak and Glenberg (2004) found adaptation effects in response to novel constructions like *This dinner needs cooked,* which are grammatical only in certain dialects of English and thus elicit immediate slowdown for speakers of standard English. Processing costs associated with the novel syntax rapidly lessened over just a small number of stimuli. In a similar vein, Coulson et al. (1998) find adaptation effects in the P600 response to ungrammatical stimuli. The P600 component is a late positivity registered in response to both garden-path ambiguity resolution as well as syntactically anomalous sentences (Osterhout and Halcomb 1992, Gouvea et al.2010). By varying the probability of encountering grammatical versus ungrammatical stimuli, Coulson et al. (1998) were able to manipulate the magnitude of the P600; when ungrammatical stimuli were more probable in the context than grammatical stimuli (in an 80%-to-20% distribution), the amplitude of the P600 response was reduced in response to the ungrammatical sentences and *increased* in response to the grammatical sentences. Importantly, the latter finding – the increase in P600 effects for improbable but grammatical sentences – parallels Fine et al.'s finding that *a priori* more probable main-verb continuations were read more slowly within contexts that strongly supported the relative clause continuation. This type of finding cannot be captured if the relevant measures index e.g. diagnosis of a bad parse or syntactic repair, and thus are probably more closely linked to general surprise associated with encountering improbable stimuli, whether grammatical or not.

These facts suggest that the earliest slowdown observed – primarily at the word after the auxiliary *did* – may index anomaly detection, or recognition that the grammar-based parse has failed, which may become easier over the course of the experiment as the comprehender adjusts to the distribution of anomalous items and is less surprised by them. This type of slightly delayed slowdown is not at all unusual for detection of grammatical violations. Many anomalies – such as phrase structure violations, unlicensed NPIs, or mismatching morphological agreement – seem to be associated with slowdown at the word *following* the critical region in self-paced reading experiments (Warren et al. 2006 for unlicensed NPIs; Pearlmutter, Garnsey & Bock 1999 and Wagers, Lau & Phillips 2009 for verbal agreement, but see also Ditman et al. 2007). This finding is corroborated by various eye-tracking experiments. Whereas disambiguation effects for garden path sentences typically emerge immediately, in early measures such as first-pass reading times at the critical region (see Clifton, Staub &

Rayner 2007 for a review), detection of grammatical anomalies (most commonly tested with agreement mismatch) is associated with marginal or no slowdown in first-pass reading times at the critical word, but significantly increased regression rates at or sometimes after that word (Braze et al. 2002, Ni et al.1998, Pearlmutter, Garnsey & Bock 1999). Unfortunately, studies have looked at reading times for only limited types of grammatical violations, and it is unclear whether readers slow down to simply come to terms with ungrammaticality versus to revisit and revise the problem (for example, to implicitly correct the problematic subject-verb agreement).

In all illusion sentences with a plurality in the *than*-clause, following anomaly detection I have proposed that there is some type of repair, given that the quantity determiner *many* cannot syntactically combine with the full DP *than*-clause subject. If there is a suitable alternative plurality, however, the repair enables comprehension of the ungrammatical illusion sentence by providing the measure function the flexibility to apply to it instead. The results from Experiment 1 suggest that a plurality of events, of which the *than*-clause subject is the agent, are a suitable host for the degree variable – resulting in a comparison of events. The results from Experiment 2 are less clear, given that we found no effects of repeatability in ratings or response times; this suggests that some alternative repair mechanism may be available when the *than*-clause contains a plural subject and non-repeatable predicate. This could entail generating a plurality of events distributively, by mapping each individual within the definite plural *than*-clause subject to a single event, or it could entail generating a non-event comparison reading by e.g. dropping the determiner (**than how many the judges retired* → *than how many judges retired*), or interpreting the DP partitively (*than how many of the judges retired*). The details of the range of possible repair operations remains to be determined, but we will begin to tackle this subject in Chapter 4.

The second phase of the response to illusion sentences – at the second word of the spillover region – involved a slowdown that was smaller in magnitude and not susceptible to adaptation effects at all. Rather, the magnitude of slowdown at this region was tied to the offline acceptability of the illusion sentence, with more acceptable-sounding sentences eliciting *more* slowdown. One possibility is that repair is initiated at this region, the difficulty of which does not change over the course of the experiment. The late onset of semantic repair is appropriate given that interpretive processing is associated with delayed and sustained

slowdown over several words in self-paced reading, in contrast to the more short-lived slowdown for syntactic anomalies (Ni et al.1998, De Vincenzi et al. 2003, Ditman, Holcomb & Kuperberg 2007); however, most studies have focused on pragmatic infelicities such as *At breakfast the boys would <u>plant</u> toast and jam.* A more straightforward index of semantic processing is found in studies on semantic enrichment for sentences like (18)a, whose processing patterns in some ways parallel the illusion findings.

(18) a. The author began the book...   Coercion
      b. The author read the book…   Control, "non-preferred" event
      c. The author wrote the book...   Control, "preferred" event

Interpretation of (18)a requires shifting the meaning of some book-entity into a book-event, unlike sentences with "neutral" verbs like read or write. Various studies have found evidence for increased processing difficulty associated with complement coercion (18)a when compared to controls (18)b-c, typically detected quite late in the post-target region in both SPR and eye-tracking studies (McElree et al. 2001, Traxler et al. 2005, Traxler et al. 2002). In studies that have found more immediate slowdown, the cost seems to be unrelated to the actual semantic operation itself; for example, McElree et al. (2001) find greater processing cost for events that are less easy to retrieve relative to the context (such as (18)a-b versus (18)c) while Traxler et al. (2005) find that pre-introducing the relevant event sense into the context decreases processing costs at the first region but not the second. Thus suggests an analogous multi-phrase process involving first detecting the relevant event sense and then integrating it into the semantic representation. The magnitude of this latter semantic operation is closer to what we find for illusions, but much smaller than what is observed for syntactic anomalies.

If the reading time patterns at *did*+2 index semantic repair, this could explain why greater amounts of slowdown at this region are associated with higher acceptability ratings; presumably, higher ratings are assigned to sentences that have successfully undergone repair and therefore have a meaningful interpretation. The magnitude of this effect is clearly very small in Experiment 2, where the *than*-clause subjects were all plural, suggesting either that the relevant repair operations are much easier to carry out, or else that something entirely different leads to their high levels of acceptability. For example, in cases where the *than*-clause subjects

are plural there is room not only for semantic repair to a comparison of events but also for the simpler repairs mentioned above (omission of the determiner or repair to a partitive interpretation). Changes like these do seem to impact reading times at the critical region. Recall that there were production tasks asking participants to recall the item they just read; on 30-50% of the repetition trials, participants repeated illusion sentences back without the problematic determiner. Although this could reflect late offline inferential processes, those trials tended to be associated with a large decrease in reading times for illusions at the critical region (dropped determiner: M = 472 ms, no dropped determiner: M = 579 ms), while, for example, participants who repeated items back without changing them tended to experience greater amounts of slowdown at the critical region. This suggests that certain modes of "repair" involve less, or even no, slowdown at the critical region. To avoid this problem, future work could look at reading times for illusions with plural pronouns (*we*) for which semantic reanalysis to a comparison of events is possible but other small-scale repairs or memory errors are not.

The proposal that slowdown at *did*+2 indexes repair could lead us to expect slower times at this region for "reparable" illusions (such as those with repeatable predicates) versus "irreparable" ones – a prediction that was not borne out. In practical terms, however, it is impossible to say whether the difficulty associated with repairing illusions with repeatable predicates might be equaled or surpassed by the difficulty of terminating a parse as ungrammatical. In addition, the cost difference between repair and general anomaly detection may be negligible in the context of acceptability judgment tasks, if the latter already entails some automatic process of retrieving the closest near-neighbor grammatical parses and comparing the degree of match (an operation that is presumably highly relevant both to the task of assessing acceptability on a scale, and also to determining a possible repair parse). In fact, if anything, the extended slowdown at *did*+3 associated with irreparable illusions in Experiment 1 seems likely to index prolonged effects of ungrammaticality detection. In order to investigate what online effects are associated specifically with repair, it would be necessary to directly contrast reaction times to illusion sentences that require only small-scale changes versus those that require more complex repairs.

Although the exact nature of the repair process remains mysterious, it does not likely consist of moving *more* to an adverbial position (*More people have been to Berlin than I have*

→ *People have been to Berlin more than I have*), since the illusion persists equally with degree quantifiers that are unambiguously nominal, namely *fewer* (Wellwood et al. 2009, 2012), and *as many* in the experiments here. This points to a shift that happens at the level of semantics and not syntax, a possibility we will explore in more detail in the next chapter. This could explain why the neural response to illusion sentences is somewhat different than the response to garden path sentences: whereas 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 elicited *less* activity in those regions than control comparatives (Christensen 2010). While this effect is interpreted by Christensen (2010) as support for shallow processing, it could equally be interpreted as evidence that the parser can and does differentiate between illusions and controls, but that the relevant reanalysis processes happen at different levels in the grammar. It is unsurprising then that the two constructions would activate different neural networks, since they involve entirely different processes of recovery and repair.

If a repair or reanalysis account is on the right track, then the mystery of Escher sentences shifts slightly to the question of why the operations associated with their interpretation are so subtle as to evade conscious awareness, whereas those associated with garden path sentences can be so disruptive and difficult that comprehenders may fail to ever completely recover. This relates to the widely-studied and pre-existing question of why garden path sentences themselves vary so widely in difficulty. For example, whereas the classical example in (19) is notoriously difficult to understand, those in (20)-(21) are considerably easier.

(19)   The horse raced past the barn fell.

(20)   The Australian woman saw the famous doctor had been drinking. (Sturt et al 1999)

(21)   The criminal confessed his gang harmed too many people. (Pickering & Traxler 1998)

With respect to garden path sentences, the broader literature has suggested many different factors that might influence ease of recovery, including how salient the alternative

parse is given various lexical (Garnsey et al, 1997) and pragmatic factors (Pickering & Traxler 1998), and the extent to which reanalysis changes prosodic (Bader 1998) or thematic structure (Pritchett 1988). Ease of recovery is also addressed in different ways within serial versus parallel processing architectures. Proponents of parallel processing models tend to take these types of findings to suggest that the more closely ranked two competing structural analyses are, the less difficult sentence reanalysis will be. When comprehenders are very strongly committed to the preferred but incorrect parse (because evidence overwhelmingly supports it, or the other parse decays for lack of support), the effects will be strong, and easier to consciously detect.

With respect to illusion sentences, the "preferred" parse – individual comparison – is the *only* parse sanctioned by the grammar of the comparative, so reanalysis should give rise to very strong garden path effects. There is also no special lexical or pragmatic bias against individual comparison that would make the two alternatives more closely ranked; because the shift to event comparison does not produce more dramatic slowdown (on a par with difficult garden-path sentences), it could be that the parser is not especially committed to the grammar-based parse to begin with. I have already suggested a number of reasons in section 2.3 why this may be the case: the syntax-semantics interface is complicated in the case of comparatives and may pose a special problem for incremental parsing. In particular, the quantifier *more* – the unit responsible for attaining a global representation – may not receive a (complete) semantic interpretation until the end of the sentence, where it encounters its complement clause and undergoes covert movement, and it is difficult to predict in advance what the content of the *than*-clause will include. This could mean that the parser proceeds more tentatively than usual with comparatives, or is less committed to any particular global reading, perhaps making reanalysis correspondingly easier. Additionally, due to perceptual noise or memory decay, there is a level of uncertainty already inherent in sentence processing, and sufficiently strong probabilistic bias away from the veridical analysis may cause the parser to adopt an analysis that is not wholly consistent with the actual input (e.g., Levy et al 2009, Levy 2011). In other words, comprehenders may adopt an event comparison analysis of the illusion sentence because this analysis is overall far more probable than a semantically meaningless one, and the unfaithfulness to the perceived input may be justified by the comprehender's uncertainty about whether they might have simply misread or misremembered the details of the matrix clause. Such uncertainty might be especially strong for comparative sentences in light of the complex

details of the syntax-semantics mapping, combined with decaying memory for the representational details of surface syntactic form (Potter & Lombardi 1990, Lombardi & Potter 1992) by the time the *than*-clause is encountered.

Under serial processing models, by contrast, presumably the complexity and scope of repair operations will determine the difficulty of recovery. For example, the "recycling hypothesis" of Arregui et al. (2006) proposes that the processer has mechanisms for reversing speech errors, thus enabling comprehenders to easily cope with (among other things) instances of mismatch in VP ellipsis; however, such error reversal is more costly when more operations are needed to recover the relevant VP. Perhaps then the relevant difference between costly syntactic reanalysis of difficult garden-path sentences, and the semantic repair implicated by the illusion data, is that the former necessitates changes to two levels of representation, and the latter only one. The reanalysis operations involved in processing Escher sentences also do not require major changes in the thematic structure of the sentence, a factor that has also been cited as important for determining the magnitude of slowdown.

In sum, Experiments 1 and 2 found that confronting a piece of linguistic input that mismatches a preferred analysis will cause observable processing difficulty, similar to well-known findings on garden path sentences; however, in the case of many garden-path sentences, this difficulty is quite obvious to comprehenders, while the semantic reanalysis that underlies Escher sentences appears to elicit subtle enough processing effects that comprehenders fail to report any knowledge of the problem at all. The results of Experiments 1 and 2 strongly rule out an uninterpreted heuristics approach to Escher sentences, with *all* illusion sentences eliciting some amount of slowdown irrespective of their offline detectability, while there are independent and nontrivial conceptual problems with an interpreted heuristics approach. The response profile associated with illusion sentences consists of two qualitatively different stages of processing difficulty, which I have suggested index anomaly detection and repair, respectively.

The goal of Chapter 4 is to investigate the second phase of this slowdown in more detail, by determining whether a shift to comparison of events is exclusively driven by the bottom-up ambiguity in the matrix clause associated with *many*. Although existing work has determined compellingly that Escher sentences may be perceived as a comparison of events,

we do not yet have reason to conclude that this is the *only* possible percept associated with Escher sentences. Chapter 4 will present a series of experiments demonstrating the flexibility of this repair, showing that there are at least three distinct percepts available to illusion sentences.

4 FLEXIBLE REPAIR OF ESCHER SENTENCES

4.1 Introduction

Chapter 3 found evidence in patterns of online processing that speak in favor of a repair account of Escher sentences. We also outlined a subtype of repair analysis – the event comparison hypothesis – which argues that illusion sentences are interpreted and the anomaly is detected, but that reanalysis to a comparison of events – similar to (2) – yields the perception of acceptability (Wellwood et al. 2009, 2012).

(1)   More people have been to Berlin than I have.

(2)   People have been to Berlin more than I have.

The event comparison approach speculatively connects the availability of the illusion to the fact that the grammar itself does not always make clear-cut distinctions between measurement of events and individuals on the basis of the syntactic position of the quantifier. The familiar example in (3), noted by Wellwood et al. (2009, 2012), has been analyzed as involving measurement of events (Krifka 1990) or event-object pairs (Doetjes & Honcoop 1997), in spite of the nominal position of the numeral quantifier. Other similar examples include (4), adapted from Westerstahl (1985)'s famous sentence *Many Scandinavians have won the Nobel Prize* (see also Herburger 2000)*,* which seems to involve restricting *many* with the VP – with the obligatorily distributive reading suggests that this happens only when the VP denotes a plurality of events (see e.g. Nakanishi 2007 for the connection between distributivity and VP plurality). Finally, crosslinguistic research suggests that floating numeral quantifiers appearing at the VP spine, such as *three* in (5), seem to involve measurement that depends at once on the plurality of the noun and the plurality of the verb (Nakanishi 2007), and are similarly sensitive to verbal plurality.

(3)   4000 ships passed through the lock.

(4)   Many STUDENTS lifted the piano.
         = Many of those who lifted the piano are students.   (??collective, √distributive)

(5)   Jungen haben drei ein Stuhl gebaut.   (??collective, √distributive)
         Boys have three a stool built.
         Three boys built a stool.

Examples such as these suggest that there is a systematic relationship between measurement of individuals and measurement of events; as a result, although a shift to an event comparison interpretation of the illusion is ultimately unfaithful to the details of its actual grammar, it is unfaithful in only a minimal way. An event comparison analysis can be obtained largely using the bottom-up ambiguity associated with the measure function *many*: reinterpretation of the problematic LF involves a shift to a grammatically licensed alternative reading of the matrix clause, so that the syntax of the matrix need not be significantly disturbed. This latter property is important since the illusion persists with unambiguously nominal quantifiers (e.g. *fewer* in Wellwood et al 2009; *as many* in Experiment 1), indicating that syntactic displacement of *more* is not a crucial component of the repair. The only small-scale changes needed to repair the illusion sentence involve the positing of a silent adverbial measure function in the *than*-clause, since the DP subject itself is not a suitable host.

The evidence thus far, however, does not distinguish between an account that specifically privileges event comparison versus one where repair can broadly apply to an illusion sentence whenever there is a suitable alternative host for the degree variable. These two hypotheses differ in the degree to which grammatically sanctioned ambiguity associated with the determiner is critical to the illusion: event comparison would be largely privileged in the case that the ambiguous matrix clause facilitates the illusion. If the illusion arises due to a more general repair mechanism that applies in the absence of such ambiguity, by contrast, then a broader range of percepts may be available. The conceptual appeal of these two hypotheses thus depends on the degree to which one wishes to constrain repair operations: on one hand, widespread misinterpretations are not the norm, and in most cases comprehenders seem quite able to easily point out significant problems with a sentence's logical form. On the other hand,

because language processing needs to be robust in the face of speech errors and imperfect input, a broad-scale repair mechanism may already be warranted independently from any illusion phenomena (see Frazier 2014 for discussion).

The goal of this chapter is to reduce the hypothesis space by focusing on the role of the other pluralities within the matrix clause, independent from the effects of event plurality. If event comparison is precluded, but the illusion sentence contains e.g. a bare plural direct object suitable for cardinality measurement, will the illusion persist? In other words, is it possible to repair an illusion sentence by shifting to cardinality measurement of a different noun phrase, or to repair an illusion sentence? If bottom-up ambiguity associated with *many* is critical to the illusion, then presumably only comparatives where the VP denotes a plurality of events can be repaired, regardless of the semantic number of the other elements. Experiments 3-4 answer this question by looking for evidence for sensitivity to object plurality absent a possible shift to event comparison, by contrasting offline acceptability judgments to illusion sentences where the semantics of the object noun phrase does versus does not support cardinality measurement. Experiment 5 looks more closely at the effects of subject plurality on illusion sentences, testing whether the robust increase in acceptability associated with these illusions persists even when it does not yield a distributive plurality of events. Finally, the goal of Experiment 6 is to investigate whether subject plurality effects are predicated on *morphosyntactic* versus *conceptual* plurality, a factor that can provide more information about the nature of the illusory percept.

4.1.1 *Dissociating comparison of individuals from comparison of events*

The goal of this chapter is complicated by the fact that nominal plurality often covaries with event plurality, making it difficult to determine whether in fact a comparison of events has been ruled out. Subject plurality tends to naturally yield a plurality of events, since, as discussed, in many cases each individual member of the plurality can be mapped to at least one event. The presence of a bare plural direct object also tends to strongly increase the salience of iterative readings. For example, while the predicates *watched the movie* and *watched movies* both allow for iterative interpretations involving multiple movie-watchings per individual, the iterative interpretation is more salient in the case of the bare plural (*watched movies more*),

since its meaning is broader and may involve events of rewatching the same movie over and over, or watching different movies each time. As a result, any finding that illusion sentences with predicates like *watched the movie* are more acceptable than those with predicates like *watched movies* could therefore be attributed to the increased salience of the event reading itself, rather than the direct object.

(6)   John watched <u>the movie</u> *more* than I did.

(7)   John watched <u>movies</u> *more* than I did.

In order to disentangle the properties of the event structure with those of nominal number, these experiments focus entirely on comparatives with stative predicates. Stative predicates are sometimes thought to differ from eventive predicates in lacking event structure altogether (e.g., Kratzer 1995, Katz 2000), in part because they describe scenarios that are difficult to anchor spatially or temporally. For example, Kratzer (1995) notes that constructions that necessarily modify an event, or bind an event variable – such as (8)-(10) – tend to preclude stative predicates like *know French.*

(8)   a. Julia knows French (*quickly, *on Fridays, *in the garden).
        b. Julie speaks French (quickly, on Fridays, in the garden).

(9)   a. *When Mary knows French, she knows it well.
        b. always [knows(Mary, French)] [knows-well(Mary, French)]

(10)   a. When Mary speaks French, she speaks it well.
        b. $always_e$ [speaks(Mary, French, e)] [speaks-well(Mary, French, e)]

Importantly, stative predicates also differ from eventive predicates in terms of their ability to support cardinality measurement, which requires plurality, corresponding to a semantically plural DP in the domain of individuals (Hackl 2001), or plural VP in the domain

of events (e.g., Nakanishi 2006, Wellwood et al, 2012). Merging *more* with a stative predicate such as *love* will typically generate measurement along a non-cardinality dimension, for example yielding a degree reading measuring the magnitude of *love* in (11), but will not generate readings comparing cardinalities of events. When a degree reading is unavailable, as is the case with *own* in (12), such adverbial comparison is ungrammatical.

(11)   Julia loves the book(s) more than Sarah does.
        (degree of love, *cardinality of *loving*-events)

(12)   *Julia owns the book(s) more than Sarah does.

Krifka (1990)'s event readings are not natural in sentences with stative or individual-level predicates either, which express inherent and permanent properties of individuals (Milsark 1974, Carlson 1977). In cases where a predicate cannot apply more than one time to an individual – either because such repetition is implausible (as in the case of eventive non-repeatable predicates like (13)) or because the predicate is defined over an individual's entire lifetime (as in the case of individual-level predicates, as shown in (14)) – event measurement would produce results that are indistinguishable from measurement of individuals. Thus, even if an event-measurement determiner could be posited inside the noun phrase this determiner would have no truth-conditional consequences.

(13)   4000 passengers graduated from law school.

(14)   4000 passengers have beards.

The other phenomena that straddle individual and event comparison are also clearly affected by the nature of the predicate. Nakanishi (2006), for example, discusses at length the fact that non-floated numeral quantifiers may co-occur with an individual-level predicate like *be smart*, while floated quantifiers cannot (Nakanishi 2006), cf (15)a-b. Similarly, whereas the focus-sensitive reading of *many* survives in (16)a, it is unavailable with the individual-level

predicate in (16)b (Herburger 2000).

(15) a. [Gakusei san-nin]-ga kono kurasu-de kasikoi (koto)

          [student 3-cl]-nom this class-in smart

      b. ??Gakusei-ga kono kurasu-de san-nin kasikoi (koto)

          Student-nom this class-in 3-cl smart

          '3 students in this class are smart'

(16) a. Many [chefs]$_F$ applied to the position.

          (= Many of those who applied are chefs)

      b. Many [chefs]$_F$ know how to make a souffle.

          (≠ Many of those who know how to make a soufflé are chefs)

Notice that event measurement readings are precluded in all of these cases in spite of the fact that the subject noun phrase is plural; thus, even if stative predicates do introduce an event argument, then it is clear that pluralizing the subject still does not yield a plurality of events suitable for cardinality measurement. As a result, a shift to comparison of events should be impossible with all illusion sentences containing stative predicates, regardless of nominal number; see (17)-(18).

(17) More teenagers have scary movies than the kids do.

      → ?? *than how much the kids have scary movies*

(18) More teenagers have scary movies than the kid does.

      → ?? *than how much the kid has scary movies*

Do Escher illusions actually disappear in this type of environment, as the event comparison hypothesis predicts? And if not, is repair sensitive to the presence of alternative plural noun phrases, in either subject and object position? The answers to these questions provide

information about the nature of the possible percepts of Escher sentences, and correspondingly, the breadth of possible repairs.

4.2   Experiment 3: Effects of object plurality on illusions

We begin investigating the effects of object plurality by comparing bare plural direct objects with and without dependent plural interpretations[8]. Descriptively, dependent plural readings arise when a bare plural noun phrases falls within the scope of another plurality, and the context strongly favors a bijective (one-to-one) mapping between the elements of the two sets. (19) is a classic example: since unicycles by definition have only one wheel, the dependent interpretation is the *only* plausible reading of the plural-marked NP *wheels*.

(19)   Unicycles have wheels.

These types of cases have long been cited as evidence for divorcing the role of morphological plural-marking from a "more than one" plural semantics (Chomsky 1975), since clearly, the plural morphology on *wheels* does not entail the existence of multi-wheeled unicycles.

A focus on dependent plurality allows us to determine whether or not a shift to direct object comparison is possible in Escher sentences. When the *than*-clause contains a singular subject, such as *I,* and a morphologically plural noun phrase, the theoretical availability of a shift to comparison of cardinality of individuals depends on whether that noun phrase can be interpreted as semantically plural or not, because there is no clausemate licensing plurality to facilitate a dependent interpretation. Either a semantic plural reading must be generated, in which case that noun phrase may serve as a host for the displaced degree variable, or else the direct object must be represented as singular at LF (*have a beard*), thus precluding cardinality

---

[8] Note that terminology with respect to this phenomenon typically reflects theoretical orientation towards it. If a bare plural argument like *wheels* is called a "dependent plural" (e.g. de Mey 1981) this implies that the noun is itself somehow ambiguous (as proposed by Chomsky 1975). Others claim that noun phrase denotations are the same in either case, but that the one-wheel-per-unicycle reading arises due to natural interactions with other elements in the sentence (e.g., Zweig 2008). I have no particular stake in this debate and thus will primarily refer to dependent plurality as a *reading* or *interpretation* of the noun phrase that is in some cases pragmatically obligatory.

measurement. As a result, in cases where the semantic plural interpretation is pragmatically odd, such as in (20), and unlike (21), a shift to direct object comparison would be either entirely ill-formed, or pragmatically odd:

(20)   More teenagers have **beards** than I do *have **beards***.
        a. semantically plural: # than wh_1 (*d_1-many) I have (#d_1-many) beards
        b. semantically singular: than wh_1 (*d_1-many) I have a (*d_1-*many*) beard

(21)   More teenagers have **friends** than I do *have **friends***.
        a. semantically plural: than wh_1 (*d_1-many) I have (d_1-many) friends
        b. semantically singular: than wh_1 (*d_1-many) I have a (*d_1-*many*) friend

The choice of this manipulation is motivated in part by the fact that the two predicates – e.g., *have beards* and *have friends* – contain superficially parallel noun phrases, in both cases bare plurals. These conditions cannot be differentiated on the basis of morphological cues, which a grammar-independent processing heuristic could feasibly target. It is not the case that one condition has a determiner while the other does not; or that one condition has an overt plural marker while the othedoes not. Nor can the conditions be differentiated on syntactic grounds – from the perspective of surface syntax, both direct object noun phrases are equally suitable hosts for the measure function *many*. Rather, the two conditions are differentiated *exclusively* at logical form, which constrains the possible interpretations of the predicate, yielding pragmatic anomaly in one case but not the other. In this sense, our manipulation of direct object plurality parallels the manipulation of event plurality (Wellwood et al. 2009, 2012; see also Chapter 3), but simply targets a different element of the sentence.

To summarize, in cases where world knowledge forces a one-to-one mapping between subject and object noun phrases, cardinality comparison of the object noun phrase is impossible. As a result, a shift to direct object comparison can be easily diagnosed: it will yield measurably lower acceptability for illusion sentences with dependent plural direct objects than semantic plural direct objects.

4.2.1   Methods

4.2.1.1   Materials & Design

We collected acceptability judgments for four conditions, crossing PRESENCE OF ILLUSION (anomalous illusion vs. normal control comparative) with OBJECT READING (dependent vs. semantic), as shown in Table 4. The control conditions, by virtue of their bare plural *than*-clause subject, allow for comparison of cardinalities of individuals provided by the subject noun phrase, and therefore should not be critically affected by the properties of the direct object noun phrase either way. In the illusion conditions, by contrast, the singular *than*-clause subject rules this possibility out, and also causes dependent readings to fail. The only possibilities are to terminate the parse as ungrammatical, or shift to a pragmatically illicit semantic plural reading of *tails* so that it can be used as a source for the degree variable. In the neutral context allowing a *semantic plural* reading, the latter option is pragmatically available, rendering a shift to object comparison a plausible interpretation.

	Dependent plural interpretation	Semantic plural interpretation
Illusion	More cats have <u>striped tails</u> than the dog does.	More cats have <u>mouse toys</u> than the dog does.
Control	More cats have <u>striped tails</u> than dogs do.	More cats have <u>mouse toys</u> than dogs do.

Table 4. Experiment 3 design.

The illusion conditions were created by substituting a singular noun phrase as the *than*-clause subject (e.g., *than my/the dog does*) in the place of the bare plural noun phrase (*than dogs do*). The *than*-clause subject and matrix subject were always semantically disjunctive, thus rendering subset comparison structures (such as *More cats have striped tails than (just) the tabby*; Grant, 2013) irrelevant. Determiner type was counterbalanced so that half of the items had possessive pronouns (*my/our*) and half had definite determiners (*the*). In order to disentangle the effects of direct object number from predicate repeatability, all 24 target items

had stative predicates, in most cases the verb *have*. The matrix predicate consisted of a verb and a bare plural object noun phrase, often preceded by a single modifier.

The dependent plural conditions were defined as those where the only plausible interpretation involved a bijective (one-to-one) mapping between the sets denoted by the subject and object noun phrases. For example, we know from world knowledge that each animal has one tail, each person has one birthday, and each movie has one title. Semantic plural conditions were defined as those plausibly allowing a one-to-many reading. For example, animals may have multiple toys, people may have multiple jackets, and movies may have multiple reviews. Thus, whereas the dependent plural condition forced a one-to-one mapping, the semantic plural condition merely allowed it; as a consequence, a shift to direct object comparison was allowed in the latter condition, but not the former.

In order to control whether the *than*-clause contained any alternative set of degrees, for each item, the modifiers on the direct object noun phrases (e.g. *striped* tail, *mouse* toy) were matched for gradability. In other words, for any given item, the modifiers were either both gradable or both non-gradable (the vast majority of items – 21 of 24 – contained non-gradable modifiers).

64 filler sentences were added to the 24 target items, consisting of a variety of grammatical and ungrammatical comparative and ellipsis constructions. The target items were masked by including a number of fillers with stative predicates, nominal quantifiers, and ellipsis sites. Filler comparatives included nominal and attributive comparatives, some with island violations or ellipsis irregularities. The ellipsis conditions contained ambiguities and various degrees of syntactic mismatch.

Four lists were constructed, and the conditions rotated in a Latin-square design so that each participant saw only one version of each item. The order of fillers and targets was held constant across each list.

4.2.1.2   Participants

28 participants were recruited from Mechanical Turk and were paid $0.75 for completing the experiment. All participants identified themselves as native adult English speakers with no known reading or language disorders. Because responses were collected

anonymously over the internet, screening procedures were employed to ensure participants read instructions and completed the task as required. First, all participants were required to respond correctly to simple commands ("Please assign this sentence a 4") and to rate filler sentences that were obviously ungrammatical or grammatical appropriately and on the correct end of the scale (i.e., ungrammatical sentences at the lower end of the scale, and grammatical sentences at the upper end of the scale). Second, in order to screen out participants who did not fully read the instructions and experimental items before responding, a lower cutoff time of six minutes was defined as the minimum allowable amount of time spent on the experiment (including the time to provide consent, read instructions, and respond to all stimuli). Of the 28 participants, four were excluded due to their responses to screening questions. Within the remaining data, an error in Qualtrics allowed participants to submit incomplete surveys, leading to six missing values (representing approximately 1% of the total data).

4.2.1.3   Procedure

The experiment was administered using Qualtrics. Participants were instructed to rate the acceptability of each sentence on a scale from 1-7, to "trust [their] gut reaction to the sentence, and rate it by taking into consideration how natural it sounds and whether [they could] imagine another English speaker saying it". The directionality of the scale was explained and example sentences containing varying degrees of island violations were shown to demonstrate graded acceptability.

4.2.2   Results

Prior to analysis, the ratings for each subject were normalized using the data from all of the target and filler stimuli. The results are summarized in Table 3.

	ILLUSION: *…than the dog does*	CONTROL: *…than dogs do*
DEPENDENT PLURAL: *More cats have striped tails…*	raw: 3.22 (1.64) z-score: -.36 (.78)	raw: 5.45 (1.59) z-score: .75 (.64)
SEMANTIC PLURAL: *More cats have mouse toys…*	raw: 3.95 (1.69) z-score: -.04 (.80)	raw: 5.25 (1.52) z-score: .64 (.75)

Table 5. Experiment 3: Mean (standard deviation) of acceptability ratings by condition.

The transformed ratings were then submitted to mixed effects regression models in R using the lme4 package in R (Bates et al, 2014). A maximal random effects structure was used by default, including random intercepts as well as random slopes for all fixed effects, both by item and by subject. As is evident from the descriptive statistics in Table 5, illusion sentences were rated substantially and significantly lower than control sentences (β = -1.11, t = -7.25; $\chi^2(1)$ = 27.46, p < .001). Semantic plurals were numerically worse than dependent plurals, though not significantly (β= -.09, t = -.98), and this effect was qualified by a significant interaction that rendered illusion sentences with semantic plurals better than those with dependent plurals (β = .43, SE = .16, t = 2.77; $\chi^2(1)$ = 6.86, p < .01). A density plot of the ratings distribution by condition in Figure 11 shows that illusions with dependent plural objects are more frequently rated at the lower end of the scale, relative to those with semantic plural objects.
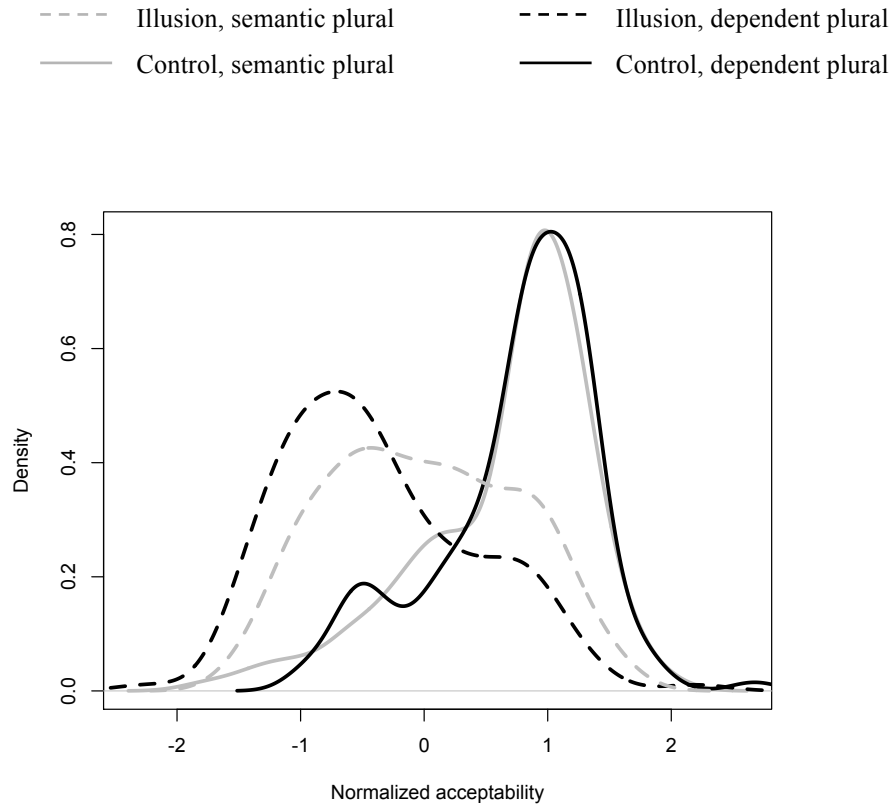
Figure 11. Experiment 3: Density plot of ratings by condition

To get a sense of how illusion sentences pattern in acceptability with respect to other sentence types, the mean z-scores for filler sentences are shown in Table 6. Descriptively, the illusion sentences were rated higher than obviously ungrammatical filler sentences containing island violations and ellipsis mismatch, and lower than grammatical sentences.

Filler type	z-score
Ungrammatical comparatives	-.74
(*Phillip has so much more homework this week than Laura is a student who has)	
Ellipsis mismatch	-.91
(*Many local banks offered the entrepreneurs a loan, and we were)	
Grammatical comparatives	.86
(The president had smarter aides than the vice president did.)	
Grammatical ellipsis	.46
(Many guests claimed the room had bedbugs and the manager did too.)	

Table 6. Experiment 3 filler ratings.

4.2.3   Discussion

The goal of this experiment was to assess whether the acceptability of an Escher sentence can be modulated by varying the plurality of its object noun phrase. More broadly, we sought to determine whether interpretation of such sentences depended solely on a shift to comparison of events, or whether it was also possible to shift to a comparison of cardinality of individuals, provided by a syntactically inaccessible noun phrase. To rule out the possibility of an event comparison interpretation we used all stative predicates with morphologically plural noun phrases, but varied whether or not the plural noun phrase was a semantically appropriate host for the displaced degree variable. If the acceptability of the anomalous illusion sentence were to be modulated by this noun phrase, we reasoned, this would suggest that event comparison is not the only possible repair, and correspondingly that the sentence may have multiple different percepts.

Although the position of *more* signals unambiguously that the cardinality measure function should be applied to the sets of individuals contributed by the subject noun phrases in the matrix and *than*-clause respectively, we found that illusion sentences were indeed sensitive to the plurality of the object noun phrase. However, it was not enough for that noun phrase to simply be morphosyntactically plural: repair was specifically facilitated by *semantically* plural

direct object noun phrases. This points to several conclusions. First, in line with prior findings suggesting relatively deep processing of illusion sentences, this experiment demonstrates that the relevant repair processes that lead to the acceptability of Escher sentences operate at LF. From the perspective of the surface syntax, any bare plural noun phrase argument is a suitable host for *more*, and yet only those predicates with specific semantic properties facilitate the illusion. Whether or not a dependent plural reading can be obtained depends not only on the interpretational properties of the predicate, but specifically those interpretational properties that arise due to syntactic dependencies among distinct lexical items. Thus, to the extent that comprehenders distinguish between semantic and dependent plural interpretations in the anomalous illusion sentences, they are clearly processing its grammar quite deeply. This provides yet another argument against superficial template-matching, where, for example, the illusion arises as a blend of the matrix clause of (22) and the *than*-clause of (23).

(22) Subject comparison: **More NPs V-ed NPs** than NPs did

(23) Object comparison: The NP V-ed more NPs **than the NP did**

(24) Illusion blend: **More NPs V-ed NPs** **than the NP did**

Second, our results implicate a broader-scale repair mechanism than has previously been proposed. Whereas the event comparison hypothesis focuses specifically on the role of the event structure in facilitating the illusion, these results suggest that verbal plurality is sufficient, but not necessary, to yield an illusion. In particular, comprehenders seem to be able to reinterpret the sentence by shifting to cardinality measurement of the object noun phrase within the *than*-clause (and possibly in the matrix clause as well). Importantly, although it is possible to shift to cardinality measurement of events in the matrix clause without displacing the measure function *many,* there is no clear way to shift to cardinality measurement of individuals provided by the object noun phrase without applying the measure function to a different constituent altogether. In other words, the illusion does not seem to be driven solely by the ambiguity associated with nominal *many*; Rather, there are (at least) two distinct available percepts for the *than*-clause, approximating the meanings in (25)-(26):

(25) –er (λd . the NP VPed d-much) …

(26) –er (λd . the NP has d-many NPs) …

One could object to our interpretation of the findings, however, on the basis of a potential confound in our design. Illusion sentences differed critically from non-illusion control comparatives in the plurality of the *than*-clause subject: the illusion sentences uniformly had singular *than*-clause subjects, whereas the control sentences uniformly had plural *than*-clause subjects. In theory, this itself could disproportionately affect the acceptability of sentences with dependent plural predicates due to complications with ellipsis resolution. In the case of the illusion sentences, the matrix predicate *have striped tails* cannot be combined with the singular *than*-clause subject on the dependent reading (i.e., # *the dog has striped tails*), because there is no higher plurality to license it. Maximally faithful ellipsis resolution will yield a pragmatically anomalous semantic plural reading, as shown in (27)b.

(27) a. # The cats have striped tails, and the dog has striped tails too.
      b. # More cats <u>have striped tails</u> than the dog does <u>have striped tails</u>
         |_____|

This problem could be avoided by making a small change to the ellipsis site – by changing the direct object to its singular form, as shown in (28) – but this would involve a mismatch between the matrix and ellipsis site, which could also affect acceptability ratings.

(28) a. The cats have striped tails, and the dog has a striped tail too.
      b. More cats <u>have striped tails</u> than the dog does <u>have a striped tail</u>
         |_____|

By contrast, in the control conditions, the plural *than*-clause subject will license dependent readings already, rendering any ellipsis mismatch in (29) unnecessary:

(29)   a. The cats <u>have striped tails</u>, and the dogs <u>have striped tails</u> too.

        b. More cats <u>have striped tails</u> than the dogs do <u>have striped tails</u>

               |_____|

In other words, it is possible that the pattern of results we observed in Experiment 1 is wholly related to the way the singular *than*-clause subject affects ellipsis resolution rather than the way it affects degree abstraction and interpretation of the illusion's anomaly. This type of account would predict that predicate type should equally affect the normal ellipsis constructions in (30):

(30)   a. The cats have striped tails and the dog does too.

        b. The cats have striped tails and the dogs do too.

      The acceptability of mismatch in ellipsis is the subject of much ongoing research, but it has long suggested that morphological number mismatch does not affect ellipsis in the way described above, and it is sometimes simply noted that there is no difference in acceptability between sentences like (30)a-b (see, e.g. Sag 1976 p. 143; Chomsky 1965, p. 180). In fact, the relative acceptability of sentences like (30)a is typically viewed as evidence for ellipsis resolution at logical form, where the dependent plural is interpreted as underlyingly singular – this would yield no mismatch. Thus, we have strong reasons to believe that the interaction observed in Experiment 3 was not actually driven by the confound described above. Nevertheless, because the effects we observe are relatively subtle it is possible that any contrast in (30) is subtle enough that comprehenders cannot reliably detect whether there is a difference using conscious introspection. This study is the first to test this judgment experimentally, in order to determine whether in fact ellipsis mismatch of the type described incurs some penalty in acceptability.

4.3   Experiment 4: Broader effects of object plurality on ellipsis resolution

The goal of Experiment 4 is to replicate the results of Experiment 3 using ellipsis constructions as controls. We expect that the results from Experiment 3 are tied to the presence of the illusion, as opposed to mismatch in ellipsis resolution, which leads us to predict an interaction such that the predicate type will have a significant effect on illusion sentences but not on the analogous ellipsis controls.

4.3.1   Methods

4.3.1.1   Materials & Design

Acceptability ratings were again collected for 24 target items, crossing presence of ILLUSION (control vs. illusion) with PLURALITY TYPE (semantic vs. dependent). However, unlike previous experiments, the control sentences in this experiment were not comparatives but VP ellipsis constructions, as shown in Table 7.

	Dependent plural	Semantic plural
Illusion	More cats have <u>striped tails</u> than the dog does.	More cats have <u>mouse toys</u> than the dog does.
Control	The cats have <u>striped tails</u>, and the dog does too.	The cats have <u>mouse toys</u> and the dog does too.

Table 7. Experiment 4 design

The illusion sentences for Experiment 4 were adapted entirely from Experiment 3. The only change was that the determiner type was not counterbalanced – instead, all of the items contained the definite article *the.* The control sentences contained the same predicates as the illusions and were always of the form *The* NP*s* VP*ed, and the* NP *did too.* Because the controls do not closely match the illusion sentences in structure, we were primarily interested in

replicating patterns of acceptability within illusions and within controls from Experiment 3, rather than looking at differences in acceptability between illusions and control sentences.

48 filler sentences were added to mask the items in this experiment, yielding a 2:1 filler-to-target ratio. Most of the fillers were adapted from Experiment 3, but were changed to fit the distribution of items in this experiment. Half of the filler items consisted of ellipsis constructions and the other half were comparatives of various types, ranging (roughly equally) from fully ungrammatical to fully grammatical.

The procedure and instructions were identical to Experiment 3.

4.3.1.2   Participants

27 participants from Mechanical Turk were paid $0.75 to complete the experiment. Screening procedures were employed as described above. 3 participants were excluded from the analysis because they spent under six minutes on the experiment (average completion time 4m 18s, compared to 15m 19s for the remaining participants). The remaining 24 participants were included in the analysis.

4.3.2   Results

As before, acceptability ratings were normalized using the ratings for all experimental items. The results are summarized in Table 8 below:

	ILLUSION: *…than the dog does*	CONTROL: *…and the dog does too*
DEPENDENT PLURAL: *(More/The) cats have striped tails…*	raw: 3.44 (1.37) z-score: -.28 (.66)	raw: 5.28 (1.59) z-score: .67 (.66)
SEMANTIC PLURAL: *(More/The) cats have mouse toys…*	raw: 3.83 (1.63) z-score: -.09 (.72)	raw: 5.15 (1.54) z-score: .60 (.62)

Table 8. Experiment 4: Mean (standard deviation) of acceptability ratings by condition.

The nature and directionality of effects is similar to those in Experiment 3: the control sentences are again rated substantially higher than illusion sentences, and the dependent plural condition is worse than the semantic plural condition in illusion sentences only.

The data were submitted to mixed effects linear regression models with fixed effects for *illusion, plurality type,* and their interaction. A maximal random effects structure was used with random intercepts to model variation across subjects and items, and random slopes for the effect of *illusion, plurality type,* and their interaction, both by items and by subjects. The descriptive patterns suggested by the data above were confirmed by the model: the presence of the illusion was a significant predictor of acceptability, with illusions significantly worse than controls (β = -.95, t = -6.7; $\chi^2(1)$ = 24.31, p < .001). Semantic plurals were not overall better than dependent plurals (actually the opposite: β = -.08, t = -1.03), but there was a significant interaction reflecting the fact that the semantic plural conditions were better than dependent plural conditions in the illusion sentences only (β = .27, t = 2.4; $\chi^2(1)$ = 5.32, p < .05). A density plot in Figure 12 shows that control sentences patterned more or less analogously, while illusion sentences with dependent plural objects tended to receive more ratings at the lower end of the scale than those with semantic plural objects.
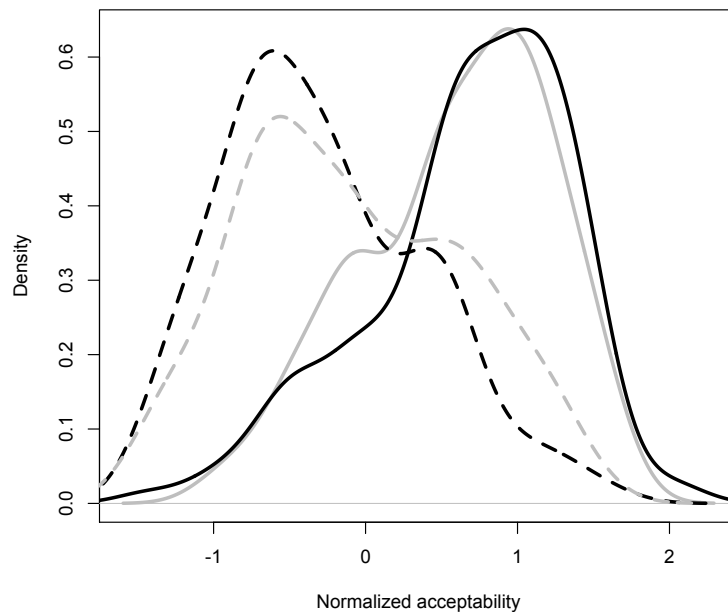
Figure 12. Experiment 4: Density plot of acceptability ratings

4.3.3   *Discussion*

The purpose of Experiment 4 was to investigate whether the general patterns found in Experiment 1 could be replicated with a set of controls matched for the number of the second noun phrase (i.e., *than **the dog** does; and **the dog** does too*). The number of this noun phrase has a direct impact on the acceptability of the ellipsis site: a plurality like *tails* would need to be changed to singular *a tail* upon reconstruction, as shown in (31), since there is no other c-commanding plurality within the clause to license a dependent reading. Noun phrases that can have semantic plural interpretations do not require such changes, as shown in (32).

110

(31)   More cats have striped tails than the dog does {# <u>have striped tails</u>, <u>have a striped</u>
        <u>tail</u>}

(32)   More cats have mouse toys than the dog does <u>have mouse toys</u>.

        Various theoretical proposals already rely on the intuitive judgment that there is no

substantive contrast between the acceptability of sentences like (31) and (32). Such facts have

been interpreted as evidence that the mechanism underlying ellipsis resolution is insensitive to

morphological number; however, this pattern has never been verified experimentally.

(33)   The cats have striped tails, and the dog does __ too.

(34)   a. The cats have mouse toys, and the dog does __ too.
        b. The cats have striped tails, and the dogs do __ too.

Experiment 4 confirmed that this factor does not cause a decline in acceptability in normal

control sentences like (33); reconstruction of the VP at the ellipsis site does not seem to be

degraded in cases where morphologically plural *tails* needs to be reconstructed as singular *tail*.

        Plurality type, however, does continue to exert significant influence on anomalous

illusion sentences: as in Experiment 3, illusion sentences in Experiment 4 were rated

significantly lower when they contained plural direct objects with an obligatory dependent

reading. This suggests that repair of Escher sentences can go beyond a simple shift to event

comparison: if event comparison were the only available repair, then all sentences in this

experiment should have been equally unacceptable, given that stative predicates cannot support

cardinality measurement, as shown in (35)-(36).

(35)   # The dog has mouse toys more than the cat does.

(36)   # The dog has striped tails more than the cat does.

Instead, we found that in contexts where event comparison is ruled out, the parser shows

sensitivity to other pluralities in the predicate that have the right semantic properties to support

111

cardinality measurement. Specifically, illusions with a predicate containing a semantically plural object noun phrase again fared better than those with a syntactically plural but semantically singular object noun phrase: presumably, only the former can host the displaced degree variable. Thus, the results from Experiments 3-4 collectively suggest that it is possible to construe Escher sentences as a comparison of cardinalities of individuals provided by a noun phrase in a different syntactic position. In the general discussion we discuss the possible LF configurations that could lead to the available percepts for Escher sentences.

4.3.3.1   Flexible repair: relevance to subject number

The findings from Experiments 3-4 suggest that Escher sentences have multiple distinct percepts: one approximating a comparison of cardinality of events and one approximating a comparison of cardinality of individuals. Experiment 5 addresses whether the availability of the latter could explain why illusion sentences with plural *than*-clause subjects, such as (37)-(38) were rated so highly in Experiment 2, even when the predicate was not iterative, as in (38).

(37)   More lawyers vacationed in Florida than the judges did.

(38)   More lawyers retired to Florida than the judges did

One possibility that we entertained was that such cases do naturally allow for event comparison readings – even if not iterative – since each individual within the plurality can be mapped distributively to a different event of *retiring* (Wellwood et al 2009, 2012). The cardinality of the sum of such events can then be measured to arrive at an event comparison reading. Alternatively, given that it is apparently possible to construe Escher sentences as a comparison of individuals, we might wonder whether comprehenders are able to compare the number of individuals provided by the two respective subject noun phrases, *lawyers* and *the judges*. In order to do so, the problematic determiner in the *than*-clause would need to be dealt with, presumably either by omitting it, (39)a, or by reinterpreting the subject noun phrase partitively, (39)b.

(39)   a. More lawyers retired to Florida than ~~the~~ judges did.

       b. More **of the** lawyers retired to Florida than the judges did.

This question speaks to the breadth of possible repairs. The cumulative findings so far suggest that anomalous illusion sentences may be flexibly reinterpreted by applying the cardinality measure function to an alternative plurality, whether it corresponds to a plurality of events or a plurality of individuals provided by a different noun phrase. Repairs of the sort outlined in (39) are fundamentally of a different type – instead of abandoning measurement of the subject noun phrase and applying the measure phrase to a different constituent altogether, they involve revisions to the subject noun phrases so as to be amenable to cardinality measurement: the percept shown in (39)a is consistent with the veridical matrix clause, but would involve changes at the anomalous subject noun phrase in the *than*-clause; while the percept shown in (39)b retains a veridical analysis of the overt material in the *than*-clause but would involve retrospective changes to the matrix clause subject.

     To determine how sentences like (39) are construed, Experiment 5 tests whether the increase in acceptability associated with plural *than*-clause subjects is critically related to its effects on the event structure, and thus whether repair operations are fundamentally limited to applying the measure function to other available pluralities. If repair is constrained in this way, subject plurality should not influence acceptability ratings when the illusion sentence contains an individual-level predicate, since in such there is no plurality of events suitable for cardinality measurement, regardless of the plurality of the subject. In other words, both of the sentences in (40) should be equally incomprehensible.

(40)   a. More chefs know how to make a soufflé than **the line cook** does.

       b. More chefs know how to make a soufflé than **the line cooks** do.

In addition, both types of illusion sentences should be similarly sensitive to object plurality, since that would provide the only other route to repair when a shift to event comparison is ruled out. A secondary goal of Experiment 5 is therefore to determine whether the plurality of

the object noun phrase affects illusions with plural *than*-clause subjects in the same way that it affects those with singular *than*-clause subjects.

4.3.3.2   Bijective readings in the context of counting

In order to test sensitivity to an alternative repair to a comparison of plural objects, Experiment 5 again contrasts illusion sentences with object noun phrases with and without pragmatically dependent plural interpretations, as shown in (41)-(42). Only the latter can support a shift to comparison of cardinalities of individuals denoted by the object noun phrase:

(41)   a. More cats have striped tails than the dogs do.
        b. # Cats have more striped tails than the dogs do.

(42)   a. More cats have mouse toys than the dogs do.
        b. Cats have more mouse toys than the dogs do.

In order to clarify why the repair in (41)b is unavailable, we must first explore some properties of dependent plurality and the way it interacts with cardinality modification. Note that, in theory, the subject of the *than*-clause contains the right properties to license dependent readings, since it is a plurality that c-commands the object noun phrase. So the infelicity of (41)b cannot be traced to properties of the *than*-clause subject; rather, regardless of the availability of a licensing noun phrase, it appears that a one-cat-to-one-tail reading simply cannot be generated when the cardinality measure function merges directly with the object noun phrase (*more/many/four* NPs). For classical accounts of dependent plurality such as Chomsky (1975), this is a natural consequence of the fact that dependent plural interpretations involve plural marking on a noun that is inherently singular: a dependent reading should then by definition be semantically incompatible with any cardinality exceeding one. This is even more clearly the case for comparatives, given that the theoretical reasons to believe that the measure function inherent in *more* combines only with pluralities (Hackl 2001); thus, an analysis that presupposes dependent plural nouns are not really plural naturally predicts the repair in (41)b to be impossible.

It should be briefly noted, however, that not all theories of dependent plurality agree that such noun phrases are interpreted as semantically singular; various modern implementations assume that plural number morphology is semantically active in such cases, and the "multiplicity" requirement associated with the plural morphology is satisfied cumulatively (de Mey 1981, Roberts 1991, Beck 2000, Zweig 2008); in other words, dependent plural interpretations are simply a subtype of cumulativity, where there is a bijective construal. Collapsing dependent plural and cumulative interpretations is a conceptually appealing move because, as Zweig (2008) notes, they share a crucial semantic property in common: in both cases the existential force of the noun phrase is distributed over, but its number specification is not. In both of (43)-(44), each student must have visited one school, but possibly no more than that. In the case of the "dependent plural" interpretation in (43), the number morphology additionally requires that the *overall* cardinality of schools visited is *more than one*; and in the case of the cumulative interpretation in (44), it requires that the *overall* cardinality is *eight.* These strong semantic parallels suggest that (43), and other similar instances of bijective plural readings, can in fact be categorized as regular plurals that receive a cumulative interpretation.

(43)   a. The prospective students visited **public schools** in California.

b. Each student visited at least one public school, *and* more than one school was visited overall.

(44)   a. The prospective students visited **eight public schools** in California.

b. Each student visited at least one public school, *and* eight public schools were visited overall.

The theoretical parsimony of assimilating cumulative and dependent plural readings is obviously undermined by their different distributions in (45)-(46), where the *overall more than one* (dependent) interpretation is available but the *overall n* (cumulative) interpretation is not. In other words, dependent plural readings may not be a subtype of cumulativity, since these

readings survive in a wider range of environments[9].

(45)   The cats have striped tails.

Intended: Each cat has at least one striped tail, *and* there is more than one tail overall.

(46)   * The cats have eight striped tails.

Intended: Each cat has at least one striped tail, *and* there are eight tails overall.

The unacceptability of (46) seems to be related to the fact that predicates like *have wheels* or *have tails* are stubbornly distributive (Schwarzschild 2009): predicates that do not require distributive readings – such as *buy n unicycles* or *like n teachers* in (47)-(48) – allow (but do not require) a bijective reading to arise through their cumulative semantics. But predicates that are obligatorily distributive – such as *have n beards* in (49) – fail to license bijective readings, presumably because a cumulative reading is fundamentally unavailable in those cases. The bare noun phrase, meanwhile, can apparently receive a bijective interpretation regardless of the details of the predicate.

(47)   a.   John, Bill, Fred, and Mark bought four unicycles.

            bijective OK: one unicycle per person

        b. $\exists$e: *buy(e) & agent(e) = J+B+F+M & $\exists$y: theme(e)=y & *unicycle(y) & |y|=4

---

[9] The observation that cumulative and dependent readings are not distributionally equivalent is not new – for example, *most, all* and *both* are also known to license dependent but not cumulative readings (a fact discussed in Zweig 2008 and elsewhere):

(i) Most/all/both men bought unicycles.   (dependent)
(ii) Most/all/both men bought four unicycles.   (*cumulative)

However, there exist accounts that explicitly take these facts into account (Champollion 2010), while I am not aware of any account that explicitly addresses the contrast in (45)-(46).

(48)   a.   John, Bill, Fred and Mark like four teachers.

           bijective OK: one teacher per person

       b. $\exists$e: *like(e) & experiencer(e) = J+B+F+M & $\exists$y: theme(e)=y &

           *teacher(y) & |y|=4

(49)   a.   # John, Bill, Fred, and Mark have four beards.

           *bijective: one beard per person

       b.   # $\exists$e: possession(e) & recipient(e)=J+B+F+M & $\exists$y: theme(e)=y &

           beard(y) & |y|=4

While an explanation of the relationship between cumulativity, dependent plurality, and distributivity is beyond the scope of this dissertation, the fact that bijective readings of plural noun phrases can be obtained in cases where cumulative readings of numerically quantified noun phrases cannot may serve as evidence in favor of dissociating the two phenomena, although this conclusion has no particular bearing on the goals of this experiment[10].

       To summarize, I have outlined an account that attributes the acceptability of illusion sentences with plural *than*-clause subjects to a shift to event comparison, and contrasted it with an account that allows for a comparison of sets of individuals contributed by the respective subject noun phrases. I have argued that a shift to event measurement is likely to be independently precluded when the predicate is stative; therefore, if comprehenders still perceive plural subjects to be better than singular ones, this would implicate a repair of another sort. By contrast, if plural subjects are better than singular subjects because of their specific

---

[10] Although Zweig (2008) suggests that the *more than one overall* meaning associated with dependent plurality is a problem for theories that posit singular or number-neutral meanings for dependent plural noun phrases, it seems to follow straightforwardly if their interpretation is analogous to a weak existential singular that is necessarily distributed over, i.e. *Unicycles each have a wheel*. In other words, there is one wheel per unicycle, and more than one unicycle, so it follows that there is more than one wheel overall, regardless of whether or not this is stipulated in the semantics. Correspondingly, (i) also seems to have a "Multiplicity Requirement" when the object is understood as nonspecific/weakly quantified, even though the object is not marked as a plurality:

(i)   The kids each like a dog.
        ?? there are three kids and one dog, and each kid likes the same dog
        OK: there are three kids and two dogs, and each kid likes one of the dogs

relationship to event structure, then the illusion sentences with plural subjects should pattern completely with illusion sentences with singular subjects in this experiment, in terms of acceptability ratings and sensitivity to direct object plurality. Experiment 5 tests these predictions by using the paradigm from Experiments 3-4 to compare reactions to illusion sentences with singular and plural subjects.

4.4   Experiment 5: Effects of subject plurality

4.4.1   Methods

A 2 x 2 within-subjects design tested the effect of PLURALITY TYPE (semantic versus dependent) and *THAN*-CLAUSE TYPE (*than the NP does* vs. *than the NPs do*), as shown in Table 9. Because Experiments 3-4 already established that plurality type has no general effect on non-anomalous control sentences, control conditions were omitted and all target stimuli were anomalous illusion sentences.

	Dependent plural	Semantic plural
Singular	More cats have <u>striped tails</u> than the dog does.	More cats have <u>mouse toys</u> than the dog does.
Plural	More cats have <u>striped tails</u> than the dogs do.	More cats have <u>mouse toys</u> than the dogs do.

Table 9. Experiment 5 design

All stimuli for this experiment were adapted from Experiment 1, except that the target items always had a definite determiner preceding the *than*-clause subject. The 24 target items were combined with 48 filler items consisting of ellipsis and comparative constructions of varying acceptability.

30 participants were recruited from Mechanical Turk for the experiment and were paid $0.75 for their participation. Data rejection procedures were adopted from prior experiments, with the minimum time to experiment completion again defined as six minutes (global average time to completion in this experiment was 15:48). Three participants spent less than six

minutes on the survey (mean time to completion: 4:07), and two used the scale in the wrong direction (i.e. assigning high ratings to ungrammatical screener sentences). The data from the remaining 25 participants were used in the analysis.

4.4.2 Results

Prior to analysis the ratings were normalized using each participants' responses to all items, including fillers. The raw and transformed ratings are shown in Table 10:

	SINGULAR *THAN*-CLAUSE: …than the dog does	PLURAL *THAN*-CLAUSE: …than the dogs do
DEPENDENT PLURAL: *More cats have striped tails…*	raw: 2.93 (1.60) z-score: -.43 (.71)	raw: 4.70 (1.66) z-score: .47 (.78)
SEMANTIC PLURAL: *More cats have mouse toys…*	raw: 3.45 (1.63) z-score: -.16 (.74)	raw: 4.71 (1.65) z-score: .49 (.77)

Table 10. Experiment 5: Mean (standard deviation) of acceptability ratings by condition.

The data, plotted by density in Figure 13 were analyzed using a mixed-effects linear regression model with fixed effects for *subject plurality* (singular vs. plural) and *object plurality* (singular vs. plural) and their interaction, and a maximal random effects structure was included with random intercepts by subjects and items and random slopes to model variability in all fixed effects across subjects and items. The model confirmed that there was a strong effect of *than*-clause subject plurality, with clauses like *than the NPs did* rated higher than *than the NP did* (β = -.64, t = -4.93; $\chi^2(1) = 30.94$, p < .001). The general effect associated with object number was not significantly different from zero (β = -.02, t = -.22) even though its addition to the model yielded a significantly better fit ($\chi^2(1) = 5.41$, p < .05); this is likely due to a possible interaction between subject and object plurality, such that object plurality modulated acceptability of illusions with singular *than*-clause subjects but not plural ones (β = -.26, SE = .14, t = -1.87; $\chi^2(1) = 3.36$, *p* = .07).

119

——— Plural subject, semantic plural object

- - - Singular subject, semantic plural object

——— Plural subject, dependent plural object

- - - Singular subject, dependent plural object
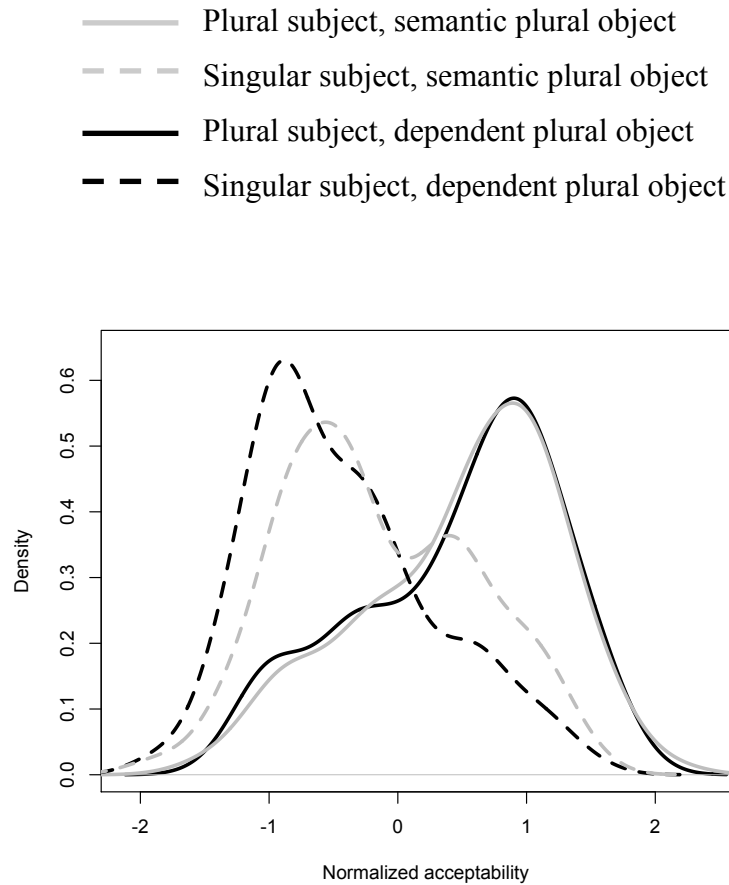
Figure 13. Experiment 5: Density plot of acceptability ratings

4.4.3   *Discussion*

The goal of this experiment was to investigate why illusions with plural *than*-clause subjects are rated higher than those with singular *than*-clause subjects. Two competing hypotheses speak to the range of ways Escher sentences may be construed: one possibility is that subject plurality facilitates event comparison, while another possibility is that it enables comparison of cardinalities of individuals provided by the two respective subject noun phrases:

(50)   Lawyers retired to Florida more than the judges did.

(51)   a. More lawyers retired to Florida than judges did.

    b. More of the lawyers retired to Florida than the judges did.

    The original account proposed by Wellwood et al (2009) ties subject plurality to its effects on event structure: in most cases, merging a plural subject will make available a plurality of events where each individual is mapped distributively to one event. The sum of these events can then be measured to arrive at an event comparison interpretation, similar to (50). This type of account might privilege event comparison as a potential repair due to ubiquitous mappings between events and individual measurement, making the syntactically adnominal position of *more* less problematic, and potentially requiring only minimal changes to the LF of the illusion sentence. By contrast, the alternative hypothesis posits a broader range of percepts for illusion sentences, along with a larger number of operations that can repair the problematic grammar. If illusion sentences with plural *than*-clause subjects are interpreted as a comparison of individuals, such as (51), then repair is not limited to simply applying the measure function to an alternative available plurality. In particular, the measure function may remain within the subject noun phrase, but the repair must then somehow deal with the problematic determiner, since *many* cannot otherwise merge with a full DP; in other words, repair operations could change the composition of the noun phrase itself to fit the measure function, rather than changing the position or role of the measure function to apply to other available pluralities.

    Crucially for this experiment, the event comparison approach predicts special sensitivity to predicate type where the alternative does not. Stative predicates are unlikely to provide a sum of events that can be measured even when the subject is plural, leading us to expect similar patterns of acceptability for illusions with singular and plural *than*-clause subjects, since neither can support event measurement. By contrast, if reanalysis depends solely on the two subject noun phrases, then the nature of the predicate is wholly irrelevant to the acceptability of the illusion. The results of this experiment provide support for the latter hypothesis: illusion sentences with plural *than*-clause subjects clearly do not behave like their counterparts with singular *than*-clause subjects, even when the effects of nominal plurality on event structure are controlled. In particular, they continue to be perceived as substantially more acceptable, and are wholly insensitive to the nature of the predicate and any pluralities it

contains. This is a logical outcome in a scenario where repair does not depend in any way on the predicate – namely, in a scenario where the cardinality of the respective subjects can be compared. This interpretation is, of course, strictly precluded by the grammar of the comparative, because there is no appropriate location within the full DP *the dogs* to host the degree variable, and the cardinality measure function requires an <e, t> predicate of plural individuals:

(52)   * How many the dogs have striped tails?

However, there are various possible ways to generate a comparison of individuals. One way is to simply omit the definite article, and interpret the comparative as if it contained two bare plural subject noun phrases[11]:

(53)   How many ~~the~~ dogs have striped tails?

(54)   a. −er (λd . d-many dogs have striped tails)(λd. d-many cats have striped tails)

b.

```
            <e, t>
          /        \
      <e, t>       <e, t>
     d-many       dogs/cats
```
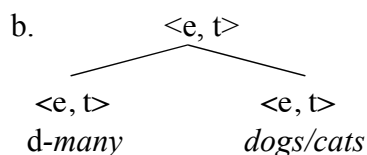
Another way is to reinterpret the matrix clause as if it also contained a definite DP, with measurement of both DPs made possible through the introduction of partitive *of,* which creates a predicate of individuals out of a type *e* denoting DP (Ladusaw 1982, de Hoop 1998, Schwarzschild 2002):

(55)   How many <u>of</u> the dogs have striped tails?

---

[11] Here for expository purposes we are assuming a non-quantificational meaning of *many,* although this particular choice does not critically affect the reasoning in this section.

(56)   a. –er (λd . d-many of the dogs have striped tails) (λd. d-many of the cats have striped tails)

b.
```
                    <e, t>
              /              \
         <e, t>              <e, t>
         d-many          /          \
                  <e, <e, t>>           e
                      of          the dogs/cats
```
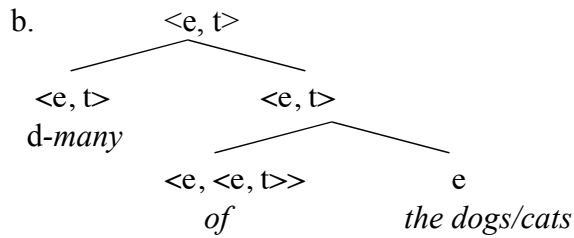
There is no immediately obvious advantage for one repair over another: both involve making relatively small-scale changes to the grammar of the noun phrase to find a location for the measure function and degree variable. However, whereas the former requires making changes to the overt material in the *than*-clause at the site of the anomaly, the latter is a retrospective change that could arise during antecedent retrieval.

Experiment 6 is designed to distinguish between percepts that would involve partitive versus non-partitive comparison of individuals. Non-partitive comparison of individuals is more constrained in that it requires a noun phrase that is both semantically and grammatically plural; by contrast, partitive comparison is possible so long as the relevant individual has contextually salient atomic parts, regardless of whether these parts are grammatically accessible. Thus whereas plural morphology is necessary in (57), it is only optional in (58).

(57)   Sally read more book(*s) than Jack did.

(58)   Sally read more of the book(s) than Jack did.

Experiment 6 looks at whether comparison of the subject noun phrases is still possible when the *than*-clause subject is notionally but not grammatically plural. We use group nouns such as *committee* as the relevant diagnostic, since these nouns famously exhibit properties of both singular and plural nouns (Landman 1989, Barker 1992, Schwarzschild 1996). Much of the confusion about the status of these nouns is due to the fact that they often, but not always, combine with the same types of predicates as true pluralities. With collective and distributive plural predicates in (59), group nouns and pluralities give rise to basically equivalent readings.

123

Yet (60) shows that some properties may apply to pluralities but not groups; and (61) shows that some properties may apply to groups without applying to pluralities.

(59) a. The committee met. = The members of the committee met.
       b. The committee is optimistic about the plan. = The members of the committee are optimistic about the plan.

(60) The members of the committee fathered two children. ≠ The committee fathered two children.

(61) The committee is old. ≠ The members of the committee are old.

Finally, as Pearson (2011) notes, in partitive constructions the group noun can behave semantically as a true plurality, since the uppermost determiner quantifies over its salient atomic parts, which in the case of a group noun like *team* correspond to the individual team members. This renders the meanings of (62)-(63) more or less equivalent.[12]

(62) a. Half of the team members are in the yearbook photo.
       b. More of the football players are in the yearbook than the tennis players.

(63) a. Half of the team (is/are) in the yearbook photo.
       b. More of the football team (is/are) in the yearbook than the tennis team.

In spite of the similarities between group nouns and pluralities, group nouns are still not grammatically suitable direct arguments for the measure function *many* (see Hackl 2001 as discussion therein). The ill-formedness of (64) tells us that comparison of cardinality of

---

[12] My judgment is that a masslike reading is also available in (63); for example, a very unskilled photographer only manages to capture the lower half of the team's bodies in the yearbook photo; however, this reading is clearly less salient.

individuals requires a true bare plural, not just a conceptually plural group noun. Failing that, cases like (65) tell us that partitive *of* may be introduced into the structure to make the salient parts of a group noun denotation available for cardinality measurement:

(64)   a. * More than two committee are meeting.
         b. * More trio than quartet was/were in the room.

(65)   More of the trio was in the room than the quartet.

If illusion sentences with plural *than*-clause subjects sound highly acceptable because listeners are tempted to simply ignore the problematic determiner, then we expect illusion sentences with a grammatically singular but conceptually plural group noun (*team*) as a subject to be no better than those with a grammatically and conceptually singular count noun (*team member*). On the other hand if a shift to a referential partitive is an available operation, the group noun may indeed be perceived as more acceptable than an ordinary singular count noun, given that its atomic parts can be made available for measurement in spite of the fact that it is not a "true" plural in the relevant sense for *many*.

4.5   Experiment 6: Effects of collectivity

4.5.1   Methods

A 2 x 2 within-subjects design tested the effect of THAN-CLAUSE SUBJECT (group noun versus singular count) on illusion and control sentences. Because comprehenders are highly sensitive to parallelism between the matrix and *than*-clause (Carlson 2001), we suspected that the parallelism between the two subject noun phrases might modulate acceptability independently from the illusion, with participants possibly finding pairs like *passengers-flight crew* to be more felicitous than e.g., *passengers-flight attendant.* In order to test the specific effects of the group noun on the illusion sentences over and beyond its effects on clausal

parallelism, VP ellipsis controls were again employed, so that the parallelism between the two subjects was kept constant across illusion and control conditions.

	Illusion	Control
Singular	More foreign-born diplomats are familiar with the trade agreement than the **American citizen** is.	The foreign-born diplomats are familiar with the trade agreement, and the **American citizen** is too.
Group	More foreign-born diplomats are familiar with the trade agreement than the **American population** is.	The foreign-born diplomats are familiar with the trade agreement, and the **American population** is too.

Table 11. Experiment 6 design

In the group noun conditions, the subject of the second clause contained a morphologically singular noun denoting a collection of individuals, such as *cast, gang,* or *tribe.* Only nouns that could optionally take plural morphology were used, in order to distinguish between singular count nouns with salient atomic parts (such as *crew*), and mass nouns with atomic parts (e.g., *furniture, fruit, jewelry,* which cannot be pluralized). The noun pairs (*flight attendant – flight crew*) were always the same length, and were usually in a *member-of* relation with one another, i.e. a flight attendant is a member of the flight crew; a gangster is a member of a gang, and a tribesman is a member of a tribe. In a few cases, the singular count noun was not a member of the group noun but a counterpart to it, i.e. *jury/judge.*

Predicate type was counterbalanced across items, partly to control for the availability of alternative repairs, and partly as an exploratory measure. Half of the items contained predicates that were entirely non-gradable and had no possible alternative host for the degree variable (*More musicians at the competition are classically trained than the choir is*), so that the only way to interpret the illusion sentences would be as a comparison of individuals. The other half of the items contained gradable adjectives which, while not technically individual-level (in terms of applying to an individual throughout their lifetime), tended to be pragmatically non-repeatable, thus not easily yielding event measurement readings (e.g., *The diplomat was (?often) familiar with the trade agreement; The cheerleader was (?often) excited about the home game*). In these cases, although the gradable adjective could theoretically introduce a measure function to serve as a host for the degree variable (*more familiar; more excited*), there

126

was still no possible location within the predicate for the measure function *many* and correspondingly no possible repair to cardinality comparison. The 24 target items were combined with 60 filler items, and were presented to participants in four lists, so that each participant saw only one condition of each item.

To increase power so that the effects of adjectival gradability could be modeled, a slightly larger number of participants were recruited from Mechanical Turk to participate in this experiment than in the prior experiments. Exclusion criteria were adopted as before: participants were excluded from analysis if they showed qualitatively abnormal patterns of acceptability (i.e., rating obvious ungrammatical filler sentences higher than grammatical ones) or spent less than six minutes in total on the experiment. Of the 38 participants who completed the experiment, the data from 6 participants were excluded for not spending sufficient time on the task (mean time to completion: 4:39, as compared to a global average of 11:55). The data from the remaining 32 participants were included in the analysis.

The experiment procedure was the same as in prior experiments. Participants completed the experiment on Qualtrics, where they were provided with instructions about rating the acceptability of the sentences on a scale from one to seven. The instructions were identical to those used in Experiments 3-5.

4.5.2   Results

Prior to analysis all ratings were normalized using the acceptability judgments provided for all of the experimental stimuli for each subject. The mean ratings by condition, raw and transformed, are shown in Table 12:

	ILLUSION:	CONTROL:
SINGULAR SUBJECT:	raw: 3.90 (1.77) z-score: -.17 (.82)	raw: 5.31 (1.45) z-score: .60 (.61)
GROUP SUBJECT:	raw: 4.47 (1.73) z-score: .15 (.83)	raw: 5.45 (1.47) z-score: .68 (.61)

Table 12. Experiment 6: Mean (standard deviation) of acceptability ratings by condition.

A mixed effects linear regression model was fitted to the standardized data using the fixed effects for *illusion, subject type* and their interaction. In addition, as an exploratory measure fixed effects for *gradability* and its interaction with *illusion* were added to the model to assess whether reactions to illusions differed across the two counterbalanced predicate types. Due to convergence errors a fully specified random effects structure could not be used. Therefore, random intercepts were included to model variability across participants and items and model comparison was used to select the random slopes that significantly improved model fit; this process justified the inclusion of random slopes to model variability in the effect of the illusion across participants and items. The resulting model revealed a main effect of presence of illusion, although the effect size was descriptively much smaller than in prior experiments (β = -.28, t = -2.01; as compared to β = -1.11 in Experiment 3 and β = -.95 in Experiment 4; $\chi^2(1)$ = 22.36, p < .001). A main effect of subject type was significant via model comparison ($\chi^2(1)$ = 19.66, p < .001) but had a very low parameter estimate and a low t-value (β = -.08, t = -1.26). This was most likely due to the presence of a significant interaction between subject type and presence of illusion (β = -.24, t = -2.71; $\chi^2(1)$ = 7.30, p < .01), such that subject type affected illusion sentences specifically. There was no main effect associated with predicate gradability, although predicate type also entered into a significant interaction with the illusion: illusions with gradable predicates were rated higher than those with nongradable predicates (β = -.47, t = -3.85; $\chi^2(1)$ = 11.49, p < .001).
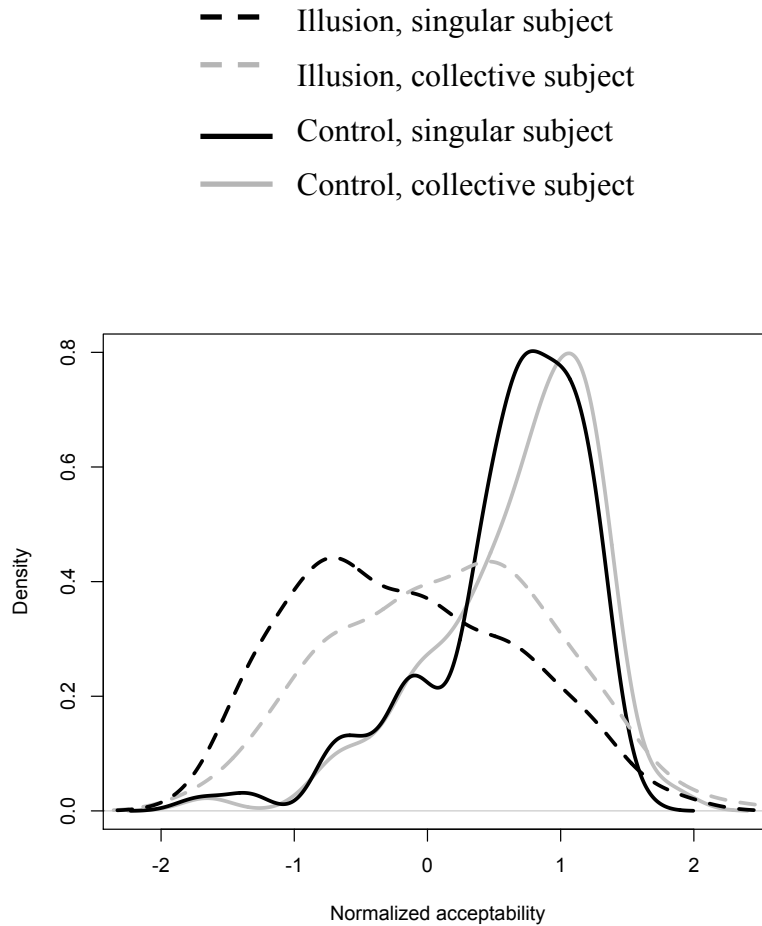
Figure 14. Experiment 6: Density plot of acceptability ratings

In order to further assess any differences associated with predicate type, Figure 15 plots the effect of the illusion by subject type, across the two predicate types. Visually it is clear that illusions and controls differed most substantially in cases where the predicate was nongradable and the *than*-clause was singular, as expected given that there is no obvious repair available in these cases. In addition, although the nature of the *than*-clause subject makes a clear difference for illusion sentences with a non-gradable predicate – with group nouns faring better than singular count nouns – its effects were less clear when the predicate was gradable: directionally, group nouns still seem to be rated a little higher, but the difference is much more subtle.
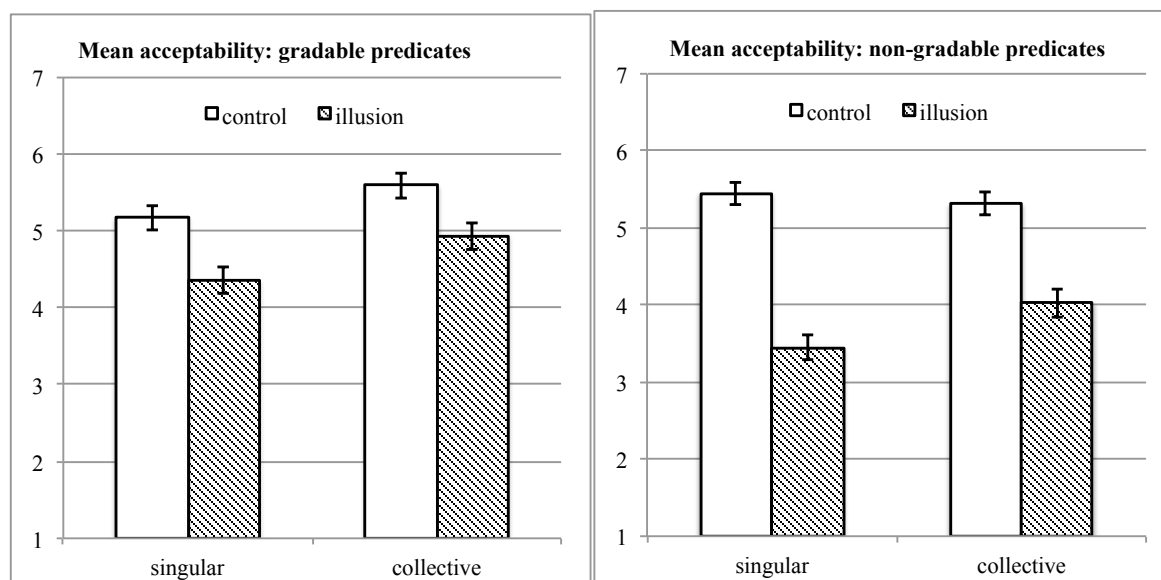
Figure 15. Effects of *illusion* and *subject type* in items with gradable predicates (left) and non-gradable predicates (right).

4.5.3 Discussion

Prior work, here and elsewhere, has repeatedly shown that Escher sentences are substantially more acceptable when they contain a plural *than*-clause subject such as *than the judges did*, even though pluralizing the noun does not affect the grammaticality of the sentence: *the judges* is no better a host for the degree variable than *the judge* is. Syntactically and semantically, there is no way for an individual-denoting DP to combine with the measure function *many*, which requires a predicate of pluralities as its first argument. Yet these types of illusions are perceived to be only minimally different from controls, a fact that seems to hold true irrespective of the internal contents of the predicate. In other words, whatever repair is responsible for the acceptability of these illusions, it does not seem to involve shifting to comparison of events, or of individuals provided by a noun phrase in a different syntactic position. Rather, the parser seems to strongly prefer to proceed with a comparison of individuals provided by the two subject noun phrases, whenever this analysis is possible.

In order to yield a well-formed LF, however, it would be necessary to make some change to the internal structure of either the matrix or *than*-clause subject. One possibility is that the problematic determiner could simply be omitted at the site of the anomaly, yielding the

130

type of comparison expected given the onset of the sentence: *more* NPs... *than* NPs. Another logical possibility is that the illusion sentence could involve partitive quantification: *more of the* NPs *... than the* NPs. In that case, a morpheme like *of* would be introduced in order to convert the individual-denoting DP into an entity that can be measured by *many*. Partitive repair would be a more widely available operation, since it can convert any individual-denoting noun phrase with salient atomic parts into a suitable argument for the measure function, regardless of whether or not the noun phrase is grammatically plural. Experiment 6 therefore investigated whether subject number would continue to modulate illusion rates when the subject noun phrase was conceptually plural but grammatically singular, as is arguably the case for group nouns such as *committee*. Such nouns pick out collections of individuals that form an atomic unit, and there are good reasons to differentiate them from actual grammatical pluralities (see in particular Barker 1992, Schwarzschild 1996). Among other things, in spite of their "notional" plurality, they are not grammatically suitable arguments for the cardinality measure function in *more,* which combines only with grammatically plural noun phrases (Hackl 2001).

In this experiment we found that illusions with group nouns indeed fared better than those with normal singular count nouns, providing some support for the availability of a partitive repair strategy. However, descriptively the pattern was not as robust as with true pluralities like *the judges*, and the effect was much subtler when the predicate was gradable. The overall smaller effect size could tentatively be caused by complications reconstructing the silent measure function in *more* as either *many* or *much*, so that repair would yield a more questionably acceptable percept. Whereas merging *many* will yield unambiguous cardinality measurement, examples like (66) suggest that such constructions heavily favor plural verbal agreement, which is only available in certain dialects of English (specifically, British English and Canadian English; see Quirk et al 1985, Barker 1992, Pearson 2011). As a result, some speakers may find this construction ill-formed, preferring instead to merge the mass quantifier *much* in the *than*-clause; in this case, however, their options are limited to the meanings in (67)b-c: either some level of mismatch between the antecedent and ellipsis site must be accepted, or else a strictly parallel analysis would be pursued by require merging *much* with *the judges*.

(66) a. How many of the team %are/??is in the picture?

  b. How much of the team is in the picture?

(67) a. how many of the judges were biased > ? how many of the jury were biased

  b. how many of the judges were biased > how much of the jury was biased

  c. ? how much of the judges were biased > how much of the jury was biased

Somewhat surprisingly, the presence of a gradable adjective had a robust effect on the illusion: illusion sentences were rated much higher when they contained a gradable adjective than when they were entirely non-gradable. Crucially, neither sentence type was especially amenable to event comparison – all of the predicates were either pragmatically non-repeatable (e.g., *be concerned about the recent criminal activity*), or else individual-level (e.g., *be allergic to peanuts*). It is possible that the gradable adjectives denoted properties that were only weakly non-repeatable – for example, it is possible to be repeatedly skeptical of a particular policy, or to be repeatedly angry about a particular tuition increase – although this reading is not especially salient considering the robustness of the effects observed here. It remains to be seen in future work what crucial property of these predicates drives their strong effect on illusion sentences, but a possible (if unexpected) explanation may be that any adjective that introduces a measure function can host the "lost" degree variable. This conclusion, if true, would have interesting implications both for the grammar and parser: on the one hand, it would suggest that repair of Escher sentences is fairly unconstrained, at least enough so to convert quantity comparison into non-quantity comparison. On the other hand, it also would indicate that the grammar of comparison has relatively deep cross-categorial and cross-dimensional parallels, possibly supporting a fundamentally unitary treatment of adjectival and nominal comparison (Wellwood 2014). Alternatively, the strong effect may indicate that the effects of event plurality on illusion sentences are related more strongly to the possible interpretation of the measure function as an adverbial *how much* – which itself can facilitate measurement of events (*… than how much I went to Berlin*) or other nonquantity dimensions (*…than how much the cheerleader is excited about the home game*).

4.6   General Discussion

The goal of this chapter was to investigate the range of percepts associated with Escher sentences, and correspondingly, the range of available repair operations. The first experiments on this illusion implicated a role for event comparison, leading to a proposal that speculatively pointed to the close relationship between nominal quantification and event measurement as the source of the illusion (Wellwood et al. 2009). In particular, the ambiguity of nominal *more* could contribute to the illusion since it can measure the noun phrase not only in terms of the absolute number of individuals but in terms of the number of event participants (as suggested by Krifka 1990), thus providing a route to event comparison largely already facilitated by the grammar; reinterpretation of illusion sentences would only require minimal changes within the silent material in the *than*-clause.

The results speak to the robustness of the illusion. If the bottom-up ambiguity associated with nominal *more* were the primary source of the illusion, then it would be expected to have primarily one percept – a comparison of events. By contrast, the experiments here have shown that it seems possible to reinterpret illusion sentences in a number of ways, and thus that there is a more general ability to recover from the grammatical anomaly in the comparative. In addition to applying the measure function to a plurality of events, it seems possible to apply it to a plurality of individuals provided by a different NP argument to the verb. Evidence in favor of this repair comes from Experiments 3 and 4, which revealed sensitivity to the semantic plurality of the object noun phrase in cases where neither the subject noun phrase nor the event structure contribute a suitable plurality for cardinality measurement. Moreover, in addition to providing the measure function with different types of arguments, Experiments 5 and 6 suggest that it may be possible to change the internal composition of the subject noun phrase so as to render it a suitable argument for the measure function. Experiment 5 found that subject plurality continues to significantly modulate illusion rates even when it has no impact on the event structure, and does so in such a strong way that the effects of object plurality are apparently rendered irrelevant. Finally Experiment 6 showed that this acceptability advantage persisted even when the subject noun phrase was only conceptually plural, but not grammatically plural – indicating that the subject noun phrase could be reinterpreted partitively by introducing a silent morpheme that maps a type e individual into a

predicate of individuals. Collectively, this suggests that the illusion sentence in (68) could theoretically be perceived with any of the meanings in (69).

(68)   More judges took vacations than the lawyers did.

(69)   a. NOMINAL COMPARISON, SUBJECT:

−er (λd . d-many of the judges took vacations) (λd . d-many of the lawyers took vacations)

b. NOMINAL COMPARISON, OBJECT:

−er(λd . judges took d-many vacations)(λd . the lawyers took d-many vacations)

c. EVENT COMPARISON:

−er (λd . judges took vacations d-much) (λd . the lawyers took vacations d-much)

In practice, however, the parser seems to be biased towards a comparison of cardinalities of individuals provided by the subject noun phrases, i.e. (69)a, when such an interpretation can feasibly be generated. In other words, the most illusory sentences tend to be perceived as equally acceptable when they allow one repair as when they allow two or more, and thus the number of possible repairs does not affect acceptability in an additive way. For example, (70)a-b are perceived as approximately similar in acceptability (although no direct comparison was made in the experiments reported here), even though (70)b allows for at least three repairs while (70)a only allows for one. The picture is similar, if somewhat less clear, in Experiment 6, where the presence of a non-quantity measure function (such as the one introduced by *happy* in (71)) seems to substantially increase ratings for illusions, and to correspondingly reduce the import of the presence of a group noun, rendering the variants in (71)a-b quite close in acceptability.

(70)   a. More judges are originally from California than the lawyers are.
        b. More judges took expensive vacations than the lawyers did.

(71)   a. More politicians are happy with the transportation workers' compensation than the union is.

b. More politicians are happy with the transportation workers' compensation than the worker is.

This suggests that repair operations may be ordered, with certain repairs favored and others attempted only as a last resort. From the perspective of the mechanisms underlying repair, it is understandable why a comparison of e.g. *judges* and *lawyers* could be a preferred analysis for (68). Recall that, in the context of a serial repair mechanism, the relative acceptability and extent of slowdown associated with illusion sentences is likely to be related to the extent of changes required (Arregui et al. 2006; Frazier 2014), which presumably can be tracked by how different the input and output of repair are. The repair at stake in (69)a is the closest possible near-neighbor analysis to the one expected prior to the point of anomaly: ontologically it is faithful to comparison of individuals; syntactically it is faithful with respect to the position of the measure function; only minor changes to the matrix subject would be necessary. In addition, from the perspective of a parallel processing architecture that tracks probabilistic information both predictively (what upcoming input is likely to be) and retrospectively (what prior input is likely to have been), such as Levy (2011) – the matrix subject noun phrase is the constituent that is furthest away from the ellipsis site, both temporally and linearly; a comprehender might naturally be more uncertain about their memories of sentence elements that are further away, leading them to more readily accept changes to the input at those locations.

In light of the range of percepts we observe here, it is somewhat puzzling that displacement of *more* does not seem to be a critical component of the illusion. Wellwood et al. (2012) report that the illusion persists with *fewer,* a strictly adnominal degree quantifier, and in Experiments 1 and 2 the same was true for *as many*. Moreover, they report that in production tasks, participants tended to make various changes to the illusion sentence to render it grammatical, yet only rarely repeated back the illusion sentences by overtly displacing *more*. However, both of these facts speak to overt displacement in the phonological form of the illusion sentence, whereas the results here suggest that the type of displacement illustrated in (69) almost certainly occurs at more abstract levels of the grammar. For example, Experiments

3 and 4 found clear differences between illusion sentences with morphosyntactically plural noun phrases that differ in semantic number, while Experiment 6 found clear differences between illusion sentences with morphosyntactically singular noun phrases that differ in conceptual number. In other words, we have elicited sensitivity to semantic number properties in cases where morphosyntactic number is held entirely constant, suggesting that it is the relatively abstract interpretational properties of the sentence that influence the illusion, not the more superficial syntactic properties. The evidence in total thus indicates that the possibility of overtly displacing *more* in the surface syntactic form of the illusion sentence is a neither necessary nor sufficient condition for the perception of acceptability.

It is occasionally suggested that Escher sentences arise because of the availability of a *just me* interpretation (*More people have been to Berlin than just me*, interpreted as *I am not the only one who has been to Berlin*; Grant 2013; see Greenberg 2009, Thomas 2011 for discussion of the semantics of additive *more*). One might thus wonder whether there is yet another repair available – a shift to subset comparison – in addition to those advocated here. This account has not yet been addressed because behavioral reaction to illusion sentences do not, in general, indicate any sensitivity to the availability of subset comparison. Wellwood et al. (2009, 2012) report that the illusion does not decrease in acceptability when the two noun phrases have disjoint rather than subset interpretations (for example, in sentences like *More boys played in the park than she did*). Moreover, as noted above, there were no significant differences between different degree quantifiers (*more, as many, fewer*), in acceptability or reaction times, even though subset comparison arises only with *more* (**Fewer students graduated from the law school than just them, *As many graduated from the law school as just him*). These facts together indicate that the additive *just me* interpretation is not a critical component of Escher illusions.

Our results are relevant to the question of the nature and quantity of possible percepts of Escher sentences. It should be noted that the presence of multiple illusory percepts does not necessarily speak to the availability of multiple distinct LF representations. As noted in Chapter 3, sensitivity to predicate repeatability is itself consistent with two general scenarios: in the first, repair of Escher sentences yields a narrow comparison of events: *the number of events of people going to Berlin exceeds the number of events of me going to Berlin.* In the second scenario, the output of repair yields comparison that is underdetermined, allowing but

not requiring an event comparison construal. The results here do not strongly differentiate among these scenarios but do add additional detail to the picture. If a repeatable predicate tends to result in a narrow comparison of events, then it must be the case that there are fundamentally at least three distinct repairs that can be applied to Escher sentences, and three percepts that can be distinguished at LF: one that situates the degree variable in an adverbial measure function such as d-*much* or d-*many times*, and two that situate the degree variable in nominal measure functions, such as d-*many* NP.

(72)   -er (λd . the judge vacationed in Florida **d-much**)
        (λd . lawyers vacationed in Florida **d-much**)

(73)   −er (λd . the dog has **d-many** toys)
        (λd . cats have **d-many** toys)

(74)   −er(λd . **d-many of the judges** vacationed in Florida)
        (λd . **d-many of the lawyers** vacationed in Florida)

An alternative possibility is that illusion sentences might have only one basic LF, underspecified with respect to the nature of the comparison. This type of situation could arise if *more* were interpreted adverbially, since adverbial elements widely give rise to quantificational variability effects. For example, the restrictor of *for the most part* is typically determined flexibly, often provided by the topical material in the sentence (von Fintel 1997, Nakanishi & Romero, 2004):

(75)   For the most part, the cats have mouse toys.
        a.   ≈ Most of what the cats have are mouse toys.
        b.   ≈ Most of the cats have mouse toys.

(76)   For the most part, the lawyers vacationed in Florida.
        a.   ≈ Most of the lawyers vacationed in Florida.
        b.   ≈ The lawyers vacationed in Florida most of the time.

If it were possible to interpret *more* as an adverbial element with quantificational variability, analogous to *for the most part* in (77)-(80), then one would naturally expect sensitivity to many of these different elements of the sentence, depending on what parts of the sentence are contextually, pragmatically or prosodically foregrounded.

(77)　For the most part, the teachers are from [Oregon]F.

　　　INDIVIDUAL QUANTIFICATION, SUBJECT: Most of the teachers are from Oregon.

(78)　For the most part, the child owns [picture]F books.

　　　INDIVIDUAL QUANTIFICATION, OBJECT: Most of the child's books are picture books.

(79)　For the most part, the teacher [gardens]F on the weekend.

　　　EVENT QUANTIFICATION: Most of the teacher's weekend events are gardening events.

(80)　For the most part, the citizen was familiar with the law.

　　　ADJECTIVAL QUANTIFICATION, SUBJECT: The citizen is familiar with the law to a large extent.

Indeed, while neither of the comparatives in (81) are particularly acceptable under an event reading, to the degree that *more* can be construed as *to a greater extent* in this position, it may be possible to derive a cardinality-of-mouse-toys reading without merging *more* directly with the object noun phrase.

(81)　a. The dog has mouse toys more (? to a greater extent) than the cat does.

　　　b. The dog has a tail more (* to a greater extent) than the cat does.

Of course, there is no obvious precedent in the grammar for this meaning for nominal *more* (and, as I noted in Chapter 3, there are good arguments in favor of global integration of the two clauses occurring within the grammar specifically), making this type of approach seem initially somewhat implausible. However, one possibility is that this type of interpretation arises at

intermediate levels of representation during the processing of comparatives, and it is the presence of this memory trace that affects the processing of Escher sentences.

To better understand this explanation, let us first explore an analogous behavioral pattern long observed in the literature on *quantifier spreading* (Inhelder & Piaget 1964, Roeper & de Villiers 1991, Phillip 1995, Crain et al. 1996, Drozd 2001, Geurts 2003), which has revealed that children tend to improperly restrict the individual quantifier *every* in sentences like (82) (Phillip 1995, Roeper & DeVilliers 1993, Sauerland 2003). Most accounts of this phenomenon propose in some manner that the child arrives at this interpretation by restricting the determiner contextually, similar to an adverb, which explains why the prevalence of quantifier spreading seems to closely depend on focus, discourse context and visual salience; however, researchers by and large associate the pattern of errors uniquely with stages of child language acquisition.

(82)   (Context: Picture of three boys each riding ponies, and one empty pony)
        Every boy is riding a pony.
        a. Quantifier spreading response: false, (not that pony)
        b. Grammar-based response: true

Recent studies have found that this issue is not exclusively associated with acquisition – adults also occasionally entertain the reading in (82)a, but they typically do so only fleetingly, or in very difficult conditions (DellaCarpini 2003, Brooks & Sekerina 2006). This suggests that a "contextual determiner restriction" pattern may be surprisingly prevalent in processing, even though this type of interpretation clearly violates the grammar of nominal quantification.

The findings on adult quantifier spreading are puzzling and have received little attention; original studies attribute the misinterpretation to shallow processing without much justification – a conclusion that is equally problematic in this case for many of the same reasons that it is problematic for Escher sentences. However, results from Slattery et al. (2013) suggest a slightly different way to think about this phenomenon. Slattery et al (2013) argue that misinterpreted garden-path sentences, first discussed by Christianson et al. (2001), do in fact have a fully complete and faithful syntactic representation; the problem is simply that memory traces of interpretations entertained at intermediate stages of processing are not consistently

discarded. This account does not rely in any crucial way on heuristics, but rather explains how incremental processing and memory interact. With respect to quantifier spreading, the analogy with garden-path sentences is especially plausible, because children – whose inhibitory control is typically less developed (Zelazo & Frye 1998; Durston et al.2002; Zysset et al. 2001) – are already known to be susceptible to the so-called 'kindergarten-path effect', failing to fully reanalyze temporary ambiguities as often as 60% of the time (Trueswell et al.1999). This could suggest that there is an intermediate stage of processing where the quantifier is represented without its accompanying first argument *pony,* and is thus temporarily restricted through context alone; if the comprehender fails to discard this intermediate representation, they will end up with an interpretation that is not fully licensed by the end product of the grammar (but not one that is derived extragrammatically, either). This scenario could extend to nominal comparative quantifiers like *more,* with the illusion arising as comprehenders experience a garden path effect at the anomaly site, and – as suggested by Slattery et al (2013) – revert to an interpretation generated at intermediate stages of processing.

While this account may seem a bit ad hoc, the foregoing discussion is simply meant to illustrate that, while the results from Chapters 3-4 reduce the hypothesis space substantially, there remain various approaches to Escher sentences that can still be adopted. One possibility is that there is deliberate repair of the illusion sentence – the analysis advocated here – such that the illusion sentence is processed compositionally, the anomaly is detected at some level, and in response a new input is fed to the compositional engine to facilitate interpretation. In this case, it is as though the comprehender is choosing to interpret the illusion sentence as its closest grammatical near-neighbor – possibly an example of speech error reversal. By contrast, one could possibly argue that there is a quantificational variability effect arising at some intermediate stage when structure-building is ongoing, and that this interpretation lingers long enough to affect the perception of the illusion sentence even after its anomaly has been detected. In this case, substantially more would need to be said about how and why comparatives are represented this way at intermediate stages of processing – since there is clearly no analogous grammatically-sanctioned meaning associated with nominal *more.* Further research is needed to differentiate among these possibilities, each of which promises to reveal slightly different but no less interesting information about the grammar, the parser, or both.

Why does this phenomenon appear exclusively within the context of the grammar of comparison? First, it is not yet clear that this is the case. There are a number of independent phenomena that arguably fall under the rubric of "acceptable ungrammaticality" (Frazier 2014), most notably other examples of elided elements with flawed antecedents, as well as examples where the parser seems to actively consider analyses not sanctioned by the grammar, including local coherence effects (Tabor et al 2004, Gibson 2006, Bicknell & Levy 2009). However, to the extent that some aspect of the illusion is specific to the grammar of comparison, the reasons may be related to the special challenges that its grammar poses for compositional online processing. As discussed in more detail in Chapter 2, left-to-right incremental processing of comparatives engenders a greater amount of uncertainty than usual because it is very difficult to proactively anticipate the content of the extraposed *than*-clause; and until that content is known, the amount of compositional processing that can occur is somewhat limited. This does not appear to affect whether the anomaly is detected – since we clearly elicited evidence of online processing difficulty – but it may explain the puzzling ease of reinterpretation, which is so subtle as to entirely evade conscious awareness.

5 INVERSION SENTENCES: OVERVIEW & BACKGROUND

5.1 Overview

The phenomenon of inversion illusions – occasionally referred to as *depth-charge* illusions due to their extremely delayed detection – is pervasively acknowledged, but ill-understood. A simple Google search suggests that the paradigmatic example of such sentences, shown in (1), may originally be due to Hippocrates; this sentence is almost universally perceived to mean that no head injuries should be ignored, even those that are very trivial.

(1)　　a. No head injury is too trivial to ignore.
　　　　b. Percept: **No** head injuries can/should be ignored, even the most trivial ones.

To the surprise of many speakers, the true meaning of this sentence is exactly the opposite of its perceived meaning, as can be demonstrated by way of analogy with other similarly structured sentences. For example, the syntactically identical sentence in (2) is typically perceived to mean that *all* missiles can be banned, not that *no* missiles can be banned:

(2)　　No missile is too small to ban.
　　　　Percept: **All** missiles can/should be banned, even the smallest ones.

The semantic analyses outlined in Chapter 2 suggest that the meaning associated with (2) is correct – i.e. it is consistent with the semantics of *too* according to its distribution elsewhere in English – while (1) is not. Assuming that the degree quantifier always scopes under individual quantifiers (Kennedy's Generalization; Heim 2000), (3)-(4) might capture the "true" meaning of these sentences.

(3) a. [[No missile is too small to ban]] = 1 iff ¬∃x: missile(x) & max{d: small(x) ≥ d}

>

max{d: ∃w' ∈ Acc(w) & x is banned in w' & small(x) ≥ d in w'}

b. There is no missile whose smallness exceeds the maximal allowable smallness for banning it.

(4) a. [[No head injury is too trivial to ignore]] = 1 iff ¬∃x: head injury(x) & max{d: trivial(x) ≥ d)} > max{d: ∃w' ∈ Acc(w) & x is ignored in w' & trivial(x) ≥ d in w'}

b. There is no head injury whose triviality exceeds the maximal allowable triviality for ignoring it.

Wason & Reich (1979) note that there are two components of the meaning of (1) that render it anomalous, and correspondingly, there are two ways that the semantics illustrated in (4) are inconsistent with its perceived meaning. The first problem concerns what we term the "internal" meaning of the sentence, namely the fact that (4) generates an implausible threshold of triviality. To illustrate this, consider first the "internal" meaning of the phrase *too small to ban*. This phrase presupposes that there is some size threshold *d*, and anything that is banned has a size greater than *d* (thus, objects of a sufficiently small size may be *too small to ban*). In the toy scenario shown in Figure 16, *too small to ban* is predicated of an object whose size is less than d_1 while all objects that can be banned have a size greater than d_1 (note that negative adjectives like *small* or *trivial* will map inversely onto the scales of size and seriousness; see Meier 2003 for discussion of the semantics of adjectival polarity in *too* sentences):

size

d₃

d₂

d₁

w1　w2　w3

x's size in worlds where
x is banned
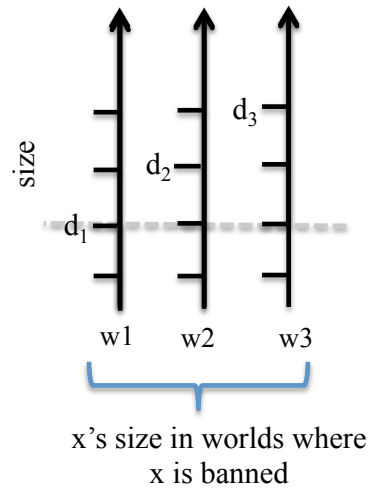
Figure 16. Size threshold for the phrase *too small to ban* (illustrated as grey dotted line)

The "internal anomaly" associated with *too trivial to ignore* in (1) concerns the fact that it projects a threshold of seriousness associated with ignoring an injury, but not in the direction that world knowledge leads us to expect: whereas someone would be more likely to ignore a head injury when it is trivial, the phrase *too trivial to ignore* implies the opposite. This is illustrated in Figure 17: in all of the worlds where an injury is ignored, it is *at least* d_1-serious; only objects that are *more* serious than d_1 can be ignored. This is clearly the opposite of what is typically true (and what is inferred by the phrase): in most circumstances, injuries are likely to be ignored when they are more trivial, not more serious.

144
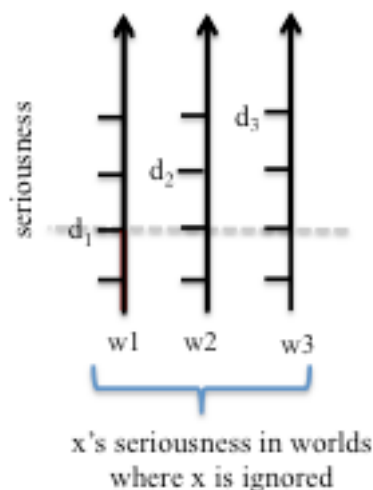
Figure 17. Seriousness threshold presupposed by *too trivial to ignore* (illustrated in grey)

Other, similarly implausible internal meanings – which are somewhat easier to process – are illustrated in (5): it does not make sense for it to be *too cloudy to rain,* for food to be *too old to go bad,* or for a jacket to be *too thick to stay warm.*

(5)   a. # It is too cloudy to rain.
        b. # This food is too old to go bad.
        c. # My jacket is too thick to stay warm.

The second problem with (1) concerns the so-called "injunction" of the sentence, what I term its "external meaning": that head injuries can or should be ignored. This is precisely the opposite of what most people infer from the sentence, and the opposite of what we usually think to be true of the world. The injunction arises because the sentence asserts that no head injury exceeds the presupposed maximum, and therefore that no head injuries fall within the "can't be ignored" portion of the scale. Similarly, the sentences shown in (6) are also implausible given their questionable suggestion that one should take illegal drugs and waste natural resources.

(6)   a. # No illegal drug is too dangerous to take. (# All illegal drugs should be taken)

        b. # No natural resource is too precious to waste. (# All natural resources should be wasted)

It is possible to dissociate these two problems by manipulating the polarity of the individual elements of the illusion sentence. The internal meaning of the illusion arises due to the relationship between the gradable predicate and the complement clause with respect to the degree quantifier – i.e. it arises in the local phrase *too trivial to ignore* – and manipulating the polarity of these elements can fix the problem. For example, changing *trivial* to *serious* changes the internal meaning of the illusion sentence, although it does not change its injunction:

(7)   a. # external, √ internal: No head injury is too **serious** to ignore.

        b. # Head injuries should be ignored, √ even the most serious ones.

The injunction, by contrast, arises because of the relationship between the noun and verb with respect to the degree quantifier – *no head injury is too _____ to ignore* – and can be addressed by manipulating the polarity of these elements. For example, changing *no* to *all* will change the illusion's injunction to match the percept, but will retain the anomalous internal meaning introduced by *too trivial to ignore*:

(8)   a. √ external, # internal: **All** head injuries are too trivial to ignore.

        b. √ Head injuries should not be ignored, # even the most trivial ones.

Thus, a change to the adjectival or nominal polarity resolves one aspect of the illusion's meaning, but the sentence retains at least one anomaly. Only the verb and degree quantifier participate simultaneously in both aspects of the sentence meaning; by changing them, we seem to arrive at a meaning that resembles the commonly reported percept of (1).

(9)   √ external, √ internal: No head injury is too trivial to **treat**.

All head injuries can/should be treated, even the most trivial ones.

(10)   √ external, √ internal: No head injury is trivial **enough** to ignore.

No head injuries can/should be ignored, even the most trivial ones.

Several observations are repeatedly made about the sentence in (1). The first is that there is very strong pragmatic bias surrounding the topic at hand: almost anyone would agree that even apparently trivial head injuries are worthy of some care and attention. This means that the grammar is encoding a message that a reasonable person would probably not utter, and people appear to rather systematically "override" that message by inverting it into something that falls in line with their own world knowledge, "effectively [letting] pragmatics overcome local semantics in forming an interpretation" (Sanford & Emmott 2012: 28). In other words, similar to accounts of depth-inversion illusions (such as the hollow mask illusion), it is sometimes thought that the cues provided by bottom-up language input are "trumped" by world knowledge.

The second observation is that the grammar of (1) is "permeated with negativity" (Wason & Reich 1979: 592). The most obvious such example is the morphologically negative individual quantifier, *no.* However, the adjective, degree quantifier, and verb are all implicitly or inherently negative. With respect to the adjective, negativity can be diagnosed by observing the adjective's interpretation in degree questions: like other negative adjectives (e.g. *short, young*), degree questions with *trivial* are necessarily norm-related. Thus, whereas the questions in (11)a-(13)a do not presuppose that John's head injury is serious, that John is tall, or that John is old, the questions in (11)b-(13)b presupposes that his head injury is trivial, that he is short, or that he is young:

(11)   a. Q: How serious is John's head injury? A: It's very trivial.

b. Q: How trivial is John's head injury? A: ? It's very serious.

(12)   a. Q: How tall is John? A: He is very short.

b. Q: How short is John? A: ? He is very tall.

(13)   a. Q: How old is John? A: He is very young.

b. Q: How young is John? A: ? He is very old.

With respect to the degree quantifier, *too* can be shown to be more negative than *enough* along several dimensions. First, the fact that *too* licenses strong and weak NPIs (14), let-alone rejoinders (15) (Schwarzschild 2008), and inferences from set to subset (16), suggests that it is inherently downward-entailing like other negative elements. *Enough,* on the other hand, introduces an upward-entailing environment, which precludes NPIs, let-alone rejoinders, and licenses inferences from subset to set.

(14)   a. None of my friends will {ever get a good job; lift a finger; work at all; go to work until Tuesday}

b. *All of my friends will {ever get a good job; lift a finger; work at all; go to work until Tuesday}

c. John is **too lazy** {to ever get a good job; to lift a finger; to work at all; to go to work until Tuesday}

d. *John is **ambitious enough** {to ever get a good job; to lift a finger; to work at all; to go to work until Tuesday}

(15)   a. This fuel should not be used in a car engine, let alone a lawn mower.

b. *This fuel should be used in a car engine, let alone a lawn mower.

c. This fuel is **too volatile** to use in a car engine, let alone a lawn mower.

d. *This fuel is **volatile enough** to use in a car engine, let alone a lawn mower.

(Schwarzschild 2008)

(16)   a. No one ate <u>dinner</u>. → No one ate <u>dinner and dessert</u>.

b. Everyone ate <u>dinner and dessert</u>. → Everyone ate <u>dinner</u>.

c. I am **too full** to eat <u>dinner</u>. → I am too full to eat <u>dinner and dessert</u>.

d. I am **hungry enough** to eat <u>dinner and dessert</u>. → I am hungry enough to eat <u>dinner</u>.

Second, *too*-constructions typically license negative implicatures, with speakers inferring that the proposition expressed by *to*-clause is false*,* while *enough* licenses positive implicatures where the content of the *to*-clause is assumed to be true (Karttunen 1971; Hacquard 2005) – both of which can be cancelled[13].

(17)   a. John was too slow to escape. +> John did not escape.

b. John was too slow to escape, but somehow he managed!

(18)   a. John was quick enough to escape. +> John escaped.

b. John was quick enough to escape, but for some reason, he didn't.

Finally, with respect to the verb*,* in spite of the lack of clear negative marking, *ignore* should be considered an inherently negative lexical item since it licenses certain NPIs in its complement clause in sentences like (19) (Klima 1964, Laka 1990), similar to other inherently negative verbs like *deny* in (20)a, and unlike e.g. *state* in (20)b. Semantically, *ignore* also denotes a lack of action (*to ignore = to not attend to*), a property that seems to be shared by the class of verbs associated with inversion illusions (e.g. *overlook; miss*).

(19)   John ignored that I was ever injured.

---

[13] This implicature may alternatively be an entailment given that perfective aspect makes such inferences difficult to cancel; the lack of clear verbal morphology in English may obscure this distinction (see Hacquard 2006 for an account of these facts, and Homer 2010 on the relationship between aspect and actuality entailments).

(20)   a. John denied that I was ever injured.

       b. * John stated that I was ever injured.

To summarize, the original and widely-cited example in (1) is strongly implausible, both "externally" and "internally", while its commonly reported inverted meaning has none of these problems. An explanation frequently offered in passing for this phenomenon, without much further explanation or systematic experimentation, is that strong pragmatic bias allows comprehenders to "override" the compositional interpretation of the sentence. This "pragmatic normalization" (Fillenbaum 1971, 1974) appears to be limited to, or facilitated by, the many instances of negation that permeate it, potentially including the negative individual quantifier (*no*), the negative adjective (*trivial*), the negative verb (*ignore*) and the degree quantifier (*too*). In the next section, we unpack the range of possible explanations about how negativity in the illusion sentences could yield the observed effects.

5.2   Three hypotheses of inversion sentences

5.2.1   *The Channel Capacity Hypothesis*

Why would negation have this particular effect on inversion sentences? As Wason and Reich (1979) note, the sheer number of negative elements might be thought to "overload the channel capacity of the individual and render the sentence incomprehensible" (p. 592). Here they are likely referring to the many psycholinguistic studies demonstrating difficulties processing negation, including longer reaction times and lower accuracy rates (Clark 1970, Just & Carpenter 1971, Clark & Chase 1972, Trabasso, Rollins & Shaughnessy 1971, Wason 1961), greater difficulty with sentence recall (Cornish & Wason 1970, Clark and Card 1969), and greater amounts of cortical activation (Carpenter et al. 1999). Such experiments often focus on the verification of declarative statements with and without sentential negation, such as (21) (from Clark & Chase 1972):

(21)   The star (is/is not) above the plus.

+   *

*   +

        Difficulty with negation is not limited to morphologically overt negative operators like *not* but also extends to inherently negative lexical items such as *hardly, scarcely* and *a few* (Just & Carpenter 1971), *absent, different,* and *conflict* (Clark 1971), negative quantifiers like *few* and *no* (Glass, Holyoak & O'Dell 1974) and verbs with negative implicatures, such as *forget* (Just and Clark 1973). In a study especially relevant to inversion sentences, Sherman (1976) asked participants to provide sensicality judgments for sentences with up to four instances of negation (nominal, verbal, adjectival, and sentential, as schematized in (22) and found steady increases in reaction times and error rates in a sensicality task as the number of negative elements increased, with special difficulty arising in cases of three or more negative elements.

(22)   {No one/everyone} {believed/doubted} that he was (not) (in-)capable of sustained effort.
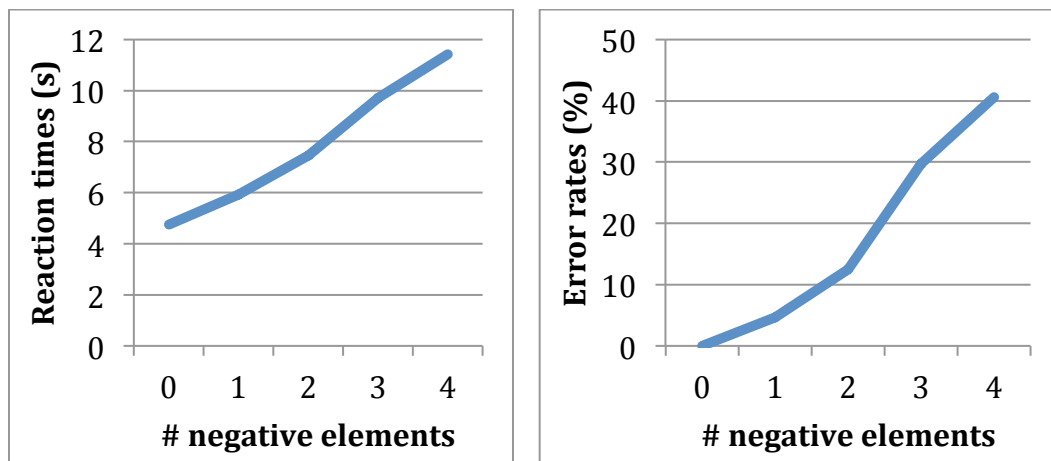

Figure 18. Results from Sherman (1976) suggest that reaction times (left) and error rates (right) rise steadily as the number of negative elements in a sentence increases.

A traditional explanation of the difficulty of comprehending negation relates to the fact that negation is commonly considered an operator applied to a proposition; negative sentences therefore contain an additional layer of meaning, and also potentially require an inner proposition to be first represented and comprehended before it can be negated (Carpenter & Just 1975, Clark 1976, Fischler et al 1983, Kaup & Zwaan 2003, Kaup et al. 2006), resulting in a substantial delay in interpretation. As evidence for this "two-step" hypothesis researchers commonly point to the negation-by-truth-value interaction for a design like (23): whereas responses to false sentences like (23)b are normally slower than those to true sentences like (23)a, the opposite pattern is observed under negation, with true negatives like (23)c more difficult to comprehend than false negatives like (23)d (Trabasso et al. 1971, Clark & Chase 1972). This result would make sense if the construction of the inner proposition, *A robin is a tree,* were first verified before interpreting it within the scope of the negation.

(23)   a. A robin is a bird.   (True affirmative)

          b. A robin is a tree.   (False affirmative)

          c. A robin is not a tree.   (True negative)

          d. A robin is not a bird.   (False negative)

The cumulative body of work on the processing of linguistic negation seems to broadly suggest that an inner proposition or concept is first represented and then inhibited – often resulting in activation of the complement set (e.g., Sanford & Moxey 2004), at least to the extent that the complement set is contextually and semantically constrained enough for such effects to be apparent (see Staab 2007 for a review). Although a number of studies have found inhibitory effects for negated concepts (e.g. inhibition of the noun *bread* after reading a phrase like *no bread;* MacDonald & Just 1989), other studies have failed to find such effects (Giora et al 2004). Hasson & Glucksberg (2006) reconciled these mixed findings by measuring inhibitory effects at different lag times: at early lag times, they found facilitory effects for concepts that were negated, but at later lag times this effect disappeared and inhibitory effects began to emerge (see also Kaup, Lüdtke & Zwaan 2006; Lüdtke et al. 2008). This type of finding is precisely as one would expect if the processing of negation involves an intermediate step of accessing the inner proposition, or the concept that is to be negated. And in fact, given

some amount of delay in the experimental task – for example, when given the opportunity to comprehend the sentence before verifying it against a picture – the interaction observed for a paradigm like (23) disappears, suggesting that participants were able to construct and negate the inner proposition before the verification task began.

The lingering effects of this inner proposition can apparently still be felt, however, for some time. Research on persuasion suggests that statements containing negation tend to lead to attitude changes in the opposite direction; for example, a statement such as *drinking is not sexy* may tend to make drinking more attractive (Christie et al 2001, Jung Grant et al 2004, Skurnik et al 2005). Relatedly, in the so-called *innuendo effect* (Wegner et al. 1981), the negation of an unflattering description of a person (e.g. *the politician was not bribed*) tends to nevertheless lead to unflattering perceptions of that person. These types of findings might be taken to suggest that the activation of the inner proposition in a negated sentence can linger in memory, influencing subsequent attitudes even after the correct meaning of the negated sentence has been inferred.

There is a very real question as to whether part or all of the processing difficulty associated with negation is related to the pragmatically odd use of negation in isolated sentences, especially in those that negate a proposition that is absurd to begin with (e.g., (23)c). Pragmatic accounts tend to emphasize the fact that negative statements are used in natural conversation to deny some salient supposition (Givon 1978, Horn 1989, Strawson 1952); processing of negation therefore will naturally be difficult if its use is pragmatically unlicensed. For example, Wason (1965) showed that in "contexts of plausible denial" – contexts where some scenario is expected to occur but does not – negation is processed more efficiently. When shown a display with seven red circles and one blue circle, participants found it easier to complete statements like (24)b than those like (24)d.

(24)	a. Exactly one circle is…	(affirmative, exceptional)
	b. Exactly one circle is not…	(negative, exceptional)
	c. Exactly seven circles are…	(affirmative, majority)
	d. Exactly seven circles are not…	(negative, majority)

In support of a pragmatic account, Lüdtke & Kaup (2006) found that the difficulty of negation

varied as a function of context, with denial easier in contexts that provide an explicit supposition for consideration. Strong supporting contexts that introduce the supposition and constrain the complement set have been found to eliminate the interaction so often observed in e.g. (23) (Nieuwland & Kuperberg 2008; Staab 2007). As far as I can tell, these results can also be accommodated by the two-step hypothesis, given some additional assumptions. In isolation the completion of two steps – accessing the "inner" proposition and then denying it – may take substantially longer than the single step required for comprehending affirmatives, especially in cases where the inner proposition is pragmatically anomalous. A suitably rich context, however, may enable comprehenders to predictively generate the semantic content of the inner proposition at an abstract level. Because it would be easier to retrieve a predicted proposition than generate one out of the blue, the initial stage of interpreting negation should be much more efficient, with shorter reaction times and higher accuracy rates.

At the same time, the status of *not* as a propositional operator is clearly insufficient for explaining the degree of processing difficulty associated with negation, especially given that it is not the only propositional operator available in the grammar. Modal quantifiers like *must* are often thought to have a similar logical structure embedding an "inner proposition", and yet it is not clear that sentences with modal quantifiers yield difficulty at the same levels. In addition, as noted, processing difficulty has been observed for negative elements that are not sentential operators (including morphologically or "inherently" negative words with no scopal mobility). Thus, the representational complexity of negative sentences does not seem to be a crucial factor in their relative difficulty.

In fact, there is some limited evidence to suggest that the difficulty associated with negation is not even strictly linguistic in nature. Carlson (1989) trained participants to provide judgments about representations of digital *logic gates,* which operate on binary inputs (0 or 1) and produce binary outputs (0 or 1). The pattern of required responses is guided by truth tables for various logical connectives, and there are both positive (Fig 14a) and negative (Fig 14b) logic gates. Subjects were trained to provide appropriate outputs by studying truth tables labeled either using explicit linguistic negation in one condition, or using nonlinguistic symbols in another. Results showed evidence of processing difficulty associated with negative conditions regardless of whether there was any explicit linguistic negation.

154
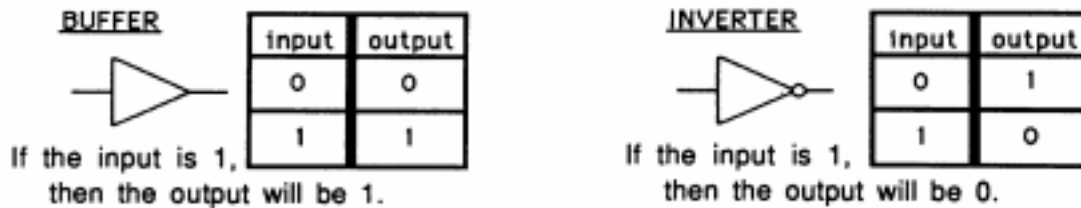
Figure 19. (a) The simplest positive logic gate used in Carlson (1989), the identity function, (b) The simplest negative logic gate. Other conditions included *and/not and, or/not or.*

The label "Channel Capacity Hypothesis" suggests that the difficulty of negation lies somehow in working memory overload, recalling Miller (1956)'s classic finding that average adults have a memory span that can store approximately seven digits, although it is also not immediately clear what aspect of processing negation would yield quantitative overload of working memory capacity. Given that the correct interpretation of negation may require accessing and then immediately inhibiting a concept, it seems reasonable to localize such difficulties to executive control and inhibitory processes rather than the storage capacity itself or other properties of the memory architecture. Indeed, this type of account is supported by studies suggesting that the processing of negated sentences such as *Adesso non premo il bottone* 'Now I don't press the button' entails substantial cortical deactivations in brain regions associated with the processing of action-related statements (Tettamanti et al 2008). Regardless of the precise cause of processing difficulty, the Channel Capacity Hypothesis would claim that the sheer number of negative elements overloads processing channels, leading the grammar-based parse to fail. The predictions of this account are relatively clear from prior psycholinguistic work on negation: as the number of negative elements within the illusion sentences increases, there should be a correspondingly steady cline in accuracy along with an increase in reaction times, with potentially greater effects for cases of three or more negative elements. The most difficult conditions in Sherman (1976) were those that contained the negative verb *doubt,* suggesting a particularly important role for the verb (also suggested by Wason & Reich 1979; Cook & Stevenson 2010; Fortuin 2014), although his findings suggest that all of the negative elements of the sentence might be expected to exert some influence independently.

A corollary of this hypothesis is that, to the extent that comprehenders are able to derive some interpretation, it must be predicted by a top-down interpretational heuristic, given

that the usual compositional processes have failed. Wason & Reich (1979) and others (Sanford & Sturt 2002, Garrod & Sanford 1998, Ferreira et al. 2002, Sanford & Emmott 2012 a.o.), surmise that plausibility is responsible; in other words, comprehenders are forced to disregard the grammar and put the pieces of the sentence together in a sensical way. This seems like a reasonable cursory explanation of the phenomenon, given that the perception of the illusion diverges from the output of the grammar, and it seemingly does so in a way that aligns better with world knowledge. As we saw, sentences where the syntax-based meaning converges with world knowledge – such as *No missile is too small to be banned* – are apparently interpreted unproblematically (*All missiles should be banned*).

Though widely adopted, this account has received somewhat mixed experimental support. Wason & Reich (1979) analyzed paraphrases for two different groups of items; four items were categorized as "pragmatic" (*All missiles should be banned; All governments can be overthrown; All dictatorships should be condemned; All weather forecasts should be mistrusted*) and four were categorized as "nonpragmatic" (*All errors should be overlooked; All messages should be ignored; All films should be missed; All books should be put down*). Although the results trended in the right direction (72% of pragmatic items were paraphrased correctly, as compared to 47% of the nonpragmatic items), the variation among items was large, with pragmatic items ranging from 50-100% accuracy, and nonpragmatic items ranging from 19%-69%; and the items were not otherwise matched across conditions. In a follow-up study, Natsopoulos (1985) re-tested pragmatic bias with different types of items (all internally anomalous) and a slightly different task (paraphrase selection), and found no significant differences between items with a weak pragmatic bias and those with a strong pragmatic bias (weakly biased illusions, 45% accuracy; strongly biased illusions, 38% accuracy; see Figure 20). Pragmatic bias was assessed by eliciting belief strength for each item, defined as a measure of how strongly participants held "beliefs and attitudes towards the topic expressed by the sentence under consideration"[14]. Because of facts like these, Wason (1981) has downplayed the role of plausibility and emphasized the need for further research before drawing strong

---

[14] Natsopoulous' measure thus differs from pure plausibility, as it does not assess the degree to which participants agree or disagree with the sentence, only the strength of their opinions about the topic of the sentence in absolute terms. Plausibility hypotheses would predict additional sensitivity to whether the interpretation that participants derive is consistent or inconsistent with their beliefs about the world, although this is not necessarily captured by belief strength.

conclusions. In the set of experiments reported here, we will begin to address this hypothesis in a systematic way.
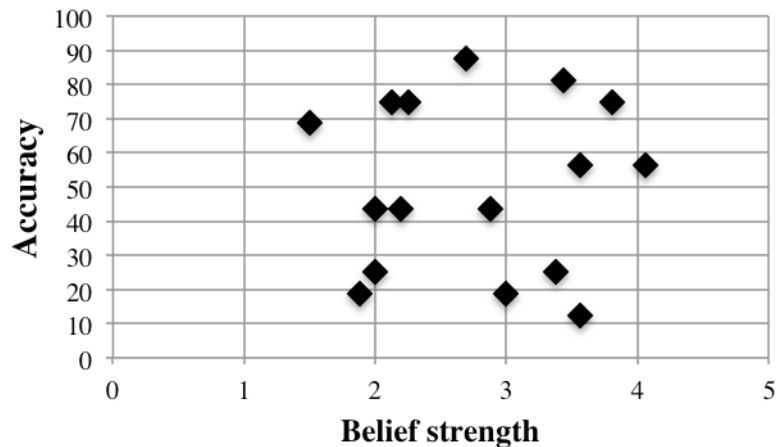
Figure 20. Relationship between belief strength and accuracy in Natsopoulous (1985)

In summary, according to the Channel Capacity Hypothesis, the processing difficulty associated with negation causes extreme computational strain, possibly leading the parser to abandon a compositional interpretation of the illusion sentence. The only way that the sentence can be successfully interpreted is by relying on grammar-independent processing heuristics, such as those that draw heavily on world knowledge. As a shallow processing account, the Channel Capacity Hypothesis thus takes a more extreme view on compositionality at the syntax-semantics interface, by claiming that the interpretation ultimately assigned to the illusion sentence largely bypasses the grammar.

5.2.2 The Change Blindness Hypothesis

A closely related way to account for inversion sentences, if slightly less extreme than the Channel Capacity Hypothesis, is to posit a role for shallow lexical processing, or shallow processing with respect to only certain smaller parts of the representation. What difficult illusion sentences such as *no head injury is too trivial to ignore* may have in common is that

the verb in the *to*-clause tends to be relatively predictable in light of the lexical and grammatical properties of the sentence onset; for example, discussions about the triviality of an injury are most obviously relevant to whether (and how) it should be treated. In other words, successful illusions seem to have high cloze probability: they set up strong expectations about the sentence completion – possibly either down to the lexical level, or to a more abstract conceptual level. Cloze probability often covaries with overall plausibility, which might contribute to the perception that the illusion is driven by plausibility heuristics. However, the two factors are potentially dissociable; for example, the two sentences *The children came inside to play* and *The children went outside to play* are both fairly plausible, but the cloze probability at the verb is likely to be much higher in the latter case (example from Federmeier et al. 2007).

High cloze probability would be especially relevant in cases where there is internal inconsistency, because in this situation the parser will be strongly biased in favor of the wrong interpretation. Because the verb is precisely the opposite in polarity from what the comprehender expects, some amount of surprise would be expected to follow, and prior expectations should be abandoned as inconsistent with the input. However, one might imagine that a sentence onset that generates very strong predictions could be more susceptible than usual to lexical retrieval errors, where words that constitute a partial match to the target are erroneously accepted – as is typically argued to be the case with the Moses illusion. Anomaly detection in the Moses Illusion is famously modulated by the semantic distance between the substitute and target, with numerous studies suggesting that changes to the target are more readily detected with increasing semantic distance between the two words, typically thought to be related to the number of features shared by the word pairs (Erickson & Mattson 1981, van Oostendorp & Kok 1990), and also with the "fit of the word to the basic situation under discussion" (Sanford 2002, p. 202), including the number and strength of associations between words in the sentence. Antonym pairs like *fail/succeed* share a great deal of conceptual overlap, with nearly identical semantic features and conceptual associations, except for their polarity. Indeed, some famous instances of the Moses illusion already involve substitution of an antonym (*After a plane crash, where should the <u>survivors</u> be buried*?). Thus, from a processing perspective, antonyms would likely be considered semantically "close enough"

(Hermann et al. 1979, Rychlak et al. 1989) to the target to yield a Moses-type illusion[15], even though they change the meaning of the sentence in a crucial way.

Inversion sentences are sometimes considered to be a linguistic analogue to change blindness (see discussion in Sanford & Sturt 2002), and perhaps the analogy is motivated by the type of account outlined here: comprehenders fail to detect that the input is no longer consistent with their expectations, because their analysis of the input is incomplete in some critical region. In this case, very strong message-level predictions presumably converge with an attentional lapse while integrating the verb; and, while antonymous verbs yield a crucial polarity change to the global meaning, they may be lexically "close enough" to the anticipated sentence completion that the comprehender simply fails to detect the substitution, proceeding instead with their initial analysis. This type of account is also probably what Pickering & Garrod (2007) have in mind when they point to the classic sentence as evidence that the production system ("emulator") dynamically generates a predicted sentence completion that in certain circumstances can override the actual input.

As far as why negation might be relevant to this process, Pickering & Garrod (2007) claim that comprehenders are especially sensitive to the output of the emulator in cases where the input is noisy (thus accounting for the phonemic restoration effect; Warren, 1970). In addition, prior work on the Moses illusion suggests that the difficulty associated with the negative elements in the sentence could facilitate shallow lexical processing of the verb. For example, Hannon and Daneman (2001) found that working memory load and other individual differences measures predicted susceptibility to the *Moses* illusion. Bohan (2008) has also proposed that lexical anomaly detection decreases as a function of memory load: in sentences like (25) comprehenders detected that *nurses* had been substituted with *patients* 97% of the time in the "low-load" condition, whereas the inclusion of the modifier *possibly quite lengthy and disruptive* in what he terms a "high-load" condition decreased detection to only 52%.

---

[15] Note that studies on the Moses illusion have not actually investigated predictability per se, but rather "contextual strength," a measure of "how many words [are] valuable cues to the targeted answer" (Hannon & Daneman 2001:452); however, this measure seems like it would be closely related to predictability.

(25)   The future of the NHS has been a major electoral issue. There is increasing concern

         from nursing unions that their members are under-paid. UNISON has threatened

         strike action if the government does not improve the present situation. However,

         critics argue that strike action could dangerously affect the people in their care.

         Would you support a national strike, (*possibly quite lengthy and disruptive*), that

         demanded a reasonable pay settlement for all **patients** in NHS hospitals?

Like the Channel Capacity Hypothesis, this type of account suggests that the perception
of the illusion sentence is driven by top-down information, and that this top-down information
becomes especially relevant due to the additional computational burden associated with
processing negation. However, the relevant information is provided by the most likely analysis
that is consistent with the compositional meaning of the onset of the illusion sentence (in this
case, *no head injury is too trivial*), rather than by considering the most likely interpretation
involving the lexical elements, strictly using world knowledge.

5.2.3   The Hypernegation/Ambiguity Hypothesis

Inversion sentences are also occasionally cited in the context of the grammatical
phenomenon of *hypernegation* (Horn 2009; Liberman 2006), i.e., the widespread tendency to
conflate the logical force of multiple negative elements. If such an operation were applied to
the illusion sentence at critical regions, either intentionally or unintentionally, it could
conceivably yield the perceived meaning, although no work has systematically explored this
possibility.

The most well-known case of hypernegation is that of *negative concord* (Labov 1972),
which involves the interpretation of multiple clausemate negative elements as a single instance
of logical negation (Laka 1990, Zanuttini 1991, Haegeman 1995, Zeijlstra 2004):

(26)   Gianni **non** ha visto **niente**.   (Italian)
         Gianni not has seen nothing

(27)   **No** he dit **res**.   (Catalan)

not have said nothing

(28)   Balázs **nem** látott **semmit**.   (Hungarian)

Balázs **not** saw **nothing**

(29)   **Nessuno** ha telefonata a **nessuno**.   (Italian)

Nobody has telephoned to **nobody**.

'Nobody telephoned anybody.'

(30)   T ee **niemand niets** gezied.   (Flemish)

It has **nobody nothing** said.

'Nobody said anything'

(31)   **Personne ne** mange **rien**.   (French)

Nobody not eat **nothing**

'Nobody eats anything'

(32)   **Nikdo n**edá **nikomu nic.**   (Czech)

Nobody neg-gives **nobody nothing**

'Nobody gives anything to anybody'

Negative concord languages can be typologically categorized in various ways (see e.g. Den Besten 1989, Van der Wouden 1994, Giannakidou 1997, 2000); however, negative concord has various properties that are stable crosslinguistically, such as clause-boundedness (Zanuttini 1991, Progovac 1988, Deprez 1997, Giannakidou 1998, 2000, a.o.): a clausal boundary between two negative elements will yield a double negative reading with the force of the two logical negatives independently represented in the semantics, rather than a negative concord reading with only the force of a single negative element:

(33) **Nessuno** ha detto **niente**.

nobody has said **nothing**

'Nobody said anything.'   (= single negation reading)

(34) **Nessuno** ha detto [che non era successo **niente.**]

Nobody has said [that NEG was happened **nothing**]

'Nobody said that nothing had happened.' (= double negation reading)

Negative concord also does not typically obtain with morphological negatives, such as negative adjectives (examples from De Swart 2009):

(35) Il n'est pas incompétent.   (French)

he neg is neg incompetent

'He is not incompetent.'

(36) Non è n'impresa imposibile.   (Italian)

neg is a enterprise impossible.

'It is not an impossible enterprise.'

      The systematic ambiguity of so-called "n-words" in negative concord languages – which sometimes contribute negative force, and other times do not – provides a complication for the syntax-semantics interface that can be resolved in various ways. One way is to localize the problem to the lexical semantics of n-words themselves, i.e. to posit a dual life for n-words as both negative and positive (Herburger 2001). Another way is to assume that n-words are not inherently negative – for example, they pattern with NPIs (Laka 1990, Ladusaw 1992, Giannakidou 1997, 2000), or open-variable Heimian indefinites bound by a higher negative quantifier (Penka 2006, Zeijlstra 2004) – this account, however, has to posit an abstract negative operator in the syntax to account for the negative interpretation of the n-words that do have negative force. Lastly, n-words could be thought to be universally negative quantifiers whose negative force is "absorbed" by a higher element in certain circumstances (Zanuttini 1991, Haegeman 1995, De Swart & Sag 2002), an account that adds the complexity to the compositional mechanisms themselves.

Other documented cases of hypernegation involve the licensing of an expletive or pleonastic negative element within the subordinate clause of a nonveridical environment, most notably the complement clause of a negative verb. Such cases of so-called 'paratactic negation' (Jesperson 1917, Van der Wouden 1994, Zeijlstra 2004, Horn 2009) differ from traditional negative concord in that the conflation of two negative elements occurs across clausal boundaries, and does not depend on the presence of a negative quantifier. Paratactic negation can be found in older varieties of English, such as the example from Shakespeare in (37), as well as a range of other languages, as shown in (38)-(40). In modern standard English, this construction is not fully grammatical and thus sometimes perceived as a speech error, especially in examples like (42)-(45), although it is important to note that this pattern is in fact grammatically licit in other languages.

(37)   First he **denied** [you had him in **no** (= any) right].
        (Shakespeare, *Comedy of Errors)*

(38)   **Prohibieron** que saliera **nadie**   (Spanish)
        forbade that went.out **nobody**
        'They forbade that anybody went out'

(39)   Je **crains** qu'elle **ne** vienne.   (French)
        I **fear** that-she **neg** come
        'I'm afraid that she's coming.'

(40)   **Fobamai mipos** kano lathos.   (modern Greek)
        fear that-not make error
        'I am afraid to make an error'

(41)   **Don't** be **surprised** [if it **doesn't** rain]! (= if it rains)

(42)   I **miss** [**not** seeing you]. (= seeing you)

(43)   The government rushed to investigate the case thoroughly, eager to **dispel** any

notion [that it did **not** take lightly the killing of one of its citizens]

(http://languagelog.ldc.upenn.edu/nll/?p=4393)

(44)   Deliberate **failure** [**not** to make a payment before leaving this site] is a criminal

offense.

(45)   I certainly understand we need to do our best to **keep** people from **not** living on the

streets. (http://languagelog.ldc.upenn.edu/nll/?p=13726)

As Van der Wouden (1994) points out, paratactic negation is also crosslinguistically systematic in terms of its licensing environments – typically limited to the clausal complements of monotone decreasing verbs like *deny, forbid, avoid* and other nonveridical environments; objects of *until, before, without* and *unless*; and complement clauses of comparatives (though puzzlingly, excluded from clausal complements of verbs expressing doubt). This suggests that it can be integrated into a semantic analysis and treated on a par with other grammatical phenomena, with Van der Wouden (1994) advocating for a parallel treatment to NPIs given that both are precluded in environments where two negative elements yield an upward-entailing environment, cf (46)-(47).

(46)   I didn't give her/?nobody any flowers.

(47)   Je **ne** crains **pas** qu'il (*****ne**) fasse cette faute.
         'I am not afraid he will make that mistake'

The expletive negation in the clausal complement is not fully vacuous, either, since it can license NPIs in environments that otherwise preclude them, e.g. in the scope of the verb *miss* (Horn 2009):

(48)   I miss *(not) seeing you around anymore.   (= Horn 2009, ex (9))

Of course, standard English is a language that lacks negative concord and paratactic negation, which gives rise to the perception that (41)-(45) are speech errors where the speaker loses track of the number of negative elements in the sentence or blends two sentence frames (a classic explanation going as far back as Paul 1886), presumably under computational strain. But given that (41)-(45) exhibit properties consistent with a relatively common crosslinguistic pattern, I am more inclined to think they are errors of a very particular and special kind: the use of a grammatical pattern that is in principle provided by UG but unavailable in the language proper.

One might question further whether these examples can be considered errors at all, or are instead symptoms of a more fundamental change to the grammar of English. It is possible to find constructions in English where negative elements add no meaning, which could indicate that the status of these phenomena with respect to English is somewhat unstable. The sentences in (49), for example, are basically synonymous in spite of the fact that one contains sentential negation and the other does not (Partee 2004, Horn 2005). The contrasting examples in (50)-(51), drawn from a web search, show that this pattern occurs consistently with both approximatives *hardly* and *barely* and yields basically no change in meaning, although the speakers I have consulted tend to agree that the cases in (51) sound somewhat more standard.

(49) a. John **doesn't** drink **hardly anything** these days.   (= John rarely drinks)
      b. John drinks **hardly anything** these days.   (= John rarely drinks)

(50) a. I walked over to Madison Avenue and started to wait around for a bus because I **didn't have hardly any** money left and I had to start economizing on cabs and all. (from *Catcher in the Rye*) (= I had hardly any money)
      b. When I first got it the humidifier would not work at all, I put it all the way up at 8 and it **didn't use barely any** water at all. (= it used barely any water)
      c. Well, last week, we texted back and forth Monday through Friday then I **didn't hear barely anything** on Saturday and Sunday. (= I heard barely anything)

(51)   a. There was **hardly any** wind, just a slight breeze.

        b. But even some fans might not recognize the pop star in this photo of her wearing **hardly any** make-up.

        c. There was **barely any** smell.

Similarly in (52)a, the negative force of *hardly* and *no one* are conflated, and the sentence is logically equivalent to the analogous example with *almost* in (52)b, in spite of the fact that *hardly* P is commonly thought to decompose into *almost* not-P (Partee 2004). Note that interpretation of *no one* as a universal quantifier with negation applied to its scope would result in an interpretation of (52)a that has the structure of the examples in (51). This interpretation seems to be obligatory; there is no obvious way to derive a double negation reading where no person *x* is such that *x* knows hardly anything about canines (i.e., every person is knowledgeable about canines).

(52)   a. No one knows hardly anything about canine behavior.

        b. No one knows almost anything about canine behavior.

        b. ∀x: x <u>doesn't know hardly/almost anything</u> about canine behavior

The data above suggest that the negation internal to *hardly* can become semantically inert, and sometimes must; the mechanisms responsible for this seem to be associated with NPI licensing, since the relationship between a higher negative element and *hardly* survives across clausal boundaries in (53), but is disrupted by factive verbs such as *realize* in (54).

(53)   a. I don't think that John said hardly anything today.

        b. No one thinks that John said hardly anything today.

        (= I think / everyone thinks [that John (didn't say) <u>hardly anything</u>])

(54)   No parent realizes that the toddler ate hardly anything today.

(≠ Every parent realizes that the toddler didn't eat hardly anything)

To a lesser extent, the same basic facts seem to extend to the distribution of *yet.* In a 2008 *Language Log* post, Arnold Zwicky observes that *yet* also occasionally embeds under negation with no meaningful change to the semantics (though he finds these examples marginal):

(55)   a. …as the Shipping Board has **(not) yet** to use the lakes in training and
recruiting mariners, but has left that field to the United States Navy.
b. The Texans have been very quiet with their decision and head coach Bill O'Brien
has(**n't) yet** to officially announce Case Keenum as the starter.

This behavior may be related to the fact that *yet* has another life as an NPI:

(56)   John has *(not) won yet.

In fact, parallel to the behavior of *hardly* in (52), embedding *yet* under *no one* again results in conflation of the negative force implicit in *yet* and explicit in *no one*. The meanings in (57) are that no action has been taken, that no one has stepped forward, and that no one has found a solution, rather than e.g., *no action x is such that x has yet to be taken* (= all actions have already been taken):

(57)   a. There are still hundreds of members waiting to be rehired and **no action has yet
to** be taken. (= action has yet to be taken)
b. Never mind that **no one has yet to** step forward with $300 million to $500 million
the club is expected to sell for. (= $300 has yet to be offered)
c. Tchen acknowledges that **no one has yet to** hit on the right solution for women
lawyers of color. (= a solution has yet to be found)

One might hypothesize that something similar is at play in inversion sentences: comprehenders interpret implicit negation as semantically inert (either mistakenly, or not) when interpreting the illusion sentence, thus conflating the logical force of two or more negative elements of the sentence. In other words, the percept of the illusion has a grammatical representation, though not one that formal theory currently predicts. In fact, it might be possible to take the analogy between *too* and *hardly, barely* and *yet* very seriously, particularly given that all three project a scale associated with its propositional argument. The nature of the scale is either explicitly fixed (in the case of *too,* it is provided by the adjective), contextually inferred (in the case of *hardly, barely*) or fixed to the temporal domain (in the case of *yet*). The truth (*hardly, barely*) or falsity (*yet, too*) of the argument proposition is inferred from the position on the scale of the value in the actual world. In sum, it may be no accident at all that elements like *too, hardly, barely* and *yet* behave this way under negation, especially given the relatively deep semantic parallels across the class of words. This behavior, if it is the same across this set of words, could be a peculiarity associated with the grammar of English polarity; or it could be a relatively stable crosslinguistic pattern that, as far as I know, has yet to be described.

This account, of course, presents the following puzzle: why would comprehenders treat inversion sentences as if they were ambiguous – i.e. corresponding to dual readings, one where negation is semantically inert and one where it is not? One possibility is that there is a genuine ambiguity presented by the bottom-up grammatical information, such that the grammar of English indeed permits dual readings, in a way not yet understood. Alternatively, this interpretation may be somewhat accidental, but related to the relative ease of processing multiple negative elements in a concord-like configuration versus as logically independent elements. While I am not aware of any studies that support this claim, a concord configuration would obviously require fewer polarity reversals, and so would naturally be expected to be easier to compute than a double negation reading. If certain combinations of quantifiers are exceptionally difficult to process as logically independent, perhaps they tend to conflate into a concord configuration, even if, strictly speaking, this configuration is not available in the grammar of standard English. In other words, it may be possible that exceptional computational duress leads to the overapplication of an existing rule in grammar that greatly eases computational burden.

5.2.4   Predictions

In the previous sections I sketched out three different ways to interpret the phenomenon of inversion illusions, which, apart from their theoretical orientation, make several distinct, testable predictions that I will begin to address in the next chapter.

First, all three accounts are predicated on the assumption that negation is critical to the illusion, but for slightly different reasons. For the Channel Capacity and Change Blindness hypotheses, it is the raw number of negative elements and the resulting explosion of computational difficulty that likely leads to inattention to the veridical grammatical representation. It is not clear from these accounts whether the illusion should be expected to change when the nature of the negative elements, or their configurational details, are altered; presumably, so long as the level of computational complexity can be kept constant, the illusion should continue to persist at equal rates. Prior work has shown, moreover, that computational costs are incurred by negation of all different types (including both morphologically overt and implicitly negative elements; and those with and without quantificational force), and that it increases consistently with the addition of each successive negative element. This leads us to expect that each of the four negative elements in the illusion sentence contribute individually and additively to its misinterpretation; and possibly, that accuracy should depend in a clear way on the *number* of negative elements in the sentence, as opposed to, for example, their syntactic relation to one another or other specific lexical properties that do not affect computational difficulty. We will begin to investigate this question in Experiment 7 by assessing the extent to which each of the different negative elements affects the illusion.

By contrast, by making a connection to grammatical hypernegation, the Ambiguity Hypothesis constrains the phenomenon considerably more. Hypernegation is sensitive to the nature of the negative elements as well as their syntactic context: negative concord, for instance, requires *logically* negative elements, and does not affect morphologically negative adjectives or verbs, and it is also inherently clause-bounded; NPI-type relations, meanwhile, can usually be disrupted by a class of intervening elements (see Linebarger 1987; Guerzoni 2006). As a consequence, a hypernegation account predicts that two sentences that are associated with equal levels of computational complexity (for example, because they contain

equal numbers of negative elements) may nonetheless be associated with very different patterns of accuracy. Experiment 11 evaluates whether patterns of anomaly detection associated with inversion sentences are related to the possible formation of NPI-type dependencies by investigating accuracy rates for illusion sentences where *no* and *too* are separated by a clausal boundary and factive verb – a syntactic and semantic environment that would preclude such a dependency.

All three accounts allow for top-down information to play some role in the processing of the illusion sentence, but each favors different types of information, and uses the information in different ways. Under the Channel Capacity Hypothesis, there is no fully specified grammatical representation, so the details of the illusion's percept depend entirely on lexical and discourse information – putting the meanings of the words together in a way that makes the most sense given world knowledge. This predicts, for example, that absent special pragmatic cues, inversion sentences will have no consistent interpretation – when all of the possible interpretations associated with the lexical elements are equally plausible, comprehenders could interpret the sentence at chance, or possibly could become aware that they cannot understand what it means. By contrast, the Hypernegation Hypothesis predicts that illusion sentences are sensitive to top-down cues to exactly the same extent that any other ambiguity is sensitive to these cues: factors such as plausibility are widely known to influence incremental processing by mediating the resolution of representational ambiguities. However, these factors do not directly affect the meaning associated with the two different representations; they merely provide evidence in favor of one analysis over the other. Experiments 8-9 build on Wason & Reich (1979) by beginning to unpack the role of plausibility in the illusion, the only top-down cue that has been tested experimentally in prior work, to attempt a systematic exploration of the way world knowledge affects the accuracy and interpretation of the illusion sentences. However, unlike Wason & Reich (1979), we will take a closer look at how plausibility affects sentences analogous to the classic example, i.e. those containing an internal anomaly.

For the Change Blindness account, comprehenders may or may not construct a complete grammatical representation, but they do fail to distinguish systematically between the output of that process and a meaning that they strongly anticipate. This predicts that the misparsing of the critical sentences is fundamentally associated with a scenario where

message-level expectations do not converge with the actual input – i.e., with illusion sentences that are internally inconsistent (*too trivial to ignore*). If strong message-level expectations converge with the actual input, then there is no reason to expect problems correctly understanding the sentence; and if expectations are relatively neutral, comprehenders may be less likely to deviate from the input. In addition, the illusion sentence – when accepted – should correspond to a percept that is mainly consistent with the expected completion of its sentence onset (*no head injury is too trivial…*) as opposed to any meaning that the comprehender finds plausible.

To the extent that the failure to notice the inconsistency between the actual verb completion (*…to ignore*) and prior expectations is associated with shallow lexical processing, accuracy rates might be expected to vary as a function of the semantic match between the expected verb completion and the verb in the actual stimulus. This is because, as discussed above, detection of lexical substitutions (such as in the *Moses* illusion) is modulated by the semantic distance between the two words (Erickson & Mattson 1981, and many others). If participants tend to expect a verb such as *remember* or *recall* and encounter a highly related term – possibly even one opposite in polarity, such as *forget* – they may be more vulnerable to the illusion than if they encounter a term that is less related, such as *discuss.* This type of surprisal could force the comprehender to process the stimuli more carefully, and in doing so, to detect the anomaly. Moreover, the amount of surprisal might vary not only as a function of semantic distance to the expected meaning, but also as a function of the *strength* of expectations: when participants converge more confidently on a sentence meaning, they could be more reluctant to abandon it. Accordingly, the goal of Experiment 10 is to assess on one hand what meaning participants expect from the illusion sentences by asking them to provide sentence completions, and on the other hand to assess how strongly they expect it, by asking them to indicate their confidence in their proposed completion.

Finally, Shallow and Good Enough processing models are rather narrowly focused on language comprehension, given that their chief proposal is that people do not always attend to all available information due to the temporal constraints inherent in the task of language comprehension. As far as I am aware there is no obvious way to extend this model to production, where the speaker presumably has a fairly specific idea of their intended message and also has more control over the amount of time spent on speech planning. As a consequence

these models do not generally predict that non-compositional processing would persist in a production environment. If participants provide responses that set up internal anomalies in the sentence, this suggests a fundamental problem with reading *and* generating the grammar of the illusion sentences, rather than a processing problem specifically related to comprehension. Experiment 10 investigates the robustness of the illusion in a limited production environment, by asking participants to complete the verb in the illusion sentences.

6 ESSENTIAL INGREDIENTS OF INVERSION ILLUSIONS

6.1 Introduction

Given that there has been very little research attempting to disentangle the many potentially critical properties of inversion sentences, the experiments in this chapter are somewhat exploratory in nature, aiming first to establish some basic facts: of the properties commonly ascribed to the classically cited illusion sentence *No head injury is too trivial to ignore,* which are crucial, and which are only incidental properties of the original sentence?

We begin in Experiment 7 by fleshing out the role of negation at three of the critical regions of the illusion sentence: the determiner *no,* the degree quantifier *too,* and the verb *ignore.* We will not directly tackle adjectival polarity in these experiments, in part because categorizing adjectives as negative or positive along the relevant dimension itself presents certain complications. For example, in the linguistic literature a classic adjectival polarity diagnostic relates to behavior in degree questions: whereas positive adjectives yield a neutral reading (*How tall is John?* does not imply that John is tall), negative adjectives yield a norm-related reading (*How short is John?* implies that John is short). Many pairs of antonymous adjectives, however, are both norm-related (e.g., *How precious/insignificant is the memento*?); and, in some cases, the "neutral" antonym is nevertheless affectively more negative than its norm-related counterpart (e.g., *How difficult/easy was the test,* or *How complex/simple was the sentence?*). However, the possible contribution of the adjective with respect to the pattern of results obtained here will be addressed.

In addition to probing the role of polarity, a major goal of this initial set of experiments is to establish the response patterns associated with inversion sentences: what meaning do these types of sentences receive (if they have a consistent meaning)? To what extent can they be detected in the best possible circumstances – i.e. provided specific, concerted effort directed at anomaly detection? When the anomaly is detected, to what extent can participants clearly identify the source of the problem? And finally, in more practical terms, what methodology can be used to investigate the illusion in general? Prior work on inversion sentences highlights the

complications associated with task methodology. The paraphrase task used by Wason & Reich (1979), for example, can yield responses that are difficult to code for accuracy, since participants rarely indicate the internal relation between the adjective and noun on their own, and therefore may provide a correct external assertion in their paraphrase even in cases where they failed to process the internal meaning correctly. In addition, it provides participants no way to indicate when they have detected problems with the sentences, an issue that may be relevant given the oddness of some of the stimuli (e.g., *No film is too good to be missed*). A simple sensicality judgment task on its own can also be insufficient since participants may reject sentences for reasons that are wholly unrelated to the experimental manipulation; or, they may accept the sentences on an unrelated and unintended alternative interpretation. Multiple-choice questions are also problematic in that they potentially draw attention to aspects of the sentence that may have gone unnoticed, or provide participants with unnatural or difficult-to-comprehend answer choices (e.g., Natsopoulos 1985), which may themselves contain multiple instances of negation. Finally, because we do not yet know the details of the illusion's percept, there is some danger of providing participants with choices that fail to match its perceived meaning.

The experiments here use a modified sensicality judgment task, such that participants who indicated that a sentence "made sense" were asked to paraphrase it briefly, and those that indicated that a sentence "didn't make sense" were asked to correct it, or explain the problem. This combination task provides a more precise measure of anomaly detection, and also provides information about what parts of the illusion sentences are most frequently misinterpreted, what the nature of the illusory percept is, and what types of changes participants favor in correcting the error.

6.2 Experiment 7: Effects of polarity

6.2.1 Methods – Experiment 7a

6.2.1.1 Materials & Design

To determine the effect of negation on the illusion sentences, a 2 x 2 within-subjects design was used crossing VERBAL POLARITY (negative verb vs. positive verb) and DEGREE QUANTIFIER POLARITY (*too* vs. *enough*), see Table 13.

	Negative verb	Positive verb
Too	According to the politician, no social program is too wasteful to oppose. [*One should oppose social programs*]	According to the politician, no social program is too efficient to support. [*One should support social programs*]
Enough	According to the politician, no social program is efficient enough to oppose. [*One should not oppose social programs*]	According to the politician, no social program is wasteful enough to support. [*One should not support social programs*]

Table 13. Experiment 7a design.

All 16 target items were illusions; that is, in all conditions, the relationship between the adjective and verb was anomalous. Normal *too* constructions, such as *That missile is too small to be banned,* presuppose that smaller missiles are less likely to be banned (see discussion in Chapter 5)*.* In *enough*-constructions such as *That missile is small enough to allow,* the opposite relation is invoked, so that smaller missiles have a higher likelihood of being allowed. Target items uniformly contained internal anomalies: for example, *too wasteful to oppose* and *wasteful enough to support* are both implausible given that in general, people would be more likely to oppose wasteful things.

In order to separate the effects of negation from the potential effects of plausibility, several measures were taken. First, a short clause like *according to the politician* was added to

175

all items, thus relativizing the statement to an unknown person's belief worlds instead of the actual world. This change introduces more objectivity into the statement, and presumably, prevents participants from relying as heavily on their own opinions about the topic at hand, thus reducing some of the variability. Second, the effect of the two polarity reversals was such that essentially opposite propositions were tested and grouped together in each condition, thus washing out preexisting bias. For example, the two *too*-conditions expressed opposite assertions that *One should oppose social programs* and *One should support social programs.* The *enough* condition also had nearly identical opposing propositions: *One should not oppose social programs* and *One should not support social programs.*

Although various authors suggest that there is a crucial role for the negative verb (Wason & Reich 1979; Cook & Stevenson 2010, Fortuin 2014), it is not yet clear what aspect of verbal negation should be expected to influence the illusion; thus, following prior work, and as an exploratory measure, "negative" verbs included a combination of morphologically negative verbs (*misunderstand, misinterpret, discard, discourage*), verbs denoting "absence of action" (*ignore, overlook*) or that otherwise denote *not*-antonyms (e.g., *reject = not accept, lose = not win, skip = not attend*), and semantically negative verbs that license NPIs in clausal complements (*deny, forget, doubt, fail*). The full list of verb pairs included the following: *fail-pass, ban-legalize, reject-accept, discourage-encourage, misinterpret-interpret, oppose-support, deny-confirm, fail-endure* (in a relationship)*, forget-remember, skip-attend, lose-win, ignore-address*, *overlook-fix, misunderstand-understand, discard-keep, doubt-trust.* Of these sixteen negative verbs, fourteen were independently categorized as negative in the Subjectivity Lexicon (Wilson et al. 2005), while two (*skip, discard*) were not listed.

It is not possible to manipulate verbal polarity without also confounding adjectival polarity or presence of internal anomaly; for example, *too wasteful to oppose* must be compared with either *too efficient to support* (anomalous, with an adjectival confound) or *too wasteful to support* (no adjectival confound, but not matched with respect to internal anomaly). In this experiment we focus on keeping the internal anomaly constant across items. In spite of the confound in adjectival polarity, both classes of adjectives were closely matched for length, frequency (Kucera & Francis 1967), mean reaction time and mean accuracy using the English Lexicon Project (Balota et al 2007), as well as concreteness (Brysbaert et al 2014), with no significant lexical differences across adjective groups. However, we will bear in mind any

possible effects that could be associated with adjective type, with future experiments providing some evidence that this confound does not affect the qualitative patterns reported here (see in particular discussion of Experiment 10).

Each participant saw one version of each item, and the conditions rotated in a Latin Square design. The number of items was kept relatively low in order to prevent overexposure to the illusion sentences and to keep the experiment at a manageable length. The target items were mixed with a set of 18 filler sentences consisting of anomalous and normal *too* and *enough* constructions. All filler sentences had positive quantifiers to offset the target items; some fillers contained obvious internal anomalies (*the jacket is thick enough to stay cool*), externally implausible statements (e.g., *most vegetables are too unhealthy to eat*) and nonsense statements (*several memories were too salty to ingest*). The item set was randomized within blocks, with the order of items constant across all lists.
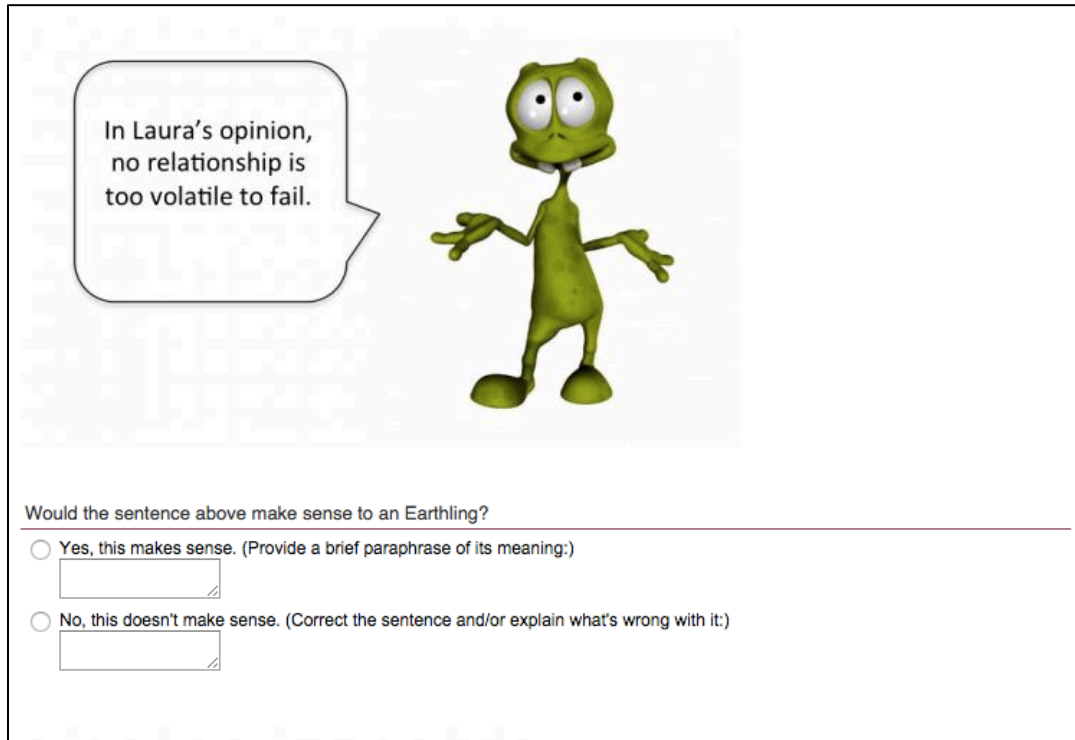
6.2.1.2 Participants

29 participants recruited from Amazon Mechanical Turk were paid $2.00 to complete the experiment on Qualtrics. All participants had a U.S. IP address and a task approval rating of at least 97%, and were self-identified adult native English speakers with no reading or language disorders. The data from one participant were excluded from the analysis for rejecting every sentence in the experiment and providing unclear or nonsensical corrections; the remaining 28 participants were included in the analysis.

6.2.1.3 Procedure

The experiment was administered online and was hosted on Qualtrics. The pretense of the experiment was that participants were to tutor an alien who had to learn how to use *too* and *enough* correctly before visiting Earth. They were instructed to give the alien feedback about his sentences by indicating whether the sentence would make sense to an Earthling. They were encouraged to provide the alien with helpful feedback by briefly explaining the meaning of sensical sentences without using *too* or *enough,* and critiquing or correcting the problem in the

nonsensical sentences. The language tutoring context was selected so that participants would be encouraged to provide specific and informative responses instead of vague critiques (such as, "this sentence is illogical" or "the wording is strange"); and the use of the alien helped to make the anomalous stimuli contextually more appropriate and to make the task somewhat more engaging. A sample trial display is shown in Figure 21.


In Laura's opinion, no relationship is too volatile to fail.

Would the sentence above make sense to an Earthling?

○ Yes, this makes sense. (Provide a brief paraphrase of its meaning:)

○ No, this doesn't make sense. (Correct the sentence and/or explain what's wrong with it:)

Figure 21. Sample trial display for Experiment 7

6.2.2   Results – Experiment 7a

6.2.2.1   Accuracy

Overall accuracy across all conditions was 62% in this study, with a wide range of variation among participants, who responded accurately between 13% and 100% of the time (a much smaller range of variation was observed for items, between 44% and 76%).

Because all of the items contained internal anomalies, in general the responses marked as *doesn't make sense* were coded as *correct*, while those marked as *makes sense* were considered *incorrect*. However, prior to analysis all responses were screened to make sure that *doesn't make sense* responses were motivated for the reasons of interest. Sentences rejected due to e.g. a general dislike of the wording (e.g. '*are' should be 'is'*) without noting the anomaly in the illusion, were marked as *incorrect*, since there was no indication that the participant had detected any other problem in the sentence. Sentences marked as *makes sense* were also screened to make sure participants were in fact susceptible to the illusion, since some participants derived alternative, potentially plausible interpretations of the illusion sentences, and a few explicitly paraphrased the anomaly in the illusion, apparently finding it unproblematic – these responses were marked as *correct*, affecting 30 responses overall. The accuracy data are summarized in Table 14:

	Negative verb	Positive verb
Too	too wasteful to oppose 28.6% (32/112)	too efficient to support 63.1% (41/111)
Enough	efficient enough to oppose 78.6% (88/112)	wasteful enough to support 78.6% (88/112)

Table 14. Experiment 7a: Accuracy by condition.

The responses were modeled using a mixed-effects logistic regression model with the lme4 package in R (Bates, Maechler and Bolker 2013). Fixed effects to model *degree quantifier polarity*, *verbal polarity*, and their interaction were included, along with the maximal possible random effects structure (the slope for the interaction term was omitted for items due to a convergence error). We report here the Wald z-values provided in glmer() output as well as p-values obtained via model comparison using the anova() function.

Overall, participants were substantially more likely to misinterpret sentences when they contained *too* than when they contained *enough* (β = -3.85, z = -4.90, p < .001; $\chi^2(1)$ = 12.97, p < .001). Accuracy did not seem to be generally modulated by verbal polarity (β = -.03, z = -1.33, p = .18), although this fixed effect significantly improved model fit ($\chi^2(1)$ = 7.68, p < .01), most likely because there was a very robust interaction between the two variables: positive verbs were associated with higher accuracy rates in sentences with *too* but not in sentences with *enough* (β = 3.05, z = 3.64, p < .001; $\chi^2(1)$ = 14.67, p < .001).

6.2.2.2   Illusion percept

Items that were marked as *makes sense* were next coded for illusion percept by examining the paraphrases provided by participants. Paraphrases were coded as either *retaining* or *inverting* the veridical meaning of the sentence. For example, the veridical meaning of a sentence such as *no social program is too wasteful to oppose* is that *all social programs should be opposed*. This meaning was sometimes inverted into its opposite assertion, for example *social programs should not be opposed*. 87% of the responses could be coded accordingly; of those, 80.3% (118/147) inverted the external assertion while 19.7% (29/147) retained it. This pattern was descriptively stronger for items with *too,* which were inverted 86% of the time, as compared to items with *enough,* which were inverted 65% of the time.

6.2.2.3   Illusion repair

Responses were next coded for suggested repairs in the cases where a sentence was identified as anomalous (*doesn't make sense*). 9.8% of the responses could not be coded either because the corrections were confusing or vague, or because the participant had correctly paraphrased the anomalous meaning of the illusion sentence without attempting to correct it in any way. The remaining responses were grouped into four categories:

The majority of responses (52%) retained the main components of the sentence and suggested more or less localized changes. Of those, most (44%) favored changing the polarity of the adjective (either by naming its antonym, or by adding *not* before the adjective). 33%

changed the polarity of the verb in the degree clause (again by naming its antonym, or by adding *not*); this was more than twice as likely to happen when the verb was negative (representing 48.2% of the changes to items with negative verbs, versus 21.2% of the changes to items with positive verbs). Only 23% retained the adjective and verb and instead suggested changes to their logical structure. Recall that phrases with *too* introduce particular presuppositions; for example, *too expensive to purchase* implies that more costly items are less likely to be purchased. A response was coded as a change to the logical structure if in such cases participants retained the other lexical items but supplied the opposite relation, often by changing the degree quantifier itself: from *too* ADJ *to* VERB to a variant such as ADJ *enough to* VERB*; so* ADJ *as to* VERB*; sufficiently* ADJ *to* VERB*; had a certain amount of* ADJ *in order to* VERB.

Approximately a third of the codeable responses (32%) pointed out the internal anomaly without suggesting any particular corrections to the sentence, for example: *A safe drug would not need to be banned*; *Stability doesn't mean your are more likely to fail*; *If advice is sound it means that you could trust it*.

14% of the responses provided a broad paraphrase that did not retain all of the elements of the original meaning. Mostly these involved providing an external assertion without incorporating the adjective meaning (*all criticism must be considered*), while a few responses focused only on the adjective (e.g., *all social programs are wasteful*). Several responses dropped the degree quantifier altogether (*some habits can be annoying to discourage*).

Finally, a handful of responses provided corrections or paraphrases that were themselves anomalous (2%), suggesting detection of the problem at some level, with difficulty understanding its precise nature.

6.2.2.4   Summary

The main goal of this experiment was to manipulate the individual negative components of the illusory sentence, to test whether the commonly intuited connection between inversion illusions and negation is indeed warranted. We found that the polarity of the degree quantifier robustly affected the illusion, with substantially lower accuracy rates observed for *too*-sentences than *enough*-sentences. Verbal polarity also affected patterns of

interpretation, although less robustly. Experiment 7b follows up on this line of inquiry by manipulating the polarity of the determiner *no* together with the degree quantifier *too,* with the same experimental paradigm used here; we address the full pattern of negation effects with respect to their implications for the hypotheses outlined in Chapter 5 in section 6.2.5.

6.2.3   Methods – Experiment 7b

6.2.3.1   Materials & Design

A 2 x 2 within-subjects design was used crossing DETERMINER POLARITY (negative vs. positive) and QUANTIFIER POLARITY (*too* vs. *enough*), see Table 15. As before, each participant saw one condition of each item, and the conditions rotated in a Latin Square design.

	Negative determiner	Positive determiner
Too	According to the politician, no social program is too wasteful to oppose. [*One should oppose social programs*]	According to the politician, all social programs are too wasteful to oppose. [*One should not oppose social programs*]
Enough	According to the politician, no social program is efficient enough to oppose. [*One should not oppose social programs*]	According to the politician, all social programs are efficient enough to oppose. [*One should oppose social programs*]

Table 15. Experiment 7b design

All conditions had negative verbs; therefore, in order to contrast *too* with *enough,* these conditions again covaried adjective type (*too wasteful* vs. *efficient enough*). The universal quantifier in the positive determiner condition was counterbalanced, with half of the items containing the quantifier *every* and half containing the quantifier *all.*

The 16 target items were accompanied by 16 filler sentences that varied in plausibility, as well as the polarity of nominal quantifier and degree quantifier. The plausible fillers had a similar syntax to the target items but were not anomalous and had a strong world knowledge bias towards the correct interpretation. The implausible fillers included obvious violations of

world knowledge (*Some cars are colorful enough to go fast; All baseball games are formal enough to wear a tuxedo*).

The procedure for Experiment 7b was identical to that of Experiment 7a: participants were again instructed to provide feedback to an alien learning how to use *too* and *enough* correctly. When they felt a sentence made sense they were asked to indicate what it meant, and when they thought it did not make sense, they were asked to correct it or explain the problem.

6.2.3.2   Participants

29 participants were recruited from Amazon Mechanical Turk and paid $2.00 to complete the experiment on Qualtrics. All participants had a U.S. IP address and a task approval rating of at least 97%, and were self-identified adult native English speakers with no reading or language disorders. The data from one participant were excluded from the analysis for rejecting every sentence in the experiment and providing unclear or nonsensical corrections; the remaining participants (n=28) were included in the analysis.

6.2.4   Results – Experiment 7b

6.2.4.1   Accuracy

Responses were coded as *correct* or *incorrect* as described for Experiment 7a. Overall accuracy rates were 67.1% for this experiment. Accuracy by condition was as follows:

	Negative determiner	Positive determiner
Too	26.8% (30/112)	81.1% (90/111)
Enough	74.1% (83/112)	86.6% (97/112)

Table 16. Experiment 7b: Accuracy by condition.

It is clear from the descriptive results that the *no…too* items again constituted the hardest condition – with less than a third of participants providing correct responses in that condition. Participants were over twice as accurate on the *no…enough* condition, effectively replicating the patterns observed in Experiment 7a. In addition, both conditions with *every/all* had very high accuracy rates regardless of degree quantifier.

The data were again modeled using mixed-effects logistic regression. The model included fixed effects modeling *determiner polarity* (positive vs. negative) and *degree quantifier* (*too* vs. *enough*) as well as random intercepts for items and subjects and a maximal random effects structure. Determiner polarity was found to significantly modulate accuracy rates, with *no* associated with lower accurate rates ($\beta = -1.69$, $z = -2.61$, $p < .01$; $\chi^2(1) = 25.01$, $p < .001$). There was a marginal main effect associated with the degree quantifier, with *too* overall slightly more difficult than *enough* ($\beta = -1.25$, $z = -1.87$, $p = .06$; $\chi^2(1) = 23.84$, $p < .001$), which was qualified by a robust interaction with determiner polarity, such that accuracy with respect to *too* was especially impaired when it combined with the negative determiner *no* ($\beta = -1.59$, $z = -2.27$, $p < .05$; $\chi^2(1) = 4.90$, $p < .05$).

6.2.4.2   Illusion percept

89.8% of the paraphrases of the illusion's meaning could be coded for the percept of the external assertion. As in Experiment 7a, when participants failed to notice the anomaly in the illusion, they had largely inverted its external assertion (representing 68.2% of responses). The vast majority of these inversions were concentrated in the *no…too* condition, where 90.9% of percepts were inverted, compared to 36.3% across the other three conditions.

6.2.4.3   Illusion repair

Of the responses that could be coded for repair type (93.7%), repairs were again categorized into four groups:

31.0% of the time, responses noticed the anomalous relationship between the adjective and verb without suggesting any particular way to change it.

Participants suggested direct and localized corrections slightly less often in this experiment -- 38.4% of the time – although the pattern of corrections largely mirrored those found in Experiment 7a: participants tended to change the polarity of the adjective (41.7% of the time) or verb (38.8% of the time), and less frequently suggested changes to the logical structure of the sentence (19.4% of the time).

29.9% of the time the responses provided a broad paraphrase that did not retain all of the elements of the original meaning, or where it was not possible to localize the changes to any one part of the sentence.

6.2.5 Discussion

Inversion illusions are often thought to arise due to the presence of numerous negative elements in the sentence (Wason & Reich 1979; Horn 2009; Liberman 2009 a.o.), in particular those associated with the nominal quantifier (*no*), adjective (*trivial*), degree quantifier (*too, enough*), and verb (*ignore*). The role of negation, however, has never been systematically tested; the goal of Experiment 7 was to begin to do just this.

As expected, we found that polarity does indeed robustly influence the illusion, with all three components exerting some influence on comprehenders' ability to correctly identify the internal anomaly in the illusion sentence. The effects of negation on inversion sentences are also clearly non-linear: negative elements of the sentence do not have consistent additive effects but rather interact strongly with each other to create an apparent explosion of difficulty at some critical point. Not all negative elements are weighted equally in this interaction; while negative verbs do contribute some difficulty, the illusion relies more crucially on the polarity of the nominal and degree quantifiers, both of which must be negative for the illusion to persist: changing the degree quantifier *too* to *enough,* or the determiner *no* to *all,* takes accuracy rates from the 30% range to the 80% range. Verbal polarity was found to have a more specific effect, only modulating acceptability in *too*-sentences, and not in *enough*-sentences; thus, the combined effect of verbal and quantifier polarity was found to be larger than the sum of each of these two effects alone. This type of superadditivity is consistent with a computational complexity account, such as the Channel Capacity Hypothesis (see Fedorenko et al. 2007; Casasanto et al. 2010 for discussion on superadditive effects). Superadditive effects

tend to arise when two sources of difficulty draw on the same cognitive resources (Sternberg 1969), causing a processing bottleneck at some critical threshold. Thus, for example, center embedding also causes more subtle difficulty until a critical threshold is reached, at which point the sentence simply becomes incomprehensible.

Unlike intractable center embedding, however, it is important to note that the data do not fully support the conclusion that participants find inversion sentences incomprehensible. Comprehension breakdown would seem to logically either imply at-chance performance, with accuracy rates at or around 50% as participants respond to the sentences more or less at random; or else consistent rejection of illusion sentences as obviously impossible to understand and therefore nonsensical. In pilot experiments, the latter was generally the type of reaction associated with complex and computationally intensive constructions; comprehenders simply did not accept sentences as sensical once they reached a certain threshold of difficulty. Illusion sentences, puzzlingly, are associated with a very different reaction: comprehenders overwhelmingly accept such sentences as sensical. In other words, it is not the case that participants are unsure whether there is a problem with the sentence because they are unable to figure out what it means; on the contrary, they seem to be relatively confident that there is *not* a problem with the sentence.

Not only are participants likely to think that there is no problem, but they are also relatively consistent in their proposed interpretation of the sentence: consistently, across Experiments 7a and 7b, they invert its external assertion in the critical condition (changing the meaning of a sentence from e.g. *All social programs should be opposed* to *No social programs should be opposed*). This matches intuitive perceptions of classic examples such as *No head injury is too trivial to ignore,* which is similarly reported to make sense under an inverted interpretation (*No head injuries should be ignored*). However, a crucial difference is that the veridical meaning of the classic example is strongly implausible (# *All head injuries can/should be ignored*), while in this experiment we attempted to avoid strong pragmatic bias, and instead used a set of items that were largely pragmatically neutral, such as (58).

(58) (According to the politician) no social program is too wasteful to oppose.
     a. **Veridical**: All social programs should be opposed.
     b. **Inverted**: No social program should be opposed.

This finding deserves some attention given that it is not immediately obvious why participants favor inversion so heavily if, as Sanford & Emmott (2012) claim, the illusion's percept arises by retrieving a standard scenario involving the noun phrase and "mapping the noun phrases and verb phrases into the slots in the scenario" (p. 28), since a truly pragmatically neutral sentence would then result in inversion only half of the time. In this case, if all of the possible events are retrieved that have a politician as an agent and a social program with some value on the scale of efficiency/wastefulness as a theme, then it will probably be the case that approximately half of those will be *opposing*-events and half will be *supporting*-events. In other words, the scenario-mapping account would likely predict inversion at chance for a sentence that is truly pragmatically neutral. In fact, with respect to *conscious* repair of the illusion sentence, participants favored changing the polarity of the adjective – a change that repairs the internal anomaly but retains the external assertion – which suggests that there is no problem in principle with the meaning of the original external assertion, or with accessing the adjectival antonym. Yet it is clear the critical condition is somehow special in that participants are biased toward accepting the internal anomaly by associated this configuration with a specific interpretation.

In addition, although negation appears to facilitate the illusion, the specific patterns observed here are in some ways inconsistent with prior work on computational complexity effects associated with negation. First, in a series of studies, Geurts & Van Der Slik (2005) showed that patterns of conflicting monotonicity – for example, sentences containing one upward and one downward monotonic quantifier – are more costly to interpret than those with multiple downward monotonic quantifiers. This would predict more difficulty and correspondingly lower accuracy rates on the *no…enough* (- +) condition than the *no…too* condition (- -), opposite of our findings. Second, whatever complexity effects we observe appear to be specific and configurational: not only are accuracy rates *exactly* the same for conditions containing differing numbers of negative elements (*no…enough…neg* and *no…enough…pos*: 78.6%), they are also quite different for conditions containing the same number of negative elements (*no…too…pos*: 63.1%; *no…enough…neg*: 78.6%). This yields a slightly different picture than the additive complexity effects reported by Sherman (1976),

which reveal a consistent linear decline in accuracy as more negative elements are introduced into the sentence, with particular difficulty associated with negative verbs.

The very fact that the anomaly inside the illusion sentences corresponds to a systematic pattern of interpretation seems highly consistent with the Ambiguity Hypothesis, especially in light of the fact that this systematic interpretation is clearly very similar to that of an *enough* sentence. If the illusion sentence is perceived as ambiguous, it seems like that it is driven by the semantics of the degree quantifier, and in particular, the way its implicit negation interacts with other adjacent negative elements, especially the c-commanding negative quantifier *no*. Because the degree quantifier participates simultaneously in the internal and external meaning of the illusion sentence, a change to its polarity will repair both aspects of the sentence at once – in line with the findings. If this semantic instability is related to grammatically-sanctioned hypernegation operations, such as negative concord, such operations can effect a change to logically negative elements but not, for example, morphologically negative adjectives. In other words, the adjective is not a suitable candidate for polarity reversal, and so the percept shown in (59)a would be correspondingly unavailable.

(59)   a. No social program is too <u>efficient</u> to oppose. (*All programs should be opposed*)
         b. No social program is wasteful <u>enough</u> to oppose. (*No programs should be opposed*)

Equally possible, however, is that high inversion rates are driven by the way the sentence onset, *No X is too Y,* affects expectations about the overall sentence meaning. The onset *no social program is too wasteful* leads us to expect a positive opinion about social programs, consistent with the inverted interpretation *No social program should be opposed.* In other words, comprehenders ultimately pursue a meaning they strongly expect, even when it happens to be inconsistent with the input. The sentence is accepted due to an attentional lapse at the verb, with comprehenders largely standing by the meaning they originally anticipated.

However, it is also possible that the items in this experiment were not, in fact, pragmatically neutral in the relevant way, thus resulting in widespread bias away from the veridical external meaning. In particular, although the veridical paraphrase *all social programs should be opposed* is not obviously implausible like the paraphrase *all head injuries should be*

ignored, perhaps the relevant metric concerns the *difference* in plausibility between veridical and inverted interpretations. Thus, if e.g. paraphrases like (59)b were in general more acceptable than those like (59)a, the results of Experiment 7 would follow: the superadditive effects of negation broadly implicate computational overload, and the high inversion rate suggests that there is sufficiently strong pragmatic bias to favor the nonveridical inverted interpretation.

In summary, Experiment 7 confirmed various widely cited intuitions about inversion illusions that have been used to motivate the hypotheses outlined in Chapter 5. First, we determined that both logically negative elements *no* and *too,* and, to a lesser extent, the negative verb, negatively impact the veridical processing of illusion sentences. In addition, while those who detected the internal anomaly tended to repair its meaning with a local change to the adjectival polarity – thus retaining the original external meaning – those who failed to detect the internal anomaly largely reported an inverted external meaning. The remaining experiments in this chapter are aimed at distinguishing among the three approaches outlined above.

Experiment 8 begins by gathering more precise information about the plausibility of the item set in Experiment 7, in order to determine whether and how pragmatic bias might be driving the response patters. We will elicit plausibility scores not only for the veridical meaning of item and condition (e.g., *No social program should be opposed*), but also its various possible inverted meanings (*All social programs should be opposed; No social program should be supported*); these scores can then be related to the patterns of accuracy and inversion in Experiment 7.

6.3   Experiment 8: Effects of plausibility

6.3.1   Methods

In order to gather information about the plausibility of each of the various possible percepts of the stimuli used in Experiment 7, plausibility scores on a scale from 1-7 were elicited for each of the eight conditions shown in (60)-(61), which collectively represent all of

the veridical meanings for the conditions tested in Experiment 7, and all possible inverted meanings for those conditions. The meaning of *too* and *enough* constructions includes an implicit modal quantifier which is ambiguous in force (i.e., can be realized in paraphrases with *can* or *should,* depending on sentence context); as a result, plausibility scores were elicited for both types of paraphrases.

(60)   **Verb: negative**
Veridical paraphrases of: *No social program is too* ADJ *to oppose.*
a. According to the politician, all social programs can be opposed.
b. According to the politician, all social programs should be opposed.

Veridical paraphrases of: *No social program is* ADJ *enough to oppose.*
c. According to the politician, no social program can be opposed.
d. According to the politician, no social program should be opposed.

(61)   **Verb: positive**
Veridical paraphrases of: *No social program is too* ADJ *to support.*
a. According to the politician, all social programs can be supported.
b. According to the politician, all social programs should be supported.

Veridical paraphrases of: *No social program is* ADJ *enough to support.*
c. According to the politician, no social program can be supported.
d. According to the politician, no social program should be supported.

The 2 x 2 x 2 design included two within-subjects factors, MODAL QUANTIFIER (*can* vs. *should*) and DETERMINER POLARITY (*all* vs. *no*), and one between-subjects factor, VERBAL POLARITY (positive vs. negative verb). As a result, a different set of participants provided responses to the paraphrases shown in (60) to those in (61).

The stimuli in this experiment included paraphrases not only of the 20 separate target items used in Experiments 7-8, but also paraphrases of 16 other items intended for use in a separate pilot experiment, for a total of 36 items. The filler sentences consisted of an equal number of strongly plausible (*the pediatrician advises that all toddlers should avoid second-hand smoke*) and strongly implausible paraphrases (*the Buddhist believes that all insects should be disrespected*), so that each list contained a set of items that varied in plausibility. The conditions rotated in a Latin Square design; each participant saw one version of each item. The order of items was consistent across lists, so that only the item conditions changed.

48 participants recruited from Mechanical Turk were paid $0.40 for completing the experiment on Qualtrics. 24 participants rated the conditions in (60), and 24 rated the conditions in (61). All were self-identified adult native English speakers with no known reading or language disorders, with US or Canada IP addresses and a task approval rating of 97% or higher. Participants were instructed to "assess the degree to which the sentence is (in)consistent with [their] knowledge or beliefs about the world, i.e. the extent to which it describes a realistic scenario" and to assign a rating accordingly, using low numbers for implausible sentences and high numbers for plausible ones.

6.3.2   Results

Prior to analysis all ratings were normalized using the distribution of ratings assigned by each participant to all items in the experiment. Each item and condition had two veridical paraphrases: one containing *can* and one containing *should*. We assumed that participants would fill in the implicit modality of a *too/enough* sentence using whichever modality they found to be more felicitous. Thus, for example, the sentence *no test is too* ADJ *to fail* would tend to be interpreted as *all tests **can** be failed,* as opposed to *all tests **should** be failed*, since the first paraphrase is more plausible. Accordingly, the mean plausibility rating for each item and condition was associated with whichever of the two paraphrases was rated higher (in this case, *all tests **can** be failed*).

Next, two different metrics were defined to determine the plausibility of the inverted interpretation of the illusion sentences. The "inverted degree quantifier" percept was defined as the paraphrase corresponding to the equivalent sentence with the polarity of the degree quantifier changed, as shown in (63). The "inverted verb" percept was defined as the paraphrase corresponding to the equivalent sentence with the polarity of the verb changed, as shown in (64). Thus, in sum, a sentence like (62) was associated with plausibility metrics for three different percepts: the *veridical* percept shown in (62)b, the *inverted degree quantifier* percept shown in (63)b, and the *inverted verb* percept shown in (64)b.

(62)   a. *According to the politician, no social program is **too wasteful to oppose.***

b. Paraphrase: All social programs should be opposed.

(63)   a. Inverted degree quantifier: *No social program is wasteful **enough** to oppose.*

b. Paraphrase: No social program should be opposed.

(64)   a. Inverted verb: *No social program is too wasteful to **support**.*

b. Paraphrase: All social programs should be supported.

Overall the target stimuli across both experiments rated fairly high in raw plausibility as compared, for example, to the mean plausibility of *all head injuries can/should be ignored* ($M = 1.67$) and other similarly implausible filler sentences ($M = 2.56$). Table 17 summarizes the plausibility of the veridical meanings of each of the conditions from Experiment 7:

Paraphrase	(Maximal) raw plausibility
'All social programs should be opposed.' NO-TOO-NEG / ALL-ENOUGH-NEG	4.31 (1.11)
'No social program should be opposed.' NO-ENOUGH-NEG / ALL-TOO-NEG	4.85 (.79)
'All social programs should be supported.' NO-TOO-POS	4.74 (1.22)
'No social program should be supported.' NO-ENOUGH-POS	4.0 (.92)

Table 17. Experiment 8: Mean (standard deviation) of plausibility ratings by condition.

Our first goal was to inspect whether in the critical condition (*no-too-neg*) there was any particular bias away from the veridical meaning of the sentence. The veridical paraphrases were found to be slightly less plausible than the inverted paraphrases (veridical: $M = .11$; inverted degree quantifier: $M = .39$; inverted verb: $M = .37$); two-tailed paired t-tests comparing mean plausibility scores by item revealed that this difference was either significant or marginally significant (veridical vs. inverted verb paraphrase: $t(19) = -2.32$, $p = .03$; veridical vs. inverted degree quantifier paraphrase: $t(19) = -1.80$, $p = .087$), suggesting that

there was some moderate pragmatic bias away from the veridical interpretation in this condition, in spite of its overall relative acceptability.

Next, the accuracy data for this condition were pooled across Experiments 7a-7b and mixed logistic regression models were created to assess whether any of the three plausibility metrics had predictive value for accuracy rates or illusion percept. Because the three metrics turned out not to be highly collinear, the data were modeled with separate fixed effects for *veridical paraphrase, inverted degree quantifier paraphrase,* and *inverted verb paraphrase* as well as random intercepts to model variability in means by *experiments, items* and *participants*. Random slopes to model different effects of the plausibility of the veridical and inverted degree quantifier paraphrases by participants were also included (plausibility metrics did not vary within items). Plausibility metrics were each mean-centered prior to analysis to ease model interpretation.

The overall plausibility of the assertion did not substantially affect accuracy (β = .51, z = .5, p = .62; $\chi^2(1)$ = .19, p = .66), nor did the plausibility of the inverted verb meaning (β = .92, p = .30). However, accuracy rates were significantly modulated by the plausibility of the inverted degree quantifier meaning (β = -2.57, z = -2.02, p < .05; $\chi^2(1)$ = 10.11, p = .001). In particular, Figure 22 illustrates a general pattern such that increases in veridical plausibility tended to lead to increases in accuracy, while increases in the plausibility of the inverted interpretation were linked to a decrease in accuracy.
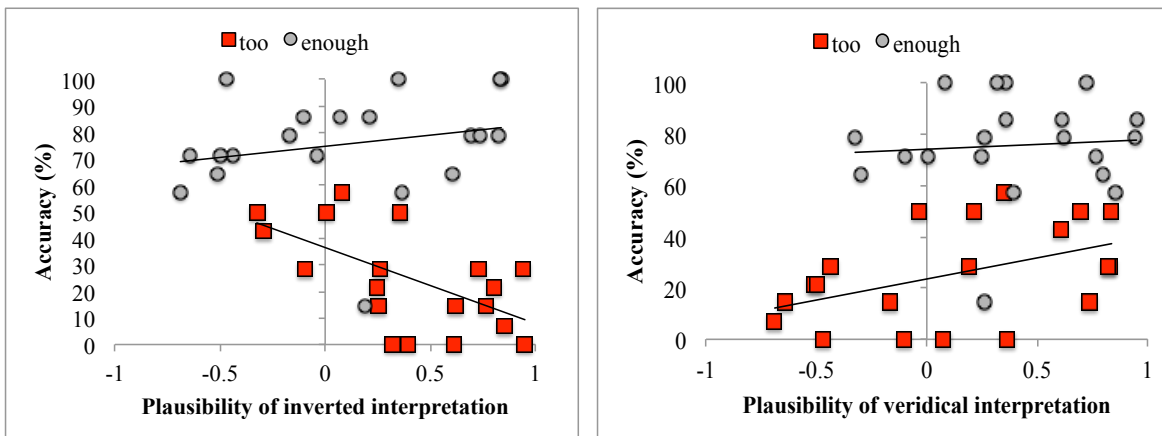
Figure 22. Accuracy rates as a function of the plausibility of the inverted percept (left) and the veridical percept (right).

The plausibility of the inverted degree quantifier meaning had no bearing on the way the illusion's meaning was perceived in the critical condition, however (*p*s > .5), as illustrated by Figure 23 (although visual inspection of the by-items accuracy and inverted degree quantifier plausibility implies a possible relationship in the non-critical condition, *no-enough-neg*):
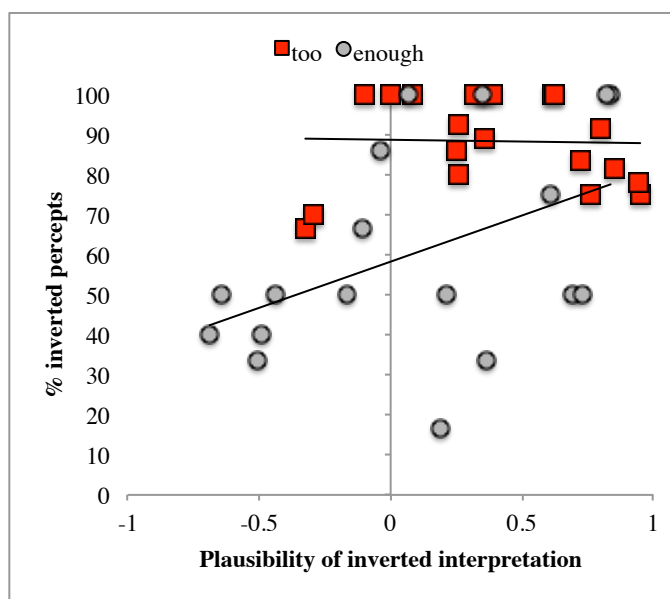

Figure 23. Percept type by plausibility

6.3.3 Discussion

Experiment 7 demonstrated the robustness of inversion sentences in the critical *no-too-neg* condition: comprehension of such sentences is seriously and systematically impaired, with a very strong bias towards accepting the anomalous sentences under an alternative interpretation. If there is total comprehension breakdown resulting from the overwhelming number of negative elements, then an explanation of the systematic acceptance and misinterpretation of these sentences is needed. The goal of this experiment was to determine whether this pattern might be driven by strong pragmatic bias, as suggested by Wason and Reich (1979). To test this hypothesis we obtained plausibility ratings to determine the degree of pragmatic bias associated with the items used in Experiment 7 and used these ratings to investigate whether and how this pragmatic bias affected response patterns.

The veridical external meanings of the experimental stimuli for Experiment 7 were found to be largely quite plausible – placing at the upper end of the scale, and thus clearly not pragmatically infelicitous in the same way as the original sentence *no head injury is too trivial to ignore* (# *all head injuries should be ignored*). This indicates that the illusion is not driven by the need to "pragmatically normalize" a strange external meaning (Fillenbaum 1974), since there is in principle no real problem with the original external meaning of the sentences. However, in spite of the relative acceptability of the target items, Experiment 8 revealed that there was a moderate amount of bias *away* from the veridical meaning in the critical condition, towards the opposite assertion. For example, the veridical paraphrase of *no social program is too wasteful to oppose* (in (65)a), while not implausible, was rated slightly lower than the paraphrases for conditions inverting the polarity of the degree quantifier (*wasteful enough;* (65)b) and verb (*to support;* (65)c), respectively.

(65)   a. Veridical: According to the politician, all social programs should be opposed.
        b. Inverted degree quantifier: According to the politician, no social program should be opposed.
        c. Inverted verb: According to the politician, all social programs should be supported.

This type of pragmatic bias did tend to affect accuracy rates. In particular, in cases where the distracting inverted meaning was very plausible, there was a corresponding decline in accuracy. Although there was no significant predictive relationship between plausibility of the *veridical* meaning and accuracy rates, there was a visually apparent positive relationship trending more weakly towards significance in the regression model, suggesting that accuracy rates may be impacted by the relative plausibility of both the veridical and inverted meanings: the effect of the illusion is likely to be greater for sentences with implausible veridical meanings and highly plausible inverted meanings. Importantly, however, although accuracy rates may be affected in this direction, comprehenders were clearly not very good at interpreting illusion sentences even in the best of circumstances – even items that had highly plausible veridical meanings and relatively implausible inverted meanings were still processed correctly only approximately at chance.

The fact that paraphrases of the positive verb condition (*no-too-pos*) were found to be more plausible than those in the negative verb condition (*no-too-neg*) suggests that at least some of the effects of verbal polarity elicited in Experiment 7a may actually be due to plausibility, rather than verbal polarity itself. Although it is sometimes suggested that verbs like *ignore* and *overlook* are a critical component of the illusion (for example, Wason & Reich 1979; Cook & Stevenson 2010; Fortuin 2014), perhaps because they imply absence of action, which is presumably more difficult to comprehend – the results here argue that their importance is probably more closely associated with their affective negativity: one is more likely to be warned *not* to ignore, miss or overlook something, and thus these types of verbs carry inherent bias towards the inverted interpretation. In favor of this explanation, a simple search on the Corpus of Contemporary American English (Davies 2008) reveals that these types of verbs, when combined with a modal quantifier *can/should,* are almost twice as likely to be negated:

(66)   a. [can/should] (always) [miss/overlook/ignore]: 528 tokens

          b. [can/should] [n't/not/never] [miss/ignore/overlook]: 1023 tokens

The specific plausibility effects observed here, while in some ways reminiscent of those reported by Wason & Reich (1979), differ in one important way. In Experiment 7 all stimuli were internally anomalous, so that accuracy was measured by looking at rates of conscious anomaly detection, rather than by inspecting paraphrases only. Wason and Reich, by contrast, focus mainly on how plausibility affects the percept of supposedly normal, internally consistent *no-too* sentences, by assuming that a veridical external paraphrase corresponds to a veridical internal paraphrase. Because Wason & Reich focus mainly on how plausibility affects the illusion's perceived external meaning, their results can more easily be interpreted as evidence in favor of the Channel Capacity Hypothesis, which posits processing overload followed by retrieval of the most plausible scenario: the perception of the sentence arises from a process that is blind to the internal logical form because that logical form simply cannot be processed; however, because there are sufficient extragrammatical cues, an understanding of the sentence's probable meaning can nonetheless be generated. This type of explanation predicts

196

that extragrammatical cues predict sentence meaning, rather than (or possibly in addition to) rates of internal anomaly detection.

On the contrary, we found no evidence that plausibility impacts the *interpretation* of the illusion sentence, merely that it impacts the accessibility of the veridical grammar. In the cases where the internal anomaly passes undetected, the percept of the illusion does not seem to be associated with the most plausible scenario; rather, comprehenders who accept the illusion sentence almost always perceive its meaning as inverted, regardless of how plausible that meaning is. This state of affairs is at odds with the account described above: if there is no access to logical form, and if real world knowledge also has no impact on the sentence's meaning, then what causes comprehenders to consistently and confidently select the inverted interpretation? In case the lack of relationship between plausibility and illusory percept is a simple issue of power, we retest the effects of plausibility in Experiment 9 by using an item set that is fully balanced, with no pragmatic bias towards or away from the veridical meaning. If plausibility indeed affects the illusory percept in a way that we were not able to detect here, then we expect inversion rates to be truly at chance levels in Experiment 9. This allows us to determine the extent to which an overwhelming preference for the inverted meaning is a more fundamental property of the phenomenon, versus an incidental property associated with the particular item set.

If pragmatic bias does not drive inversion rates – as the results here suggest – then what does? Experiment 9 looks more closely at the connection between inversion and the internal anomaly, which are assumed to be closely connected in the Change Blindness and Ambiguity Hypotheses. We test this by comparing reactions to *no-too-neg* sentences that contain an internal anomaly, and those that do not, as shown in (67). If external inversion rates are associated exclusively with the internal anomaly (i.e., both components of the sentence's meaning covary), these two conditions should be perceived differently, with (67)a perceived to mean that social programs should not be opposed, and (67)b perceived to mean that they should be opposed (the veridical meaning).

(67)   a. According to the politician, no social program is <u>too wasteful to oppose</u>.

b. According to the politician, no social program is <u>too efficient to oppose</u>.

Veridical external meaning: 'All social programs can/should be opposed.'

From the perspective of the Channel Capacity Hypothesis, there is no special reason for comprehenders to perceive the meaning of these sentences any differently, given that the veridical external meanings are exactly equivalent. In either case, comprehension breakdown would be expected due to the presence of multiple negative elements, leading to the adoption of whichever meaning is the most plausible.

6.4 Experiment 9: Effects of internal anomaly

6.4.1 Methods

The effect of the internal anomaly on illusion rates was measured by comparing responses to *no-too-neg* sentences in two conditions: those containing an internal anomaly (*too wasteful to oppose*) and those not containing an internal anomaly (*too efficient to oppose*).

(68)　Internal anomaly:

　　　According to the politician, no social program is <u>too wasteful to oppose</u>.

(69)　No internal anomaly:

　　　According to the politician, no social program is <u>too efficient to oppose</u>.

The internal anomaly arises due to the relationship between the adjective and verb. Normal (non-anomalous) *too*-constructions presuppose that the proposition denoted by the *to*-clause is less likely to hold of objects with greater amounts of the gradable property – i.e., that more efficient social programs are less likely to be opposed. Items with an "internal anomaly" are characterized as those that violate this presupposition. For example, *too wasteful to oppose* is anomalous because social programs that are more wasteful are usually more likely to be opposed. In order to eliminate the anomaly, a change to either element is needed – either the adjective (*too efficient to oppose*) or the verb (*too wasteful to support*). Because verbal polarity may independently modulate illusion rates, the verb was kept consistent across conditions and

the adjective was manipulated instead. This also eliminated the possibility of the two conditions differing in plausibility, since in both cases the external meaning remains the same (*all social programs should be opposed*).

Because the goal of this experiment was to see how the illusion would be interpreted in a pragmatically fully neutral context, it was important to ensure that the set of items was not biased towards or against an inverted interpretation. At this point it is not clear whether the inverted interpretation bears a closer resemblance to the same construction with a change in verbal polarity (*all social programs should be supported*) or with a change in degree quantifier polarity (*no social programs should be opposed*) – although the results here speak somewhat more favorably towards the former – so the set of items were only considered to be pragmatically neutral if the veridical interpretation was not significantly different in plausibility from *either* of the two possible percepts.

Using the normalized plausibility data from Experiment 8, a plausibility score was first obtained for the veridical interpretations and then compared to paraphrases for the two separate inverted meanings; two two-tailed paired t-tests revealed no significant difference between the plausibility of the veridical interpretations ($M = .40$) and the inverted degree quantifier meanings ($M = .28$; $t(9) = 0.71$, $p = .5$) or the inverted verb meanings ($M = .46$; $t(9) = -0.43$, $p = .67$) for the item set, indicating that they were collectively pragmatically neutral.

(70)   *No social program is too _____ to oppose.*
      a. Veridical: All social programs should be opposed. ($M = .40$)
      b. Inverted degree quantifier: No social program can be opposed. ($M = .28$)
      c. Inverted verb: All social programs should be supported. ($M = .46$)

There were five items of each condition for a total of 10 target items. Each participant saw one version of each item, with the conditions distributed across two separate lists. The number of items in the experiment was kept as low as possible in order to keep the task at a manageable length, and to avoid overexposure to the illusion sentences. The 10 target items were accompanied by 20 filler sentences consisting of easy-to-process *too* and *enough* sentences of various types.

The procedure was essentially the same as before: participants were again instructed to provide feedback to an alien learning how to use *too* and *enough* correctly. When they felt a sentence made sense they were asked to indicate what it meant, and when they thought it did not make sense, they were asked to correct it or explain the problem.

25 participants were recruited from Amazon Mechanical Turk to complete the experiment on Qualtrics. All participants had a U.S. or Canada-based IP address and a task approval rating of at least 97%, and were self-identified adult native English speakers with no reading or language disorders. Participants were paid $2.00 for completing the experiment. One participant was excluded from the analysis for rejecting every sentence in the experiment and providing unclear or nonsensical corrections; the remaining participants (n=24) were included in the final analysis. Due to an error in the experimental materials, six participants saw an item in the wrong condition, leading to slightly more data for the anomaly condition than the non-anomaly condition; this affected twelve datapoints.

6.4.2   Results

Prior to analysis, the responses were coded according to whether participants had interpreted the sentence accurately or not. In general, "makes sense" was considered to be an accurate response to the non-anomalous sentences (*no social program is too efficient to oppose*), while "doesn't make sense" response was considered to be accurate for the anomalous sentences (*no social program is too wasteful to oppose*). However, in practical terms, participants sometimes accepted sentences that they interpreted properly, and sometimes rejected sentences for unrelated reasons. 91.25% of the responses were coded as *correct* or *incorrect.* Of the 8.75% (21) responses that could not be coded, most were filtered out because a participant had rejected a sentence because they had found its external meaning implausible.

Overall, participants processed the sentences correctly approximately at chance, with a 47% global accuracy rate, although accuracy clearly varied by condition: 65.9% (67) of the responses to non-anomalous sentences were coded as correct while only 30.8% (36) of the responses to internally anomalous sentences were coded as correct. Viewed differently, comprehenders rejected about a third of the sentences, across both conditions, even though such rejections were justified only in the cases where there was an internal anomaly. Accuracy

rates were modeled using a mixed effects logistic regression model with a fixed effect for *internal anomaly* (anomalous vs. non-anomalous), random intercepts for *participant* and *item,* and random slopes to model the effect of the internal anomaly by participant and item; the results confirmed that accuracy rates in the non-anomaly condition were significantly higher than accuracy rates in the anomaly condition ($\beta = 2.87$, $z = 3.06$, $p < .01$; $\chi^2(1) = 8.85$, $p < .01$).

Of the sentences accepted as "makes sense", the perceived meaning of the sentence depended strongly on the condition. All paraphrases were coded according to whether they retained or inverted the veridical external meaning (*all social programs should be opposed*). Nine responses (6% of the data) could not be clearly coded as such. Inversion was heavily favored in the case of the internally anomalous sentences, which were perceived with an inverted meaning 82.1% of the time. By contrast, the non-anomalous sentences were paraphrased with their veridical meaning a majority of the time (75.4%). A mixed effects logistic regression model with the same parameters outlined above revealed that inversion rates were significantly different across conditions, with inversion significantly more common in the anomaly condition ($\beta = 3.78$, $z = 4.23$, $p < .001$; $\chi^2(1) = 16.91$, $p < .001$).

In cases where the target sentence was rejected, in addition to suggesting changes that effectively repaired anomalous illusion sentences, participants also suggested specific changes to non-anomalous sentences, including changes to the polarity of the adjective or degree quantifier, that rendered them anomalous ("should read [no event is] *too inconvenient to skip*"), and in some cases even suggested that the sentence had an internal anomaly when it did not. Most of the corrections to non-anomalous sentences (62%) did not technically affect the external assertion, suggesting that rejections were not somehow indirectly associated with the plausibility of the sentence but rather specifically related to problems parsing the relationship between the verb and adjective. These types of corrections did not seem to be due to confusion arising from repeated exposure to anomalous illusion sentences, given that rejection rates of non-anomalous sentences was higher in the first half of the experiment (rejection rate: 39.6%) than in the second half (rejection rate: 29.63%); in addition, this type of response was not item-specific, and affected all of the target items to some extent.

6.4.3 Discussion

The goal of Experiment 9 was to determine, broadly, the extent to which the pattern of responses in Experiment 7 are due to the internal anomaly on the one hand, and plausibility on the other hand. Even after controlling closely for pragmatic bias, we found that comprehenders were still clearly biased towards accepting internally anomalous sentences under a particular interpretation – it does not seem to matter how (im)plausible that interpretation is. These findings speak against the feasibility of the Channel Capacity Hypothesis, whose main claim is that comprehension breakdown impairs the generation of a full grammatical representation, so that the sentence can only be interpreted with respect to the most plausible scenario. As we have noted elsewhere, if there is total comprehension breakdown, it is not only puzzling why comprehenders are so confident in the sensicality of the illusion sentences, but it is also puzzling why they are so consistent in their proposed interpretation of an item set, especially one that in this case is pragmatically neutral. Thus, although the bias towards accepting illusion sentences can be modulated by plausibility, it seems more appropriate to think of plausibility as a comprehension *aid*: poor anomaly detection appears to be the rule, although detection becomes slightly easier in cases where the inverted meaning is somewhat implausible.

On the other hand, the presence of an internal anomaly had a substantial effect on the accuracy rates and perceived meanings of *no-too* sentences. When there was an internal anomaly, comprehenders were much less likely to process the sentence veridically, instead tending to accept the internal anomaly and invert the meaning at the same time. By contrast, sentences with no internal anomaly were usually correctly accepted and associated with their grammar-based meaning. This finding provides more evidence that there is a systematic connection between the internal and external perception of a *no-too* sentence: accepting the internal anomaly implies inverting the external meaning, and inverting the external meaning implies the existence of an internal anomaly.

Our findings stand at odds with prior work: while Wason & Reich (1979) found that internally consistent but externally implausible sentences such as *No message is too urgent to be ignored* were paraphrased incorrectly a majority of the time, Experiment 9 found almost no indication that it was possible to accept internally consistent sentences under an inverted external meaning. It is important to note, however, that the task used by Wason & Reich is

202

blind to perceptions of the sentence's internal meaning and overall acceptability. What is likely occurring in these cases is that comprehenders are forced to contend with two situations, neither of which they find ideal: keeping the sentence internally consistent entails accepting an implausible veridical external meaning that *messages should be ignored*; while inverting the sentence's external meaning creates an internal anomaly since the adjective and verb are in the wrong relationship with one another to license that interpretation (# *the more urgent, the more likely to be ignored*). No matter how they contend with the sentence, comprehenders cannot arrive at a pragmatically normal interpretation of it; since Wason & Reich only elicit paraphrases without allowing comprehenders to indicate whether there is a problem with the meaning of the sentence, there is no way for comprehenders to explicitly indicate this type of conflict.

To investigate the feasibility of this explanation, in an informal *post hoc* questionnaire we asked 15 people were asked to judge the sensicality of the sentence *No message is too urgent to be ignored* and then either paraphrase or correct the sentence. Given the opportunity to explicitly reject this sentence, comprehenders nearly always did – in most cases, by mistakenly pointing out that the adjective and verb were not in the right configuration (*If a message is urgent, it should not be ignored*). In other words, rather than pointing out the implausible external assertion, they seem to feel pressure to change the adjective and verb to be consistent with the pattern *the more ADJ, the more V* – which presumably also would allow them to invert the problematic external meaning. Only one of fifteen people genuinely inverted the sentence in the way implied by Wason & Reich (*The sentence makes sense because every message should be taken serious since it can have the possibility of been* [sic] *a life or death situation*). Thus, the pattern of responses reported by Wason & Reich is somewhat misleading – comprehenders clearly feel that something is wrong with this sentence, and while they would like to invert the external meaning of it, they also know that the adjective and verb are not in the right configuration to do so.

From the perspective of the Ambiguity hypothesis, the relationship between internal and external interpretation follows naturally if illusions arise when the negative force either of the degree quantifier or the verb is neutralized – since these elements of the sentence participate simultaneously in both aspects of the sentence's meaning. However, it is also to some extent consistent with the predictions of the Change Blindness hypothesis, where strong

message-level expectations do not converge with the input, but the comprehender fails to notice the change that has occurred to the meaning of the sentence. Different responses to internally anomalous and non-anomalous sentences follow naturally from the different expectations set up by opposite adjectives. The predictions generated by the onset *no social program is too efficient* facilitate the veridical meaning that *all social programs should be opposed,* so that there is no conflict between expectations and the content of the *to*-clause, *to oppose.* The predictions generated by the onset *no social program is too wasteful* on the other hand are at odds with *to oppose,* leading to mismatch and therefore misinterpretation:

(71) a. No social program is too wasteful…   (positive view of social programs)
      b. No social program is too efficient…   (negative view of social programs)

Comprehenders might be thought to largely prefer to retain the meaning consistent with initial expectations, regardless of how plausible it is and regardless of whether it is consistent with the content of the *to*-clause.

The problem with this latter account, however, is that normal (non-anomalous) sentences were also regularly rejected, with participants correcting them so as to create internal anomaly where there was none before. Although there was a clear preference for accepting non-anomalous sentences rather than rejecting them in this way, this pattern of responses was also not rare or isolated to particular people: fully *half* of participants rejected the filler sentence *No situation is too dire to be positive* in Experiment 7b, almost unanimously claiming that *to be negative/pessimistic* was the more sensible completion. The Change Blindness Hypothesis is inherently committed to the claim that illusions arise when comprehenders fail to attend to a critical component of the input, leading to higher-than-expected acceptance rates; however, this is logically distinct from a scenario where comprehenders attend to something in the input that is simply not there, leading to higher-than-expected rejection rates. In other words, people behave as if misanalysis of the *no-too* sentence is pervasive whether or not the sentence contains an anomaly, a scenario most consistent with the Ambiguity Hypothesis.

Experiment 10 addresses a limitation in interpreting the latter result: because Experiment 9 contained a mixture of internally anomalous sentences and internally non-anomalous sentences, it could have induced a level of confusion about the sentence form that

would not necessarily exist otherwise. Experiment 10 tests the Change Blindness Hypothesis using a slightly different approach, one that investigates whether participants actively distort the grammar of *no-too* sentences without ever exposing them to anomalous sentences. In particular, we will look for evidence for the misinterpretation of illusion sentences in production, using a sentence completion task. The version of the Change Blindness Hypothesis we outlined is crucially associated with comprehension failure at the verb or *to*-clause; one way to explain this systematic comprehension failure is by analogy to other well-known cases of incomplete lexical retrieval, namely the shallow lexical processing entailed by the Moses Illusion. Shallow processing, as noted, has more obvious applications to comprehension than production, since it is generally justified by the need to comprehend speech quickly and efficiently. The idea of "shallow processing" with respect to speech production makes less sense, given that people do not usually produce sentences in a grammatically sloppy way in order to save time.

To test whether inversion illusions are a comprehension-specific phenomenon, Experiments 10 uses a cloze task to elicit sentence completions from participants. We then examine the extent to which participants are able to produce sentences that are internally consistent, i.e. the extent to which the internally inconsistent structure is imposed on sentences absent any special exposure to anomalous illusion sentences. In addition the responses from this task will be used to examine whether and how accuracy rates in comprehension are driven by *confidence* in the anticipated sentence meaning along with the degree of *semantic overlap* between the actual completion and the anticipated one, measured by path distance between pairs of word senses in the lexical database WordNet (Fellbaum 1998). In other words, we will assess on one hand *what* meaning participants expect from the illusion sentences, and on the other hand to assess *how strongly* they expect it. Shallow processing is often thought to occur in situations where there is a high degree of confidence (provided by a strong supportive context) coupled with a high degree of semantic similarity: the comprehender knows more or less what to expect, and the substituted word is "good enough" to proceed with their expected analysis. Thus, the Change Blindness Hypothesis might expect illusion rates to vary from item to item, but to peak for items where there are both high levels of confidence as well as high amounts of conceptual overlap between the anticipated completion and the actual input.

6.5   Experiment 10: Effects of task type

6.5.1   Methods

The target items and experimental conditions were adapted from Experiment 7a, except that the stimuli were presented with the final verb missing, e.g. *No social program is too wasteful to _____.* The condition labels in this experiment, however, are complicated by task-specific differences, making it somewhat easier to simply refer to conditions with the letters (a-d). For example, the *negative verb* conditions (a, c) in the comprehension task of Experiment 7a probed how participants dealt with the pragmatically anomalous negative verb ending (e.g., *No social program is too wasteful to oppose*: #one is less likely to oppose wasteful social programs). A pragmatically felicitous ending would have been positive (e.g. *support*: one is less likely to support wasteful social programs). As such, the *target* for correct production will be a positive verb such as *support* for conditions (a, c), and a negative verb such as *oppose* for conditions (b, d). This is illustrated in Table 18 below.

	Correct positive completion	Correct negative completion
Too	(a) No social program is too wasteful to _____ (# oppose; √ support)	(b) No social program is too efficient to _____ (√ oppose; # support)
Enough	(c) No social program is efficient enough to _____ (# oppose; √ support)	(d) No social program is wasteful enough to _____ (√ oppose; # support)

Table 18. Experiment 10 design.

Each participant saw one version of each item; the conditions rotated across four lists in a Latin Square Design. In addition to the 16 target sentences, there were 16 filler sentences, ten containing *too* or *enough* and most with relatively predictable completions (e.g., *Harry was suspicious about the guitar's price; it seemed too good to be _____*). There were no anomalies in any of the filler sentences. The order of targets and fillers was randomized.

Participants were told that their task was to try to predict how a sentence would end. They were asked to select a word to complete the sentence that they felt best reflected what the

speaker intended to say, by providing simple, one-word answers as much as possible. In addition, they were also asked to indicate on an eight-point scale how confident they felt that they were able to identify "what the speaker was trying to say".

40 native English speakers were recruited from Amazon Mechanical Turk to complete the experiment on Qualtrics. All participants had a U.S. IP address and a task approval rating of at least 97% and were self-identified adult native English speakers with no reading or language disorders.

6.5.2   Results

6.5.2.1   Effects of task type on accuracy

Prior to analysis, all responses were first coded for accuracy – i.e., whether or not participants provided a completion that yielded an internal anomaly. In order to distinguish accurate vs. inaccurate responses, two independent coders were provided with a template such as *When a social program is wasteful, one is* <u>more</u>/**less**/*?* *likely to* {fund} *it,* for each response provided in the experiment*. They were instructed to select which captured the meaning better, *more likely* or *less likely;* if they did not know, they were asked to select the question mark. The responses provided were then checked for agreement; any discrepancies were coded as *?*. The agreeing responses were compared with the original degree quantifier in the test item; since *too*-constructions typically corresponded to a *less likely* relation (*this program is too efficient to oppose*: when a program is efficient, one is <u>less</u> likely to oppose it), any items coded with *less likely* were identified as correct for *too* and any items coded as *more likely* were identified as incorrect. *Enough*-constructions correspond to the *more likely* relation (*this test is efficient enough to support*: when a program is efficient, one is <u>more</u> likely to support it) so the opposite schema was used, with *more likely*-responses coded as accurate and *less likely*-responses coded as inaccurate.

6.7% of responses could not be coded for accuracy. The overall accuracy of the remaining responses was 65.4%. The distribution of responses across conditions for Experiment 10 is outlined in Table 19 and compared with the results from Experiment 7a.

Conditions	Comprehension (Experiment 7)	Production (Experiment 10)
(a) No social program is too wasteful to	… #oppose. 28.6%	…#oppose /…√support 35.3% / 64.6%
(b) No social program is too efficient to	… #support. 63.1%	…√oppose / …#support 38.8% / 61.2%
(c) No social program is efficient enough to	… #oppose. 78.6%	…#oppose / …√support 17.2% / 82.8%
(d) No social program is wasteful enough to	… #support. 78.6%	…√oppose / …#support 76% / 24%

Table 19. Comparison of accuracy rates across experiments.

In order to assess what effect the type of presentation had on the illusion, the data from both experiments were pooled and analyzed with a new mixed effects logistic regression model using the lme4 package in R, with fixed effects for DEGREE QUANTIFIER (*too, enough*), COMPLETION TYPE (conditions (a, c) vs (b, d)) and an added factor EXPERIMENT (comprehension/Experiment 7a, production/Experiment 10), and all possible interactions between these factors. In addition, random effects modeling variability across PARTICIPANTS and ITEMS were included. Due to the complexity of the model, a data-driven approach was taken to selecting random slopes, using forward model comparison with the anova() function in R; this justified only the inclusion of random slopes for the effect of verb by items; no other random slopes were found to significantly improve the model fit.

The results show a significant effect of degree quantifier type, with *too* associated with significantly lower accuracy as compared to *enough* (*too:* 49.0%, *enough:* 78.9%, $\beta = -.64$, z = -2.34, $p < .05$; $\chi^2(1) = 97.94$, $p < .001$), replicating the results from Experiment 7. There were several effects that were associated specifically with the *too*-sentences. First, within the *too*-conditions, accuracy rates were lower in comprehension than production ($\beta = -2.10$, z = -4.65, $p < .001$; $\chi^2(1) = 52.72$, $p < .001$; $\chi^2(1) = 3.63$, $p = .06$), although there were no general accuracy differences across experiments. Verbal polarity was also found to modulate accuracy in the case of the *too*-sentences, but not in the case of the *enough*-sentences ($\beta = -1.09$, z = -2.83, $p < .01$), an effect that did not, however, significantly improve model fit ($p = .79$). Finally, the particular interaction between verbal polarity and degree quantifier polarity

208

changing directions across the two experiments: accuracy rates were higher for condition (b) than condition (a) in comprehension, but higher for condition (a) than condition (b) in production (β = 2.95, z = 4.78, p < .001; χ^2(1) = 23.23, p < .001).

6.5.2.2   Semantic distance

Semantic similarity measures were obtained by comparing the semantic distance between the completions provided by participants and the actual verbs from the stimuli in Experiment 7a. The semantic similarity metrics thus indexed the degree to which participants' anticipated meanings matched the input. Semantic distance was elicited using information from WordNet (Fellbaum 1998), a lexical database that organizes words by *synsets* or word senses organized as nodes containing synonymous words. Synsets are related to each other through various semantic relationships (most notably the subset/superset relation) yielding a hierarchically organized structure with specific concepts at the bottom and general concepts at the top of the hierarchy.

There are many available metrics for determining the distance between two synsets in WordNet, each of which conceptualizes semantic distance in a slightly different way (see Pedersen et al. 2004, Meng et al. 2013 for reviews). Path-based measures generally calculate the distance between two nodes. Some of these measures (e.g., wup, Wu & Palmer 1994) are relativized to the specificity of the concepts to reflect the fact that there are intuitive differences in similarity between specific pairs of words, such as *robin/bluejay* and general ones, like *bird/mammal*. Overall specificity may therefore be thought of as related to the depth from the root node to the closest parent node of the two concepts (the *least common subsumer* or LCS). Other approaches measure conceptual specificity using corpus-based similarity metrics (lin, Lin 1998; and jcn, Jiang & Conrath 1997), which assume that more specific concepts will occur less frequently in a sense-tagged corpus such as SemCor (Miller et al 1993).

These different approaches are not always evaluated with respect to how well they model perceptual judgments or behavioral patterns in language processing; rather they may be evaluated on theoretical grounds (Wei 1993, Lin 1998) or with respect to some particular NLP application (e.g. Budanitsky & Hirst 2001). In addition, the success of the measure depends largely on the domain of inquiry, with certain measures performing better than others in

particular contexts. For this reason it has been argued that human judgments are best modeled using a combination of multiple metrics (Ballatore et al 2012), and indeed such an approach has had success at modeling EEG data (Crangle et al 2013). We therefore generated similarity metrics by averaging four different measures, one associated with overall path distance (*path*), two associated with path distance relativized to overall specificity (*wup, lch*) and one incorporating information content (*lin*). These metrics were extracted using the Natural Language Toolkit (NLTK; Bird, Klein & Loper 2009). Certain measures (e.g. *jcn*) were excluded because they yield infinite values in the case of exact synset matches, of which there were many in this experiment.

In lieu of specifying particular senses for each participants' response, similarity metrics were instead derived for *all* possible verbal senses of the response word; the distance to the best fitting sense was selected (i.e., the similarity measure for the closest word sense, as suggested by Resnik 1995). This assumed that, for example, if participant provided a response like *No event was too convenient to miss*, *miss* was likely intended to mean 'fail to attend an event or activity' – a sense highly related to *skip* – rather than 'suffer from the lack of' (e.g., *he misses his mother*). Similarity measures were obtained pairwise for the completion against both the negative and positive verb. For example, if a participant provided the completion *No event was too convenient to miss,* two similarity measures were derived for the fit between *miss-skip* and *miss-attend,* and the maximal of these two values was selected (*skip*).

Figure 24 shows sample similarity scores for one item, calculated for each different type of measure (*path, wup, lch, lin*) and their combination (global mean). For presentational reasons the raw scores have been standardized to make scaling comparable. The metrics do a reasonably good (though clearly not perfect) job of distinguishing between words that are vs. are not synonymous with the target word, and as expected, they do not privilege responses that are exact matches (*keep, discard*) over non-exact matching synonyms (*throw out*, *save*).
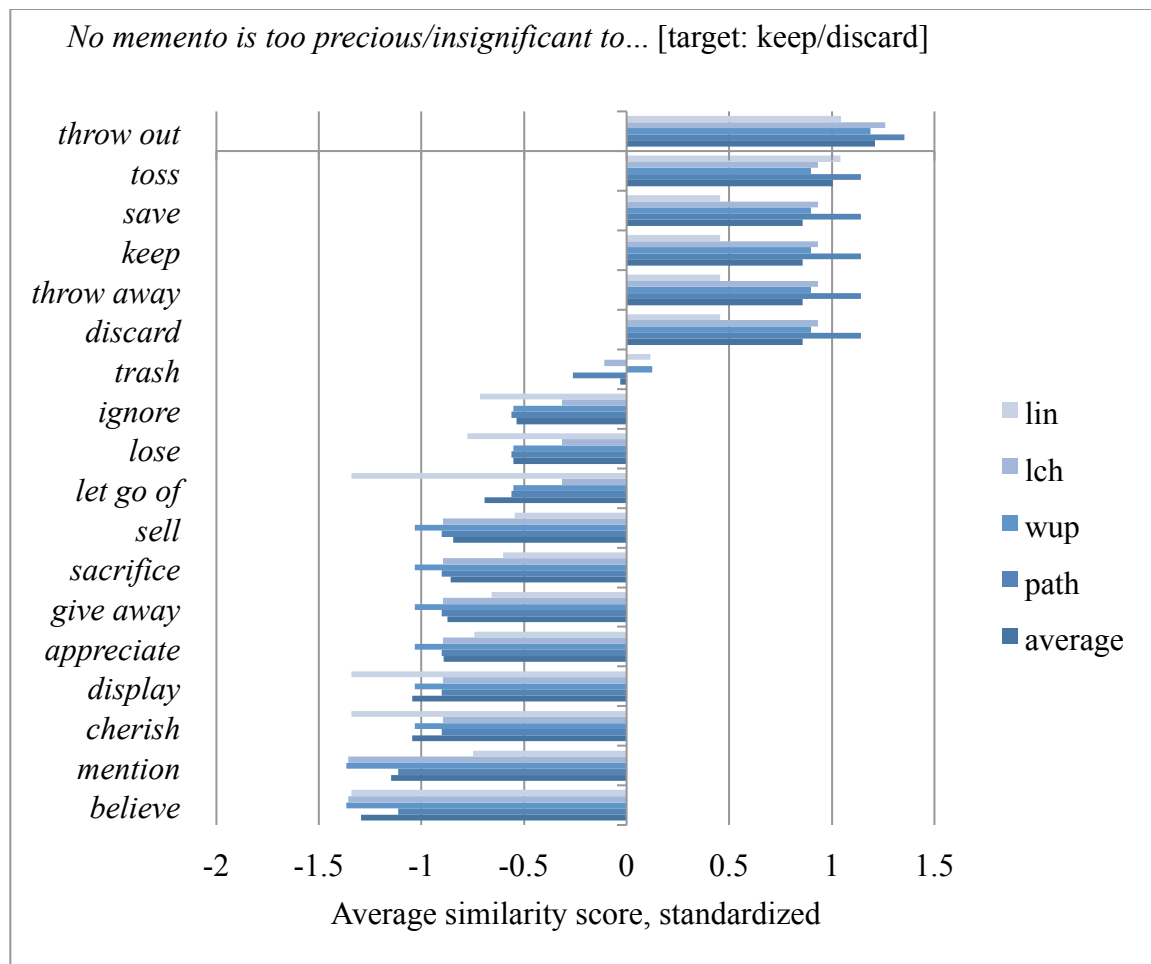
Figure 24. Sample similarity to target, using calculated distance between word senses in WordNet

Averaging over items, similarity metrics were found to have a robustly positive relationship with lexical cloze probability, with semantic similarity increasing with the number of responses that were exact matches to the target ($r(14) = .93$, $p < .001$), although it is visually apparent in Figure 25a that the similarity metrics additionally make finer-grained distinctions among items where there were no exact lexical matches at all, and also collapse distinct lexical items that refer to the same concept (e.g. *toss* and *discard*). There was also a negative relationship with the overall number of unique responses provided for each item, with similarity to the target decreasing as the range of completions becomes more diverse ($r(14) = -.65$, $p < .01$); see Figure 25b.
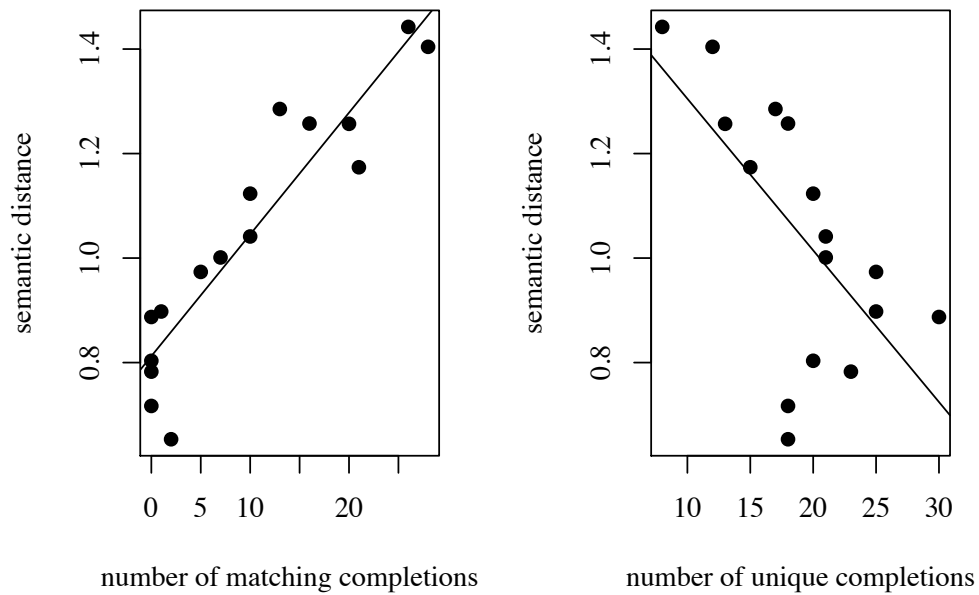
Figure 25. (a) For items where the average similarity to the target was high, participants were more likely to suggest exact lexical matches to the target. (b) For items where the average similarity to target was high, participants also tended to suggest a smaller number of completions (i.e., they were in more agreement with one another).

6.5.2.3   Confidence

One of the goals of this experiment was to evaluate whether the illusion is affected by the strength of expectations about the meaning of the illusion sentence. In this case, confidence measures indicated how sure participants felt that they had identified the speaker's intended message. Averaging across conditions, confidence levels were found to be negatively related to the overall number of unique completions provided for each item ($r(14) = -.53$, $p < .05$), suggesting that confidence scores were increasingly penalized when a greater number of sentence completions were held under consideration, consistent with the experimental instructions.

Within the critical condition specifically, confidence was positively associated with plausibility, with participants' confidence in their completions to items like (72)a related to the plausibility of the meaning suggested by the grammar of the sentence onset, such as (72)b

$(r(14) = .59, p < .05)$.

(72) a. No social program is too wasteful to _____.

      b. No social program is too wasteful to support.

        [*All social programs should be supported*]
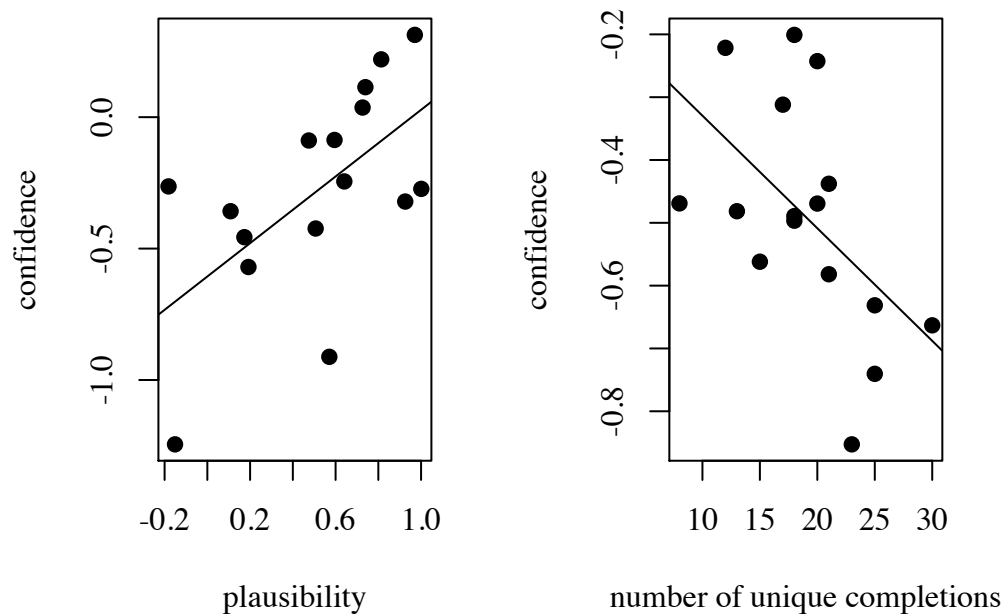
Figure 26. Confidence as a function of plausibility (critical condition only) and variety of possible completions (all conditions).

Overall, participants were moderately confident about the completions they provided for illusion sentences at $M = 3.64$ ($SD = 2.04$), although less so than they were about the filler completions, and were numerically more confident on the trials where they provided a response coded as *correct* (correct: $M = 3.86$, $SD = 2.01$; incorrect: $M = 3.38$, $SD = 2.02$). In order to filter out differences in how participants used the confidence scale, prior to analysis the confidence scores were normalized to z-scores, and adjusted such that values over two standard deviations from the mean were assigned cutoff values.

6.5.2.4   Effects of confidence & semantic distance on accuracy

Predictions: According to the Change Blindness Hypothesis hypothesis, the comprehender generates expectations about the meaning of the sentence and the nature of its completion, and is susceptible to the illusion when they fail to detect the mismatch between expectations and input. Two predictions follow from this hypothesis: first, comprehenders may be more reluctant to abandon a sentence meaning that they are strongly committed to (Christianson et al 2001; Slattery et al 2013) versus one that they are relatively unsure about; and second, comprehenders may be less likely to detect the mismatch between two lexical elements when they have greater amounts of semantic overlap (Erickson & Mattson 1981). Additionally, we reasoned that large amounts of semantic overlap could make sentences especially illusory in cases where expectations are particularly strong: when participants converge more confidently on a completion that is highly related to the target, they will be less likely to abandon it. In other words, the Change Blindness Hypothesis would likely predict a negative relationship between accuracy and both confidence and semantic distance, as well as a potential interaction between them reflecting superadditive effects.

To test this prediction the accuracy data from the critical condition were pooled across three experiments – Experiment 7a, Experiment 7b, and Experiment 9 – and entered into a mixed effects logistic regression model with fixed effects for two centered continuous predictors, the *confidence* and *semantic distance* associated with each item, as well as their interaction term. In spite of their interrelatedness, confidence scores and semantic distance metrics were not found to be strongly collinear, neither in the critical condition ($r(14) = .38$, p = .15), nor across conditions ($r(14) = .19$, p = .47), justifying their inclusion as separate predictors in the model. The accuracy data from Experiment 10 were not included in this analysis due to the lack of any particular hypothesis about whether or how surprisal would affect accuracy in production. Confidence and similarity did not vary by item so only variability in accuracy across subject means was included in the random effects structure; there was not enough data to include all three random slopes by participants, so the interaction term was omitted. As a result, the final model included random intercepts for items and participants, and random slopes modeling variable effects of similarity and confidence across participants.

Table 20 presents the descriptive pattern, by dichotomizing the continuous variables for ease of interpretation into "high/low confidence" and "high/low similarity" conditions.

	High similarity	Low similarity
High confidence	29%	29%
Low confidence	11%	38%

Table 20. Accuracy rates as a function of similarity and confidence (continuous variables are shown dichotomized into high/low bins for clarity).

The analysis generally confirmed the observable pattern in Table 20: confidence and similarity, on their own, did not tend to modulate accuracy significantly (*confidence*: $\beta = 2.01$, $z = 1.8$, $p = .07$; $\chi^2(1) = 1.47$, $p = .23$; *similarity*: $\beta = -.84$, $z = -1.78$, $p = .07$; $\chi^2(1) = .31$, $p = .58$). However, there was a significant interaction among the two variables, with high levels of semantic overlap detrimental to accuracy when participants were not very confident about how to complete the item, but beneficial when participants were very confident ($\beta = 2.36$, $z = 2.16$, $p < .05$; $\chi^2(1) = 5,04$, $p < .05$). In other words, we did observe an interaction between semantic distance and confidence levels, but not one that leads to superadditive difficulty when both are high.

6.5.3   Discussion

The experiments in this chapter have repeatedly found that perceptions of the internal and external meaning of a *no-too-neg* are correlated. In most cases, sentences with *too* presuppose that objects ranking higher on some scale are less likely to be predicated of the verb in the *to*-clause: social programs that are *more* efficient are *less* likely to be opposed. When the adjective and verb support this relationship – as in (73)a – comprehenders tend to construe the sentence veridically. When the adjective and verb are more plausibly related to each other as *the more ADJ, the more V* (the more wasteful a program is, the more likely it is to be opposed), as in (74)a, they also tend to report an inverted external perception of it, such as (74)b – even when world knowledge provides no special bias in favor of this meaning.

(73)   a. No social program is too efficient to oppose. [THE MORE ADJ, THE LESS V]

b. Veridical percept: All social programs should be opposed.

(74)   a. No social program is too wasteful to oppose. [THE MORE ADJ, THE MORE V]

b. Inverted percept: No social program should be opposed. / All social programs should be supported.

We entertained the possibility that this interpretational correlation was due to the different expectations generated by the sentence onset in either case. Sentences with *too* can be readily interpreted even without a *to*-clause when its content is highly predicable: *No mountain is too high, No price is too high, No task is too big.* Similarly, the sentence onset *no social program is too efficient* can be generally understood as expressing a negative opinion about social programs before the *to*-clause is encountered. Illusory perceptions of the sentence, then, may be related to attentional lapse: comprehenders who strongly anticipate a meaning and fail to fully attend to the details of the input continue to pursue their original analysis, even when it is inconsistent with the input. For example, participants may be so confident in the anticipated meaning that they fail to discard it, or else the final verb might be considered semantically "close enough" to its antonym so as to evade awareness – thus making inversion sentences fundamentally similar to shallow lexical processing in the Moses illusion.

To assess this hypothesis we gathered several types of data about illusion sentences. First, by asking participants to supply the sentence completion, the task used in Experiment 10 drew explicit attention to the critical region of the sentence; if the illusion were caused by shallow processing at the verb, for example by strong prior expectations causing the system to "override" the actual input, then the illusion should more or less disappear in a task that allows participants to provide their own verb. Indeed, it is unclear whether shallow processing mechanisms are warranted in production, and if so, how they work and what exactly they accomplish. Second, the fill-in-the-blank task provided us with a means to assess the extent to which expected responses fit with the actual sentence ending – thereby providing some index of semantic surprisal. We also asked participants to indicate how confident they were in their completion, since very strong contextual predictions could lead to shallow lexical processing

and therefore decreasing anomaly detection. These two measures could plausibly interact with one another; for example, high levels of confidence might lead to increased anomaly detection in cases where the sentence ending is a poor match with expectations, but decreased anomaly detection in cases where the sentence ending is a "good enough" fit.

In spite of the task change, overall accuracy rates in Experiment 10 were only minimally better than Experiment 7a, suggesting that misrepresentation of the illusion sentence is modality independent and occurs irrespective of the attention devoted to the critical region. In addition, cloze results and confidence scores failed to provide evidence for the Change Blindness hypothesis. Confidence measures, which were broadly related to the diversity of completions offered, tended to be *positively* associated with accuracy (although not strongly so), thus aiding comprehenders in correctly parsing the grammar rather than biasing them towards "overriding" the grammar. In addition, although high semantic overlap between the expected completion and the input did decrease accuracy, this effect was strictly limited to situations where participants were not very confident in how the sentence would end. This effect likely reflects the difficulty inherent in disentangling the various possible top-down effects, all of which are loosely correlated. For example, confidence scores were also moderately related to plausibility, so that comprehenders tended to be less confident about completions that yielded less plausible meanings. In the low-confidence, high-similarity cases, comprehenders are clearly converging on similar completions, so they are likely penalizing confidence scores primarily because they find the predicted meaning somewhat implausible and therefore do not weight it as heavily as a highly predicable *and* highly plausible completion. In other words, the interaction between confidence and similarity is probably more closely tied to the plausibility facts already observed than the particular top-down effects associated with the Change Blindness Hypothesis.

It is possible that these types of factors might wield greater influence on accuracy rates in a task that more readily facilitates shallow processing, instead of the somewhat intensive and time-consuming paraphrase task conducted in Experiment 7. However, this then would still support only an ancillary role for cloze bias and would not explain the *core* phenomenon underlying inversion sentences – since the illusion clearly persists robustly even in a context that explicitly warns participants of the possibility of sentence anomalies; primes them with a number of other *too* and *enough* sentences, including several with easy-to-detect anomalies of

a similar type (e.g., *The jacket is thick enough to stay cool*); and requires them to consider the LF of the sentence in detail by identifying particular locations where the meaning is wrong. However, it is of course also possible a different method for measuring semantic similarity, for example statistical techniques for assessing patterns of co-occurrence of lexical items across collections of documents (Latent Semantic Analysis), might have greater success.

To summarize, contrary to the predictions of the shallow processing models, this experiment found that inversion illusions persist in production, and are not affected in a clear way by cloze bias. In general, throughout the experiments in this chapter we have found very little evidence that would suggest the illusory percept is determined through top-down cues related to world knowledge or sentential predictions. The goal of Experiment 11 is to shift focus to the predictions of the Hypernegation Hypothesis, which explains inversion illusions by way of a systematic grammatical ambiguity allowing the negation implicit in *too* to be interpreted as semantically inert, as if it were an NPI licensed by the negative determiner *no.*

Effects of verbal polarity in production

Experiment 10 elicited an interesting reversal of the effect of verb polarity. Whereas condition (a) was more difficult than condition (b) in Experiment 7a, precisely the opposite was true in Experiment 10. This effect, however, is consistent with an overall bias against negative verbal predicates, whether due to their semantic negativity or their affective negativity (thus leading to lower plausibility). In Experiment 7a, participants found it easier to detect the anomaly when the sentence contained a positive verb (condition b) than a negative one (condition a). This effect reverses in production: accurate production of a negative verb in condition (b) was more difficult than the accurate production of a positive verb in condition (a). Thus although the results seem contradictory, they are actually largely consistent in view of the different tasks: comprehenders are biased against comprehension of negative verbs, and speakers are biased against production of negative verbs.

Finally, the results here rule out the possibility that the verb polarity interaction observed in Experiment 7a is actually an adjective polarity interaction. Due to inherent properties of the items, the adjective types could not be matched across items, resulting in an unavoidable confound. One possible interpretation of the data in Experiment 7a is that the

degree quantifier enters into an interaction with the adjective, not the verb (or that all three elements are associated with the dramatic drop in accuracy in the critical condition). The results from Experiment 10 speak to the fact that the interaction is primarily associated with the properties of the verb, given that the adjectives that were presented were identical across conditions in the two experiments and yet the pattern of results changed so drastically in the opposite direction.

6.6   Experiment 11: Effects of NPI intervention

The Hypernegation Hypothesis differs crucially from shallow processing accounts in that it expects the illusion to be modulated by the grammatical factors that govern NPI licensing, negative concord, or paratactic negation. All of these phenomena involve the grammatically sanctioned logical conflation of two negative elements into one, sometimes called *duplex negatio negat* (Horn 2009). For this account it is not the sheer quantity of negation in the sentence *per se* that yields the illusion, but rather the specific relationship between the negative elements. As a result, it should be possible to elicit very different patterns of accuracy for sentences containing precisely the same number of negative elements.

Experiment 11 will test the effect of the illusion in an environment that precludes the formation of local and non-local dependencies between *no* and *too,* the negative elements that are the most critical to the illusion. Chapter 5 noted that negative concord, a local relation between two negative elements, is crosslinguistically clause-bounded: in languages that allow two negatives to be conflated into one in logical form, these negative elements must occur within the same clause. NPI dependencies are not clause-bounded in the same way as negative concord; rather, negative polarity items are broadly constrained to downward-entailing (Ladusaw 1980) or nonveridical (Giannakidou 1998, Zwarts 1995) environments, which would be introduced by the negative determiner *no* in illusion sentences, thus allowing inferences from set to subset in (75).

(75)   a. No book is too long for Phil.

       b. No book is too long for Phil *and also* too expensive for Sally.

"Factive" verbs (Kiparsky & Kiparsky 1971) such as *know* or *realize* are strongly veridical, presupposing the truth of their complement clause and thus precluding NPI licensing in their scope (see among others Giannakidou 1999, 2006; De Cuba 2007; Homer 2008 on the relationship between factivity and NPI licensing), even when there is a c-commanding negative item like *no* or *not*. Thus, in (76) the NPI *a red cent* is licensed in the clausal complement of *don't think,* but not the factive *don't know*[16]:

(76)　　a. I didn't think that John had a red cent.
　　　　b. * I didn't know that John had a red cent.

Experiment 11 focuses on minimal pairs similar to (77). In (77)a, unlike (77)b, *no* and *too* are separated not only by a clausal boundary but also by a factive verb, thus precluding negative concord and NPI licensing between the critical elements *no* and *too*:

(77)　　a. *No* politician <u>knew</u> that the social program was ***too*** wasteful to oppose.
　　　　b. The politician <u>knew</u> that ***no*** social program was ***too*** wasteful to oppose.

　　　Shallow processing accounts of the illusion do not make any special predictions about how clausal boundaries or factive verbs affect illusion rates. Indeed, if the effect of negation is associated with computational overload, it is unclear that one would want to make such predictions, since negative elements distributed across separate clauses, such as (78), remain intuitively difficult to understand:

(78)　　**No** senator realized that the president was **not** aware that **no** citizen would vote for
　　　　the health law.

---

[16] Among the class of factive verbs, NPI licensing is differently affected by emotive factive verbs such as *regret* or *be surprised* (Linebarger 1980)*,* which license some weak NPIs, versus epistemic factive verbs such as *know,* which do not. By and large, the items used here are drawn from the second class of verbs; see the Appendix for specific item properties.

In light of the results from Experiment 7-10, Experiment 11 takes for granted the interpretational patterns associated with anomaly detection; accordingly, we switch from the modified paraphrase task used in Experiments 7-10 to an anomaly detection paradigm whereby participants are asked to pick out sentences that contain word substitutions. This task, which is much less time-consuming, allows for the inclusion of a broader range of sentences to mask the illusions, including semantic illusions of other types.

6.6.1   Methods

Experiment 11 used a very simple within-subjects design consisting of two conditions – *licensing* and *non-licensing* – repeated here in (79). In the *non-licensing* conditions, the negative determiner *no* was situated in the subject of the higher clause, separated from *too* by the factive verb and clausal boundary; in the *licensing* condition the negative determiner was situated in the embedded clause instead. In both cases the truth of the complement clause was presupposed by an embedding factive verb, such as *know*. Other than the order of the nominal quantifiers *no* and *the*, the two conditions contained the same lexical elements in the same syntactic configuration, so that the quantity of negation did not vary across conditions.

(79)   a. NON-LICENSING: ***No*** politician <u>knew</u> that the social program was ***too*** wasteful to oppose.

b. LICENSING: The politician <u>knew</u> that ***no*** social program was ***too*** wasteful to oppose.

All of the target stimuli contained internal anomalies, so that the relationship between adjective and verb presupposed by *too* (the more ADJ, the less V) was violated; in addition, all sentences introduced the negative determiner *no,* the negative degree quantifier *too,* and a negative verb. In other words, aside from the added syntactic complexity associated with the embedded

clause, the target stimuli in the dependency condition were basically analogous to the sentences from the critical condition in Experiment 7.

As a result of the experimental manipulation, the external assertions associated with the embedded clause were slightly different across conditions. In order to minimize possible pragmatic effects on accuracy rates, the stimuli were normed by eliciting plausibility ratings for eight different paraphrases per item, varying the modal quantifier (*could* / *should*), the condition (licensing / non-licensing), and the paraphrase by condition (veridical / inverted). 32 participants rated the plausibility the sentences like (80)-(81) on a scale from one to seven.

(80)   Paraphrases, non-licensing condition:

a. No politician knew that the social program could not be opposed. [verid., *could*]

b. No politician knew that the social program should not be opposed. [verid., *should*]

c. No politician knew that the social program could be opposed. [inverted, *could*]

d. No politician knew that the social program should be opposed. [inverted, *should*]

(81)   Paraphrases, licensing condition:

e. The politician knew that all social programs could be opposed. [verid., *could*]

f. The politician knew that all social programs should be opposed. [verid., *should*]

g. The politician knew that no social program could be opposed. [inverted, *could*]

h. The politician knew that no social program should be opposed. [inverted, *should*]

Four scores per item were selected, corresponding to the plausibility of the veridical and inverted meanings of each condition, respectively, using the mean rating for whichever modal quantifier was found to be more plausible. The final item set was then selected from the normed stimuli in such a way as to minimize differences in veridical plausibility and inverted plausibility across the two conditions. A 2 x 2 repeated measures ANOVA, comparing the by-item mean plausibility scores, crossing *paraphrase type* (veridical vs. inverted) with condition (licensing vs. nonlincensing), found that the inverted paraphrases of the final item set were rated marginally higher than the veridical paraphrases in both conditions, but that there were no general plausibility differences across the two conditions, and the differences between veridical and inverted paraphrases also did not vary across conditions (*p*-values > .5). The veridical

paraphrases of the licensing condition were actually numerically *higher* (M = .35) than those for the non-licensing condition (M = .26) (though as mentioned, not significantly so) – as a result, to the extent that accuracy rates across conditions are differently affected by plausibility, they will trend towards higher accuracy in the dependency condition, biasing the results *away* from our hypothesis.

The 10 target stimuli were accompanied by 60 fillers of a variety of types, approximately half of which were anomalous. The target items rotated in a Latin Square design such that participants saw one condition of each item; the order of targets and fillers was held constant across the two lists. Ten fillers corresponded to very complex reversed-role passive sentences, with the thematic role reversal rendering the sentences implausible (e.g., *The commission said that the reports stating that the **soldier** was protected by the **child** in the battle were submitted too late to be taken into consideration*). Ten fillers had Moses-type lexical substitutions (*After **Snow White** pricked her finger, she slept for 100 years*), rendering those sentences false. The remaining fillers consisted primarily of non-anomalous *too* and *enough* sentences, and non-anomalous counterparts to the reversible passives and Moses illusions. Four of these fillers were sentences from the non-critical conditions in Experiment 7, which were included in order to eyeball in general the extent to which the task change would affect accuracy rates, if at all.

26 participants recruited from Mechanical Turk were paid $1.25 to complete the experiment on Qualtrics. All were self-identified adult native English speakers with no known language or reading disorders; participation was automatically restricted to workers with a US-based IP address and a 97% approval rating.

Participants were told that the experiment tested how easily people detect problems with very complex sentences. They were told that in some of the sentences, "one or more words have been swapped for similar, but incorrect, alternatives – rendering the sentence false or nonsensical." Their task was to detect which sentences contained these substitutions by selecting either of two options – either "There are no problems with this sentence" or "The following word(s) are incorrect:" – and if applicable, identifying the problem word(s) by typing them into an adjacent text box. Participants were told that some of the problems were tricky to identify and not to worry about exhaustively identifying all of them, or about editing

for grammar or clarity, but to simply read sentences carefully and note the times when they observed an obvious problem.

6.6.2   Results

Overall, the illusion sentences in this experiment were rejected 50% of the time. By contrast, complex role-reversal sentences such as *The nurse forgot that the letter saying that the doctor was treated by the patient in the surgery had already been sent to the insurance company* were rejected 42% of the time; *Moses* illusions (such as *After Snow White pricked her finger, she slept for 100 years*) were rejected 46% of the time. Items in the non-licensing condition were rejected almost twice as often (64% of the time) as those in the licensing condition (33% of the time). (For reference, the four *no…enough…neg* and *all…too…neg* filler items were rejected at an approximately similar rate as the non-dependency items, at 71% of the time, suggesting overall lower accuracy rates associated with this task). A mixed effects logistic regression model was fitted to the data with one fixed effect to investigate the possibility of *licensing* (licensing vs. non-licensing) on rejection rates. Random intercepts for participants and items were also included, along with random slopes for the effect of the condition across participants and items. The model confirmed that the possibility of NPI/concord licensing robustly modulated rejection rates for illusion sentences, with comprehenders much more likely to reject sentences where semantic dependencies between *no* and *too* were precluded (β = -1.80, z = -4.39, *p* < .001; χ^2(1) = 12.76, *p* < .001).

When participants indicated that there was a problem with the sentence, they were asked to indicate the word or words in the sentence they found to be problematic. In the majority of cases, participants who rejected illusion sentences pointed out that the adjective was incorrect (57% of corrections), mirroring the pattern of suggested corrections in prior experiments. In a smaller number of cases participants pointed to the verb (35% of the time) or degree quantifier (29% of the time), and at times rejected the negative determiner *no* (21% of the time) even though a change to the determiner would not necessarily bear on the internal anomaly in the sentence.

6.6.3   Discussion

Experiment 7 showed that inversion illusions are somehow caused by the quantity of negation in the sentence, especially in terms of the nominal quantifier *no* and degree quantifier *too*. The hypotheses outlined in Chapter 5 differ in terms of whether the effects of negation are broadly associated with computational complexity, versus whether they are more specifically associated with the logical conflation of two negative elements using operations made available in the grammar. The latter account predicts that any dependency between *no* and *too* will be governed by those properties that crosslinguistically influence negative concord and negative polarity licensing: the former is subject to structural constraints such as clause-boundedness, while the latter is heavily influenced by semantic properties such as nonveridicality (Giannakidou 2002), downward-entailingness (Ladusaw 1980) or Strawson entailment (von Fintel 1999). A pure computational complexity account, by contrast, is unlikely to predict such differences given that the quantity of negation remains the same either way. In fact, structural properties such as clause-boundedness should be by definition irrelevant if the grammatical parse is terminated and the illusory percept arises solely through extragrammatical inferences (the Channel Capacity Hypothesis).

Experiment 11 found evidence that the properties that influence concord and NPI licensing indeed strongly affect illusion sentences. A clausal boundary and intervening factive verb preclude possible negative dependencies between *no* and *too,* which apparently makes the internal anomaly much easier for comprehenders to detect. This provides strong evidence in favor of the Hypernegation Hypothesis over the shallow processing approaches, since it is the only approach that postulates a percept that is generated via existing grammatical mechanisms. To the extent that the illusion arises due to the complexity associated with negation, the evidence here suggests that that complexity is specifically associated with the processing of two negative elements as logically independent in an environment that has the right grammatical properties to support *duplex negatio negat.*

225

6.7   General discussion

6.7.1   Towards a theory of inversion sentences

Inversion illusions, like Escher sentences, seem to be telling us that comprehenders can arrive surprisingly easily at interpretations that are at odds with the output of logical semantics. They differ from Escher sentences, however, in that the structure underlying this illusion is entirely licit from the perspective of the grammar; in principle, composition should proceed unproblematically, although its output may be pragmatically odd. Nevertheless, errors are frequent, and they persist unless memory-intensive and conscious effort is applied, with many comprehenders failing to ever come to terms with the compositional meaning.

This chapter set out to disentangle some of the critical properties of inversion illusions, exploring how these properties interact with each other and whether and how they fit with the various proposals outlined in Chapter 6. All of these accounts are based on the same fundamental observations about the classic sentence *No head injury is too trivial to be ignored*: first, there is an overwhelming amount of negation; second, the veridical meaning is doubly implausible; third, the illusion is "inverted" in perception. However, these three "facts" have never been systematically confirmed; and, although they are usually cited as evidence for shallow processing, they are theoretically consistent with multiple possible proposals. In broad strokes, the experiments here confirm the importance of these three properties of illusion sentences; however, we have found that the details are most consistent with the Hypernegation Hypothesis as compared to either of the two possible shallow processing hypotheses.

Throughout the experiments here a particular behavioral response associated with illusion sentences with internal anomalies was consistently elicited. First, if comprehenders experience processing overload due to the multiple negative elements in the sentence, they show no overt signs that they are aware of it. This reaction is a stark and puzzling contrast to the usual perception of complex sentences. Separate pilot experiments, for example, repeatedly found that comprehenders are overwhelmingly likely to reject sentences as nonsensical when they find them computationally intractable: 79% rejected the multiple center embedding

construction *The old woman who the information that the frightened child survived the crash had comforted looked for the rescue worker* while 68% rejected the difficult garden path sentence *The cotton clothing is usually made of grows in Mississippi,* even though participants were provided the option of specifying "don't know" if a sentence was so complex that they could not determine its meaning, and were not only explicitly instructed to differentiate between sentences that were actually nonsensical and those that were very complex by using this response, but also offered financial incentive to do so (in the form of a bonus). Sentences that contain similar numbers of negative elements as the illusion sentence in a slightly different configuration (e.g., *No head injury isn't serious enough to ignore*) typically fall into the same category: they are overwhelmingly rejected due to their incomprehensibility. A computational complexity theory of this phenomenon therefore needs to account for the fact that comprehenders rarely reject illusion sentences due to their complexity: the sentence is either accepted as a normal and sensical English sentence, or it is rejected due to its internal anomaly.

Second, comprehenders appear to be fairly consistent in their proposed interpretation of the internally anomalous illusion sentence: overwhelmingly, they invert its external assertion. In the classic example *no head injury is too trivial to ignore* this inversion yields the pragmatically sensible meaning that head injuries should not be ignored; however, the plausibility of the external assertion turns out to be wholly irrelevant to rates of inversion. The close relationship between internal and external meaning is bidirectional: internally anomalous sentences that are perceived as sensical are usually inverted, while internally consistent sentences almost never are. Elevated levels of inversion also do not generalize to all similar sentence types – outside of the *no-too* conditions, internally anomalous sentences were inverted almost exactly at chance (48%). This indicates that the details of the relationship between the nominal and degree quantifier are somehow essential to this pattern.

It is tempting to attribute the inverted percept wholly to the persistence of the initial parse in the face of mismatching input – a version of the Change Blindness Hypothesis – as such an account would clearly be motivated by other well-known similar phenomena. For example, the Moses illusion is an example of strong expectations – generated by sentential context – seeming to override the actual content of the input. Likewise, it has been widely shown that garden path reanalysis results in "lingering" misinterpretations that are consistent with the initial, supposedly overridden parse (Christianson et al, 2001, 2006). As we have

noted, however, these are fundamentally *comprehension* errors where a semantic analysis is pursed that happens not to be sanctioned by the grammar. Such errors are not generally observed in production: people do not produce garden path sentences like *While Anna dressed the baby played,* intending for the comprehender to infer that *Anna dressed the baby.* And they especially do not actively reject sentences that are already consistent with this parse – such as *While Anna dressed the baby, it played* – by mistakenly insisting that the intended meaning that *Anna dressed the baby* should require a different syntax: *While Anna dressed the baby played.*

   With respect to illusion sentences, people seem to insist on a mapping between syntax and semantics that (at least as far as the semantic literature is concerned) does not exist – in some cases, generating sentences themselves that contain anomalies, and in other cases imposing anomalies onto normal *no-too* sentences, even when there is no particular problem that needs to be "corrected". Facts like these tell us that illusions are not caused by situations where comprehenders are ignoring or failing to notice part of the input, nor by situations where the implausible input is implicitly corrected. Rather, they behave as if they fundamentally misunderstand the logical form of sentences of the form *No X is too Y to Z,* treating the sentence as if it were ambiguous, with two (slightly noisy) form-to-meaning mappings: when the adjective and verb are in a relationship implying *the more ADJ, the more VERB*, similar to a *so…that* or *enough…to* construction, an inverted meaning is obtained; when the adjective and verb imply that *the more ADJ, the less VERB,* the non-inverted meaning is obtained. In cases where people are rejecting internally consistent sentences, it is probably because they independently favor the inverted percept, but the adjective and verb are in the wrong configuration to support that reading.

(82) a. No social program is too wasteful to oppose. [THE MORE ADJ, THE MORE V]

   b. Veridical external percept: Social programs shouldn't be opposed

(83) a. No social program is too efficient to oppose. [THE MORE ADJ, THE LESS V]

   b. Veridical external percept: Social programs should be opposed

The well-cited role of plausibility turns out to indeed affect interpretation, but mostly to the extent that it adjudicates between the two available mappings above. Recall that there was a crucial relationship between plausibility and the detection of the internal anomaly: when the inverted meaning was particularly plausible, the internal anomaly was usually not detected; when the inverted meaning was not very plausible, however, comprehenders had an easier time detecting the internal anomaly. This suggests that the route to a veridical external meaning necessarily involves parsing the adjective and verb in the correct way, *the more ADJ, the less V.* As a consequence, to the extent that the comprehender is determined to retain the veridical interpretation – because, for instance, they find it strongly plausible – they will be forced to reconcile with its problematic internal relationship. When they are not determined to retain this interpretation – for example, they find the inverted interpretation strongly plausible – the meaning can be inverted, but only if the adjective and verb are in the internal configuration *the more ADJ, the more V.* Our results are somewhat in line with Fortuin (2014), who analyzes *No X is too Y to Z* sentences as fundamentally ambiguous, with selection of one meaning over another driven mainly by pragmatic or rhetorical factors. However, although pragmatic factors certainly played some role, the illusion clearly persisted even in cases where the veridical interpretation was more plausible than the inverted one (with accuracy rates around 38% in the *no-too-neg* condition and 55% in the *no-too-pos* condition for such cases, across experiments).

Moreover, contra Fortuin (2014) and other accounts that focus primarily on corpus data rather than psycholinguistic experimentation (e.g., Cook & Stevenson 2010), we found the interpretational patterns to be mostly associated with the logical operators *no* and *too,* rather than the negative verb. Accuracy rates for *no-too* sentences with positive verbs were overall higher than those for negative verbs, but it should be noted that comprehension in this condition was also generally aided by its higher plausibility; accuracy was still as low as 14% on certain items, so interpretation is clearly somewhat unstable regardless of verbal polarity. Importantly, grammatical operations such as negative concord are likely to be sensitive to the distinction between logical and merely morphological negativity, while, for example, from the perspective of computational difficulty, negative verbs would be expected to incur an equal or possibly greater cost (e.g., Sherman 1976), and indeed are often speculated to be critical to the illusion (Wason & Reich 1979; Cook & Stevenson 2010; Fortuin 2014).

One way to explain the interaction between *no* and *too* is to propose a semantics for *too*-expressions that is analogous to *enough* or *so…that* expressions, but with negation inside the clausal complement, so that the logical form of (84)a resembles that of (84)b.

(84)   a. The runner is too fast to catch.

          b. The runner is so fast that one cannot catch her.

An "inverted" interpretation arises when implicit clausal negation is interpreted as semantically inert – as if it were an NPI or an instance of expletive negation, instead of true logical negation – thus altering both the internal and external meaning of the expression, see (85). In other words, the percept of the illusion sentence arises from a grammatical representation that is generally complete, but the negative polarity of the degree quantifier is semantically unstable when embedded underneath a higher negative logical operator (*no*). This account is strongly supported by the finding that the illusion is much easier to detect when these two elements are separated by structural and semantic properties that preclude dependency formation, namely a factive verb and clausal boundary. This suggests that the illusion is treated by the parser as a case of hypernegation similar to others available within or beyond English, although the conflation of two negative elements is apparently only optional in this case – unlike other grammatically-sanctioned hypernegation operations, such as negative concord.

The temptation to invert may be especially strong in the face of various forces of influence: first, verbal polarity probably does not directly affect the availability of the inverted interpretation, but perhaps does contribute to the general difficulty associated with the veridical meaning, in part because of its computational complexity and in part because of its affective negativity. Indeed, the veridical meaning of (85)b is intuitively nearly as difficult to process as that of (85)a:

(85)   a. No head injury is too trivial to ignore.

          b. Veridical meaning: No head injury is so trivial that one should **not** ignore it.

          c. Inverted percept: No head injury is so trivial that one should $\varnothing$ ignore it.

As discussed in Chapter 5, this pattern of interpretation is already associated with other scalar words like *hardly* and *barely,* which similarly have double lives – including an informal use where the implicit negation encoded in the lexical semantics is perceived as semantically inert, and a more standard use where it is not.

(86) a. John drinks **hardly anything** these days. (= John rarely drinks)

b. John **doesn't** drink **hardly anything** these days. (= John rarely drinks)

Indeed, there may a special tendency to interpret logical negation within a comparative clause as semantically inert, given the crosslinguistic tendency for comparative quantifiers to license vacuous paratactic negation in their clausal complements, e.g. (87)-(88).

(87) Maria è più alta di quanto (non) lo sia Giovanni.   (Italian)

Maria is more tall of how-much neg it is Giovanni

'Maria is taller than Giovanni (isn't)'

(88) Marie est plus grande que (ne) l'est Jean.   (French)

Marie is more tall that neg it is Jean

'Marie is taller than Jean (isn't)'

By invoking computational difficulty to explain the role of verbal negation, this account might more aptly be termed the "Accidental Ambiguity Hypothesis"; in other words, although the operation we posit is made available by UG, or even in some dialects of English, it is not clear that it is directly sanctioned within the grammar of standard English. For example, comprehenders do seem to *initially* prefer a veridical interpretation of the sentence onset prior to processing the *to*-clause: Experiment 10 found an association between confidence and plausibility, such that participants felt more confident when completing a *no-too* sentence if they found the *veridical meaning* suggested by the sentence onset to be more plausible. In other words, comprehenders appeared to weight their expectations according to the likelihood of the veridical completion, including the degree to which it was supported by world knowledge. If both interpretations were equally available, then confidence might be expected

231

to be high in cases were *either* available meaning is strongly supported by world knowledge: either the veridical meaning is highly plausible, or the inverted meaning is highly plausible, or there is a large difference between the two values, allowing for confident selection between competing interpretations. However, this was not the case.

To the extent that inversion illusions are driven by misapplied NPI licensing mechanisms, these sentences could form a natural class with a separate grammatical illusion involving the spurious licensing of NPIs by adjacent negative elements, even when those elements are not in the right structural configuration to support NPI licensing (Drenhaus, Saddy & Frisch 2005; Vasisth et al 2008; Xiang, Dillon & Phillips 2009): although the embedded determiner *no* is not in the right position to license the NPI *ever* in (89)b-c, patterns of online processing suggest that the parser initially finds no problem with (89)b, unlike (89)c.

(89)   a. No bills that the senators voted for will ever become law.

        b. * The bills that no senators voted for will ever become law.

        c. * The bills that the senators voted for will ever become law.

These cases seem to suggest similarly that processing mechanisms may at times lead to the acceptance of grammatically unlicensed NPI dependencies. While these two phenomena appear to complement one another – one involving an NPI licensed in the wrong environment, and the other involving a non-NPI treated as an NPI in the right environment – behaviorally the reactions to the two illusion types are quite different. NPI illusions like (89)b are much more temporally unstable, with the illusion rapidly and progressively disappearing with time (Parker & Phillips, to appear), whereas the perception of inversion sentences seems to endure indefinitely absent sufficient conscious introspection. Thus, the superficial similarity between these two phenomena does not necessarily mean that they have the same basic underlying source.

To summarize, the experiments here provide some critical preliminary information about inversion sentences. We have confirmed that many intuitions about the classic illusion sentence turn out to be correct, though in more complicated and subtle ways than the shallow processing accounts would tend to predict. Plausibility and top-down effects unquestionably play some role in the processing of illusion sentences, but not in the way usually assumed;

negation also plays a crucial role in the illusion, but is probably only indirectly associated with its associated computational complexity. In general, the patterns observed in this chapter suggest that inversion sentences are treated as if they were systematically ambiguous. The source of this ambiguity remains an open question: are inversion sentences evidence for a more fundamental diachronic shift within the grammar of English, or do they arise solely due to processing constraints related to the greater ease of the hypernegation reading? Perhaps it is no accident that the logical form of *no-too-neg* sentences falls at the upper bounds of computational tractability; the "accidental ambiguity" proposed here could be thought of as a compositional route that substantially lessens the computational burden associated with the difficult construction, perhaps in the same way that it is (at least intuitively) far easier to automatically obtain a negative concord reading of *John didn't see nobody* than the more effortful double negation reading, in spite of the fact that the latter is the only meaning available within standard English. This would imply, compellingly, that processing constraints may lead to the adoption of certain grammatical operations made available by UG, even when they are not available in the language proper; the effortlessness of this shift may be related precisely to the fact that the grammatical parse does *not* break down in the face of computational duress, contrary to the claims within the shallow processing literature.

7 CONCLUSION

Semantic illusions, when they are addressed in the literature, tend to be construed as evidence for interpretational heuristics that can generate meaning over and above what the grammar provides. It is sometimes thought that these illusions arise in cases where the parser fails to fully implement the grammar-based parse, leading interpretational heuristics alone to fill in the meaning of the sentence. We set out to evaluate these claims, given that they challenge the widespread assumption of incremental compositionality, against a range of competing theories. We found consistently, however, that it is simply not possible to have a complete understanding of comparative illusions without referencing properties intimately related to their grammar. From the perspective of extralinguistic cognition, there is no clear way to explain how or why semantic plurality modulates acceptability in Escher sentences, or why clausal boundaries or factive presuppositions would increase anomaly detection in inversion sentences. These properties are by definition important because of the way they affect the logical semantics. But what does it mean for an illusory percept to be *sensitive* to the grammar, when it is not strictly licensed by the grammar? Proponents of shallow processing accounts are right to point out that these sentences are telling us something important about incremental compositionality, even if the results here do not support the immediate conclusion that compositionality is insufficient as a means of generating meaning in real time. Our results do suggest, however, that there are nontrivial complexities associated with the mapping between form and meaning, ostensibly due to processing considerations related to perceptual uncertainty, speech error detection or general computational limitations.

With respect to Escher sentences, I argued for broad-scale repair of the grammatical anomaly, in light of the computational difficulties encountered at the critical region together with the range of possible percepts associated with these puzzling sentences. This account suggests that the illusory meaning is generated compositionally but that, when composition of the veridical input fails, reinterpretation of the LF is initiated, leading to nonveridical sentence perceptions. Escher sentences are perhaps generated by speech error reversal mechanisms built into the processer, allowing it to navigate noisy, ill-formed or uncertain input, and thus could be incorporated more broadly into recent proposals by Frazier (2014) or Levy et al (2009);

however, the apparent ease of reinterpreting these sentences may also be related to properties of their logical form, which I have argued may present special difficulties for real-time implementation with an incremental left-to-right parser. Crucially, however, the repair mechanisms conform to the semantic link between plurality and cardinality measurement, underscoring the importance of grammatical analysis in accounting for these illusions. In addition, by systematically disentangling morphosyntactic plurality from conceptual or semantic plurality, we were able to localize the relevant operations specifically to interpretational components of the grammar.

Inversion sentences yield a slightly different picture: they seem to be treated as though they were representationally ambiguous between (only) two possible readings, so it does not seem to be the case that a broad repair mechanism is "cleaning up" the problematic internal anomaly (e.g., #*too wasteful to oppose*) in any way that it can. Rather, the illusory percept seems to be tied to the possibility of forming a logical dependency between *no* and *too* that conflates their negative force – an instance of the crosslinguistically common pattern of *duplex negatio negat* (Horn 2009). There is some very preliminary evidence to suggest that the veridical meaning is initially favored, suggesting that the parser may later recruit a grammatical operation whose status in English is questionable, in order to ease the computational burden associated with the many negative elements. The emerging picture of inversion sentences is that there is ambiguity in the bottom-up input – in production and comprehension – and that either a double negation ("veridical") or a hypernegation ("inverted") analysis may be selected on the basis of other available information, including top-down cues such as plausibility. This gives rise to the impression that the illusion is *caused* by pragmatic bias, although in reality, the inverted analysis already seems to be available whether or not this bias exists. Structural priming effects – which arise only for grammatically licensed sentences (Sprouse 2007) – could be used to confirm that the illusory percept indeed arises from an abstract grammatical representation, as suggested here. The ambiguity associated with *no…too* sentences might stem from inherent properties associated with a class of implicitly negative scalar terms like *hardly, barely* and *yet* in English; or, it could be a special type of processing error leading to the adoption of a grammatical operation available within UG, but not in standard English; crosslinguistic research would be helpful in distinguishing among these compelling possibilities.

Since Escher and inversion illusions do not share a specific common source, one might wonder whether they are fundamentally unrelated phenomena. There are several facts that speak against this conclusion. In particular, in both cases, the specific interpretational properties associated with the grammar of comparison seem to figure crucially into the illusion itself, which, broadly speaking, arises as a consequence of the interaction between the extraposed degree clause and the syntactically covert interpretational elements associated with the comparative operator – the negation implicit in *too* and the measure function implicit in *more.* Moreover, we have reason to believe that these silent elements may already yield systematic ambiguity across languages, associated with NPI licensing and event measurement respectively, so that in both cases there could be intermediate stages of processing where the parser is already considering alternatives to the eventual grammar-based analysis. Although issues related to e.g. event ambiguities and NPI licensing are more often discussed with respect to their implications for the grammar, comparative illusions may be telling us that these formal properties of the syntax-semantics interface have important ramifications for real-time online processing; and equally, that the processing patterns observed here can provide important information about the properties of this level of the grammar.

8 BIBLIOGRAPHY

Anderson, C. (2004). *The Structure and Real-Time Comprehension of Quantifier Scope Ambiguity.* Northwestern University.

Arregui, A., Clifton Jr, C., Frazier, L., & Moulton, K. (2006). Processing elided verb phrases with flawed antecedents: The recycling hypothesis. *Journal of Memory and Language*, *55* (2), 232-246.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Using R.* Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. H., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Badecker, W., & Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *28* (4), 748-769.

Bader, M. (1998). Prosodic influences on reading syntactically ambiguous sentences. In *Reanalysis in Sentence Processing* (pp. 1-46). Springer.

Ballatore, A., Wilson, D. C., & Bertolotto, M. (2013). The similarity jury: Combining expert judgements on geographic concepts. In P. V. S. Castano (Ed.), *Advances in Conceptual Modeling* (pp. 231-240). Berlin Heidelberg: Springer.

Balota, D. A. (2007). The English lexicon project. *Behavior Research Methods*, *39* (3), 445-459.

Barker, C. (1992). Group terms in English: representing groups as atoms. *Journal of Semantics*, *9* (1), 69-93.

Barker, C. (1999). Quantification & individuation. *Linguistic Inquiry*, *30*, 683–691.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4.

Beck, S. (2012). DegP scope revisited. *Natural Language Semantics*, *20* (3), 227-272.

Bever, T. G., & Townsend, D. J. (2001). *Sentence comprehension.* Cambridge, MA: MIT Press.

Bever, T. (1970). The cognitive basis for linguistic structures. In *Cognition and the Development of Language.* New York: Wiley & Sons .

Bhatt, R., & Pancheva, R. (2004). Late merger of degree clauses. *Linguistic Inquiry*, *35* (1), 1-45.

Bhatt, R., & Takahashi, S. (2007). Direct comparisons: resurrecting the direct analysis of phrasal comparatives. *Proceedings of SALT*, *17*, pp. 19-36.

Bhatt, R., & Takahashi, S. (2011). Reduced and unreduced phrasal comparatives. *Natural Language & Linguistic Theory, 29*, 581-620.

Bicknell, K., & Levy, R. (2009). A model of local coherence effects in human sentence processing as consequences of updates from bottom-up prior to posterior beliefs. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, (pp. 665–673).

Bierwisch, M. (1989). The semantics of gradation. *Dimensional Adjectives*, *71*, 261.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18* (3), 355-387.

Bock, J. K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*, 45-93.

Bohan, J. (2008). *Depth of Processing and Semantic Anomalies.* University of Glasgow.

Bohan, J., & Sanford, A. (2008). Semantic anomalies at the borderline of consciousness: An eye-tracking investigation. *The Quarterly Journal of Experimental Psychology*, *61* (2), 232-239.

Braze, D., Shankweiler, D., Ni, W., & Palumbo, L. C. (2002). Readers' eye movements distinguish anomalies of form and content. *Journal of Psycholinguistic Research*, *31* (1), 25-44.

Breakstone, M., Cremers, A., Fox, D., & Hackl, M. (2011). On the analysis of scope ambiguities in comparative constructions: converging evidence from real-time sentence processing and offline data. *Proceedings of SALT 21.*

Bredart, S., & Docquier, M. (1989). The Moses illusion: A follow-up on the focalization effect. *Current Psychology of Cognition, 9.*

Bredart, S., & Modolo, K. (1988). Moses strikes again: Focalization effect on a semantic illusion. *Acta Psychologica*, *67* (2), 135-144.

Bresnan, J. (1973). Syntax of the comparative clause construction in English. *Linguistic Inquiry*, *4*, 275-343.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904-911.

Budanitsky, A., & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources*, *2.*

C, C., Perreau-Guimaraes, M., & Suppes, P. (2013). Structural similarities between brain and linguistic data provide evidence of semantic relations in the brain. *PLoS ONE*, *8* (6).

Carlson, K. (2001). *Parallelism and Prosody in the Processing of Ellipsis Sentences.* University of Massachusetts Amherst.

Carlson, R. (1989). Processing nonlinguistic negation. *The American Journal of Psychology*, *102* (2), 211-224.

Carpenter, P. A., Just, M. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1999). Time course of fMRI-activation in language and spatial networks during sentence comprehension. *Neuroimage* .

Carpenter, P., & Just, M. (1975). Sentence comprehension: A psycholinguistic model of verification. *Psychological Review, 82*, 45-76.

Casasanto, L. S., Hoffemeister, P., & Sag, I. (2010). Understanding acceptability judgments: Additivity and working memory effects. *Proceedings of the Cognitive Science Society.*

Cherniak, C. (1986). Limits for knowledge. *Philosophical Studies, 49* (1), 1-18.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge: MIT Press.

Chomsky, N. (1977). On WH-movement. In T. W. Peter Culicover (Ed.), *Formal syntax.* New York: Academic Press.

Chow, W.-Y., & Phillips, C. (2013). No semantic illusions in the "Semantic P600" phenomenon: ERP evidence from Mandarin Chinese. *Brain Research*, *1506*, 76-93.

Christensen, K. R. (2010). Syntactic reconstruction and reanalysis, semantic dead ends, and prefrontal cortex. *Brain and Cognition*, *73* (1), 41-50.

Christianson, K., Hollingworth, A., Halliwell, J., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*, 368-407.

Christie, J., Kozup, J. C., Smith, S., Fisher, D., Burton, S., & Creyer, E. (2001). The effects of bar sponsored alcohol beverage promotions across binge and non-binge drinkers. *Journal of Public Policy and Marketing*, *20*, 240-253.

Clark, H. (1970). How we understand negation. *COBRE Workshop on Cognitive Organization and Psychological Processes.* Huntington Beach, CA.

Clark, H. (1976). *Semantics and Comprehension.* Mouton De Gruyter.

Clark, H. (1971). The chronometric study of meaning components. *CRNS Colloque International sur les Problemes Actuels de Psycholinguistique.* Paris.

Clark, H., & Card, S. (1969). Role of semantics in remembering comparative sentences. *Journal of Experimental Psychology*, *82* (3), 545.

Clark, H., & Chase, W. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472-517.

Clifton, C., & Staub, A. (2008). Parallelism and competition in syntactic ambiguity resolution. *Language and Linguistics Compass*, *2* (2), 234-250.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In *Eye movements: A window on mind and brain* (pp. 341-372).

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In *Eye Movements: A Window on Mind and Brain* (pp. 341-472).

Cook, P., & Stevenson, S. (2010). No sentence is too confusing to ignore. *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, (pp. 61-69).

Corblin, F. (1996). Multiple negation processing in natural language. *Theoria*, *62* (3), 214-259.

Cornish, E., & Wason, P. (1970). The recall of affirmative and negative sentences in an incidental learning task. *The Quarterly Journal of Experimental Psychology*, *20* (2), 109-114.

Corver, N. (1997). Much support as a last resort. *Linguistic Inquiry*, *28*, 119-164.

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes*, *13* (1), 21-58.

Cresswell, M. J. (1976). The semantics of degree. In B. Partee (Ed.), *Montague Grammar* (pp. 261-292). New York: Academic Press.

Dalrymple, M., Shieber, S. M., & Pereira, F. C. (1991). Ellipsis and higher-order unification. *Linguistics and Philosophy*, *14* (4), 388-452.

Davies, M. (2008). The corpus of contemporary American English: 450 million words, 1990-present.

De Cuba, C. (2007). *On (Non)Factivity, Clausal Complementation and the CP-Field.* Stony Brook University .

De Swart, H. (2009). *Expression and Interpretation of Negation: An OT Typology.* Springer.

De Swart, H., & Sag, I. (2002). Negation and negative concord in Romance. *Linguistics and Philosophy*, *25* (4), 373-417.

De Vincenzi, M., Job, R., Di Matteo, R., Angrilli, A., Penolazzi, B., L, C., et al. (2003). Differences in the perception and time course of syntactic and semantic violations. *Brain & Language*, *85* (2), 280-296.

Deglin, V. L., & Kinsbourne, M. (1996). Divergent thinking styles of the hemispheres: How syllogisms are solved during transitory hemisphere suppression. *Brain and Cognition*, *31* (3), 285-307.

Den Besten, J. B. (1989). *Studies in West Germanic Syntax.* University of Tilburg.

Déprez, V. (1997). A non-unified analysis of negative concord. In D. F. al. (Ed.), *Negation and Polarity: Syntax and Semantics* (pp. 53-75). Amsterdam: John Benjamins Publishing.

Ditman, T., Holcomb, P. J., & Kuperberg, G. R. (2007). An investigation of concurrent ERP and self-paced reading methodologies. *Psychophysiology*, *44* (6), 927-935.

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

Drenhaus, H., Saddy, D., & Frisch, S. (2005). Processing negative polarity items: When negation comes through the backdoor. In S. K. Reis (Ed.), *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives.* Berlin: De Gruyter.

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, *27* (4), 429-446.

Endriss, C., & Hinterwimmer, S. (2007). Selective vs. unselective quantification over the atomic parts of plural entities: A comparison of for the most part and usually. *Plurality, Unity and Structure in Semantics*.

Enochson, K., & Culbertson, J. (2015). Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PLoS ONE*, *10* (3).

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20* (5), 540-551.

Evans, J. S. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255-278.

Evans, J. S., & Over, D. E. (1996). *Rationality and Reasoning.* Psychology Press.

Fedorenko, E., Gibson, E., & Rohde, D. (2007). The nature of working memory in linguistic, arithmetic and spatial integration processes. *Journal of Memory and Language*, *56* (2), 246–269.

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

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47* (2), 164-203.

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1* (1-2), 71-83.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11* (1), 11-15.

Fillenbaum, S. (1974). Pragmatic normalization: Further results for some conjunctive and disjunctive sentences . *Journal of Experimental Psychology*, *102* (4), 574-578.

Fillenbaum, S. (1971). Processing and recall of compatible and incompatible question and answer pairs. *Language and Speech*, *14* (3), 256-265.

Fine, A. B., Florian, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, *8* (10).

Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry Jr., N. W. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology*, *20*, 400-409.

Frazier, L. (2014). Two interpretive systems for natural language? *Journal of Psycholinguistic Research*, *44* (1).

Frazier, L., & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, *29* (2), 181-200.

Garnham, A., & Oakhill, J. (1987). Interpreting elliptical verb phrases. *The Quarterly Journal of Experimental Psychology*, *39* (4), 611-627.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37* (1), 58-93.

Garrod, S., & Sanford, A. (1998). Incrementality in discourse understanding. In H. v. Goldman (Ed.), *The Construction of Mental Representations During Reading.* Mahwah: Lawrence Erlbaum Associates.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical models.* Cambridge University Press.

Geurts, B., & Van Der Slik, F. (2005). Monotonicity and processing load. *Journal of Semantics*, *22* (1), 97-117.

Giannakidou, A. (2000). Negative...Concord? *Natural Language and Linguistic Theory*, *18*, 457-523.

Giannakidou, A. (2006). Only, emotive factive verbs, and the dual nature of polarity dependency. *Language*, *82* (3).

Giannakidou, A. (1998). *Polarity Sensitivity as (Non)Veridical Dependency* (Vol. 23). John Benjamins Publishing Company.

Giannakidou, A. (1997). *The Landscape of Polarity Items.* University of Groningen.

Gibson, E. (2006). The interaction of top-down and bottom-up statistics in the resolution of syntactic category ambiguity. *Journal of Memory and Language*, *54*, 363–388.

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass*, *5*, 509–524.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". In W. S. Hewstone (Ed.), *European Review of Social Psychology* (Vol. 2).

Gigerenzer, G., & Regier, T. (1996). How do we tell an association from a rule? Comment on Sloman (1996). *Psychological Bulletin, 119* (1), 23-26.

Giora, R. (2006). Anything negatives can do affirmatives can do just as well, except for some metaphors. *Journal of Pragmatics*, *38* (7), 981-1014.

Givon, T. (1978). Negation in language: pragmatics, function, ontology. In P. Cole (Ed.), *Pragmatics.* New York: Academic Press.

Glass, A. L., Holyoak, K. J., & O'Dell, C. (1974). Production frequency and the verification of quantified statements. *Journal of Verbal Learning & Verbal Behavior*, *13*, 237-254.

Goel, V., Buchel, C., Frith, C., & Dolan, R. J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *Neuroimage*, *12*, 504-514.

Golding, E. (1981). The effect of unilateral brain lesion on reasoning. *Cortex*, *17*, 31-40.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes*, *25* (2), 149-188.

Grant, M. (2013). *The Parsing and Interpretation of Comparatives: More than Meets the Eye.* University of Massachusetts Amherst .

Greenberg, Y. (2009). Additivity in the domain of eventualities. *Proceedings of Sinn und Bedeutung.*

Guerzoni, E. (2006). Intervention effects on NPIs and feature movement: towards a unified account of intervention. *Natural Language Semantics*, *14* (4), 359-398.

Hackl, M. (2000). *Comparative quantifiers.* Massachusetts Institute of Technology.

Hackl, M. (2001). Comparative quantifiers and plural predication. *Proceedings of WCCFL XX*, (pp. 234-247).

Hackl, M., Koster-Hale, J., & Varvoutis, J. (2013). Quantification and ACD: Evidence from Real-Time Sentence Processing. *Journal of Semantics* .

Hacquard, V. (2005). Aspects of "too" and "enough" constructions. *Proceedings of SALT 15.*

Haegeman, L. (1995). *The Syntax of Negation.* Cambridge: Cambridge University Press.

Hahne, A., & Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. *Journal of Cognitive Neuroscience*, *11* (2), 194-205.

Hankamer, J. (1973). Why there are two than's in English. *Papers from the 9th Regional Meeting of the Chicago Linguistics Society*, (pp. 179-191).

Hannon, B., & Daneman, M. (2001). Susceptibility to semantic illusions: An individual-differences perspective. *Memory & Cognition*, *29* (3), 449-461.

Hasson, U., & Glucksberg, S. (2006). Does understanding negation entail affirmation?: An examination of negated metaphors. *Journal of Pragmatics*, *38* (7), 1015-1032.

Heim, I. (2000). Degree operators and scope. *Proceedings of SALT*, *10*, pp. 40-64.

Heim, I. (1985). *Notes on comparatives and related matters.* University of Texas at Austin.

Hellan, L. (1981). *Towards an integrated analysis of comparatives* (Vol. 11). Narr.

Henderson, J. M., & Hollingworth, A. (2003). Eye movements and visual memory: Detecting changes to saccade targets in scenes. *Perception & Psychophysics*, *65* (1), 58-71.

Herburger, E. (2001). The negative concord puzzle revisited. *Natural Language Semantics*, *9* (3), 289-333.

Hermann, D. J., Conti, G., Peters, D., Robbins, P. H., & Chaffin, R. J. (1979). Comprehension of antonymy and the generality of categorization models. *Journal of Experimental Psychology: Human Learning and Memory*, *5* (6), 585.

Hill, H., & Johnston, A. (2007). The hollow-face illusion: Object-specific knowledge, general assumptions or properties of the stimulus? *Perception*, *36*.

Hobbs, J. R., & Shieber, S. M. (1987). An algorithm for generating quantifier scopings. *Computational Linguistics*, *13* (1-2), 47-63.

Homer, V. (2008). Disruption of NPI licensing: the case of presuppositions. *Proceedings of SALT XVIII.*

Horn, L. (1989). *A Natural History of Negation* (Vol. 960). Chicago: University of Chicago Press.

Horn, L. (2005). Airport '86 Revisited: toward a unified indefinite any. In G. C. Pelletier, *The Partee Effect.*

Horn, L. (2009). Hypernegation, hyponegation, and parole violations. *Berkeley Linguistics Society.*

Jacobson, P., & Gibson, E. (2014). Processing of ACD gives no evidence for QR. *Proceedings of SALT 24.*

Jespersen, O. (1917). Negation in English and other languages. In *Selected Writings of Otto Jespersen* (pp. 3-151). London: George Allen & Unwin.

Jiang, J. J., & Conrath, W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of International Conference on Research in Computational Linguistics.*

Jung Grant, S., Malaviya, P., & Sternthal, B. (2004). The influence of negation on product evaluations. *Journal of Consumer Research*, *31*, 583-591.

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior, 10* (3), 244-253.

Just, M. A., & Clark, H. H. (1973). Drawing inferences from the presuppositions and implications of affirmative and negative sentences. *Journal of Verbal Learning and Verbal Behavior*, *12* (1), 21-31.

Kahneman, D. (2011). *Thinking, Fast and Slow.* Macmillan.

Karttunen, L. (1971). Implicative verbs. *Language* (47), 340-358.

Kaschak, M., & Glenberg, A. M. (2004). This construction needs learned. *Journal of Experimental Psychology: General, 133* (3), 450.

Kaup, B. A., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is that door that is not open mentally closed? *Journal of Pragmatics*, *38* (7), 1033-1050.

Kaup, B., & Zwaan, R. (2003). Effects of negation and situational presence on the accessibility of text information. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29*, 439-446.

Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods*, *41* (1), 1-12.

Kennedy, C. (2001). Polar opposition and the ontology of 'degrees'. *Linguistics and Philosophy*, *24* (1), 33-70.

Kennedy, C. (1999). *Projecting the Adjective: the syntax and semantics of gradability and comparison.* New York: Garland Press.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*, 1-45.

Kiparsky, P., & Kiparsky, C. (1971). Fact. In D. D. Jakobovits (Ed.), *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology* (pp. 345-369). Cambridge University Press.

Klein, E. (1991). Comparatives. In S. E. Forschung, & A. v. Wunderlich (Ed.). Berlin: Walter de Gruyter.

Klima, E. (1964). Negation in English. *The structure of language* . (J. F. Katz, Ed.) Englewood Cliffs, New Jersey: Prentice-Hall.

Kranz, D., Luce, S., & Tversky, A. (1971). *Foundations of Measurement* (Vols. 1, 3). New York & London: Academic Press.

Kratzer, A. (2005). On the plurality of verbs. In J. D. Heyde-Zybatow (Ed.), *Event Structures in Linguistic Form and Interpretation.* Berlin: Mouton de Gruyter.

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

Kucera, H., & Francis, W. N. (1967). *Computational Analysis of Present-day American English.* Providence: Brown Unviersity press.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23-49.

Löbner, S. (1990). *Wahr neben Falsch: Duale Operatoren als die Quantoren natürlicher Sprache.* Tübingen: Niemeyer.

Lüdtke, J., & Kaup, B. (2006). Context effects when reading negative and affirmative sentences. In R. Sun (Ed.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, (pp. 1735-1740).

Lüdtke, J., Friedrich, K. C., De Filippis, M., & Kaup, B. (2008). Event-related potential correlates of negation in a sentence-picture verification paradigm. *Journal of Cognitive Neuroscience*, *20* (8), 1355-1370.

Labov, W. (1972). Negative Attraction and Negative Concord in English grammar. *Language*, *48*, 773-818.

Ladusaw, W. (1992). Expressing negation. In C. B. Dowty (Ed.), *Proceedings of SALT II*, (pp. 237-259).

Ladusaw, W. (1980). *Polarity Sensitivity as Inherent Scope Relations.* New York: Garland.

Laka, I. (1990). *Negation in syntax: On the nature of functional categories and projections.* PhD dissertation, MIT.

Lechner, W. (2004). *Ellipsis in Comparatives.* Berlin: Mouton de Gruyter.

Lechner, W. (2001). Reduced and phrasal comparatives. *Natural Language & Linguistic Theory*, *19* (4), 683-735.

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, (pp. 1055-1065).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106(3): 1126-1177.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106* (50), 21086-21090.

Liberman, M. (2004, May 7). *Escher sentences*. Retrieved from Language Log: http://itre.cis.upenn.edu/~myl/languagelog/archives/000862.html

Liberman, M. (2009, November 28). *No wug is too dax to be zonged*. Retrieved from Language Log: http://languagelog.ldc.upenn.edu/nll/?p=1926

Liberman, M. (2006, February 26). *Why are negations so easy to fail to miss?* Retrieved from Language Log: http://languagelog.ldc.upenn.edu/nll/?p=1926

Lin, D. (1998). An information-theoretic definition of similarity. *Proceedings of the 15th International Conference on Machine Learning.*

Linebarger, M. (1987). Negative polarity and grammatical representation. *Linguistics and Philosophy, 10*, 325–387.

Lombardi, L., & Potter, M. C. (1992). The regeneration of syntax in short term memory. *Journal of Memory and Language*, *31*, 713-733.

MacDonald, M. C., & Just, M. A. (1989). Changes in activation levels with negation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *15*, 633-642.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101* (4), 676.

McElree, B., Traxler, M. J., Pickering, M. J., Seely, R. E., & Jackendoff, R. (2001). Reading time evidence for enriched composition. *Cognition*, *78* (1), B17-B25.

Meier, C. (2003). The meaning of too, enough and so...that. *Natural Language Semantics*, *11*, 69-107.

Meng, L., Huang, R., & Gu, J. (2008). A review of semantic similarity measures in Wordnet. *International Journal of Hybrid Information Technology*, *6* (1), 1-12.

Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. T. (1993). A Semantic Concordance. *Proceedings of the ARPA Workshop on Human Language Technology*.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *The Psychological Review*, *63*, 81-97.

Montalbetti, M. M. (1984). *After Binding: On the Interpretation of Pronouns*. Massachusetts Institute of Technology.

Moore, Steven. (1994). Events, objects and eventual objects. *Proceedings of WCCFL 12*.

Natsopoulos, D. (1985). A verbal illusion in two languages. *Journal of Psycholinguistic Research*, *14* (4), 385-397.

Ni, W., Fodor, J. D., Crain, S., & Shankweiler, D. (1998). Anomaly detection: Eye movement patterns. *Journal of Psycholinguistic Research*, *27* (5), 515-539.

Nieuwland, M., & Kuperberg, G. (2008). When the truth Is not too hard to handle. *Psychological Science*, *19* (12), 1213-1218.

Nissenbaum, J., & Schwarz, B. (2010). Parasitic degree phrases. *Natural Language Semantics* .

Noë, A. (2002). Is the visual world a grand illusion? *Journal of Consciousness Studies*, *9* (5-6), 1-12.

Noë, A., Pessoa, L., & Thompson, E. (2000). Beyond the grand illusion: What change blindness really teaches us about vision. *Visual Cognition*, *7* (1-3), 93-106.

Oostendorp, H., & De Mul, S. (1990). Moses beats Adam: a semantic relatedness effect on a semantic illusion. *Acta Psychologica, 74* (1), 35-46.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, *11* (6), 988-1010.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31* (6), 785-806.

Pancheva, R. (2009). More students attended FASL than CONSOLE. *Formal Approaches to Slavic Linguistics 18*. Cornell University.

Parker, D., & Phillips, C. (to appear). Negative polarity illusions and the format of hierarchical encodings in memory.

Partee, B. (2004). The Airport Squib: any, almost and superlatives. In *Compositionality in Formal Semantics*. Malden, MA: Blackwell Publishing Ltd.

Paul, H. (1886). *Principien der Sprachgeschichte*. Halle: Max Niemeyer .

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, *41*, 427-456.

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, *41* (3), 427-456.

Pearson, H. (2011). A new semantics for group nouns. *Proceedings of WCCFL*, *28*.

Pederson, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet:: Similarity: measuring the relatedness of concepts. *Proceedings of HLT-NAACL*, (pp. 38-41).

Penka, D. (2006). *Negative Indefinites*. University of Tübingen.

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

Peter, W., & Reich, S. (1979). A verbal illusion. *Quarterly Journal of Experimental Psychology*, *31* (4), 591-597.

Phillips, C. (1996). *Order and structure*. Massachusetts Institute of Technology.

Phillips, C. (1996). *Order and Structure.* Massachusetts Institute of Technology.

Phillips, C., Wagers, M., & Lau, E. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. In J. Runner (Ed.), *Experiments at the Interfaces* (pp. 153-186). Bingley, UK: Emerald Publications.

Pickering, M. J., & Branigan, H. P. (1999). Syntactic priming in language production. *Trends in Cognitive Sciences*, *3* (4), 136-141.

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24* (4), 940.

Pickering, M. J., McElree, B., & Traxler, M. J. (2005). The difficulty of coercion: A response to de Almeida. *Brain and Language*, *93* (1), 1-9.

Pickering, M. J., McElree, B., Frisson, S., Chen, L., & Traxler, M. J. (2006). Underspecification and aspectual coercion. *Discourse Processes*, *42* (2), 131-155.

Pickering, M., & Garrod, S. (2013). How tightly are production and comprehension interwoven? *Frontiers in Psychology*, *4*, 238.

Poesio, M. Relational semantics and scope ambiguity. In J. G. J. Barwise (Ed.), *Situation Theory and its Applications* (Vol. 2, pp. 469-497).

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, *29*, 633-654.

Pritchett, B. L. (1998). Garden path phenomena and the grammatical basis of language processing. *Language*, *64* (3), 539-576.

Progovac, L. (1988). *A Binding Approach to Polarity Sensitivity.* University of Southern California.

Pulvermüller, F., Shtyrov, Y., Hasting, A. S., & Carlyon, R. P. (2008). Syntax as a reflex: Neurophysiological evidence for early automaticity of grammatical processing. *Brain and Language*, *104* (3), 244-253.

Pylkkänen, L., & McElree, B. (2007). An MEG study of silent meaning. *Journal of Cognitive Neuroscience*, *19* (11), 1905-1921.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English Language.* London & New York: Longman.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14* (3), 191-201.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence.*

Rullman, H. (1995). *Maximality in the semantics of wh-constructions.* University of Massachusetts Amherst.

Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2006). Processing Reflexives and Pronouns in Picture Noun Phrases. *Cognitive Science*, 193-241.

Rychlak, J. F., Barnard, S., Williams, R. N., & Wollman, N. (1989). The recognition and cognitive utilization of oppositionality. *Journal of Psycholinguistic Research*, *18* (2), 181-199.

Sag, I. (1976). A note on verb phrase deletion. *Linguistic Inquiry*, *7* (4), 664-671.

Sag, I. (1976). *Deletion and logical form.* Massachusetts Institute of Technology.

Sanford, A. (2002). Context, Attention and Depth of Processing During Interpretation. *Mind & Language*, *17* (1-2), 188-206.

Sanford, A. J., & Emmott, C. (2012). *Mind, Brain and Narrative.* Cambridge University Press.

Sanford, A. J., & Graesser, A. C. (2006). Shallow Processing and Underspecification. *Discourse Processes*, *42* (2), 99-108.

Sanford, A. J., & Moxey, L. M. (2004). Exploring quantifiers: Pragmatics meets the psychology of comprehension. In I. A. Sperber (Ed.), *Experimental Pragmatics.* Basingstoke, Hampshire: Palgrave Macmillan.

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

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon mechanical turk for linguistic research. *Psihologija*, *43*, 441–464.

Schwarzschild, R. (2008). The semantics of comparatives and other degree constructions. *Language & Linguistics Compass*, *2* (2), 308-331.

Schwarzschild, R., & Wilkinson, K. (2002). Quantifiers in comparatives: A semantics of degree based on intervals. *Natural Language Semantics*, *10* (1), 1-41.

Selkirk, E. (1970). On the determiner systems of noun phrases and adjective phrases. Ms, MIT.

Seuren, P. (1973). The comparative. In F. K. Ruwet (Ed.), *Generative Grammar in Europe* (pp. 528-564). Dordrecht: Reidel.

Seuren, P. (1984). The comparative revisited. *Journal of Semantics*, *3*, 109-141.

Seuren, P. (1978). The structure and selection of positive and negative gradable adjectives. In W. J. D. Farkas (Ed.), *Papers from the Parasession on the Lexicon, CLS 14*, (pp. 336-346).

Sherman, M. A. (1976). Adjectival negation and the comprehension of multiply negated sentences. *Journal of Verbal Learning and Verbal Behavior*, *15* (2), 143-157.

Simons, D. J., & Levin, D. T. (1997). Change Blindness. *Trends in Cognitive Sciences*, *1* (7), 261-267.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, *5* (4), 644-649.

Skurnik, I. W., Yoon, C., Park, D., & Schwarz, N. (2005). How warnings about false claims become recommendations: Paradoxical effects of warnings on beliefs of older consumers. *Journal of Consumer Research*, *31*, 713-724.

Slattery, T. J., Sturt, P., Christianson, K., Yoshida, M., & Ferreira, F. (2013). Lingering misinterpretations of garden path sentences arise from competing syntactic representations. *Journal of Memory and Language*, *69* (2), 104-120.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119* (1), 3.

Solomon, E., & Pearlmutter, N. (2004). Semantic integration and syntactic planning in language production. *Cognitive Psychology*, *49* (1), 1-46.

Sprouse, J. (2007). *A program for experimental syntax.* University of Maryland, College Park.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43* (1), 155-167.

Staab, J. (2007). *Negation in Context: electrophysiological and behavioral investigations of negation effects in discourse processing.* UC San Diego.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23* (5), 645-665.

Stein, E. (1996). *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science.* Oxford University Press.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276-315.

Strawson, P. (1952). *Introduction to Logical Theory.* London: Methuen.

Sturt, P. (2007). Semantic re-interpretation and garden path recovery. *Cognition*, *105* (2), 477-488.

Sturt, P., Pickering, M. J., & Crocker, M. W. (1998). Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, *40* (1), 136-150.

Sturt, P., Sanford, A. J., Stewart, A., & Dawydiak, E. (2004). Linguistic focus and Good-Enough representations: an application of the changedetection paradigm. *Psychonomic Bulletin & Review*, *11* (5), 882-888.

Szabolcsi, A. (2014). Quantification and ACD: What is evidence from real-time processing evidence for? A response to Hackl et al. (2012). *Journal of Semantics*, *31*, 135-145.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*, 355–370.

Tettamanti, M., Manenti, R., Della Rosa, P., Falini, A., Perani, D., & Moro, A. (2008). Negation in the brain: Modulating action representations. *NeuroImage*, 358-367.

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

Trabasso, T., Rollins, H., & Shaughnessy, E. (1971). Storage and verification stages in processing concepts. *Cognitive Psychology*, *2* (3), 239-289.

Traxler, M. J., McElree, B., Williams, R. S., & Pickering, M. J. (2005). Context effects in coercion: Evidence from eye movements. *Journal of Memory and Language*, *53* (1), 1-25.

Traxler, M. J., Pickering, M. J., & McElree, B. (2002). Coercion in sentence processing: Evidence from eye-movements and self-paced reading. *Journal of Memory and Language*, *47* (4), 530-547.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19* (3), 28.

Tunstall, S. L. (1998). *The Interpretation of Quantifiers: Semantics & Processing.* University of Massachusetts, Amherst.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185* (4157), 1124-1131.

Van der Wouden, T. (1994). *Negative Contexts.* University of Groningen.

Van Gompel, R. P., Pickering, M. J., Pearson, J., & Jacob, G. (2006). The activation of inappropriate analyses in garden-path sentences: Evidence from structural priming. *Journal of Memory and Language*, *55* (3), 335=362.

Van Oostendorp, H., & Kok, I. (1990). Failing to notice errors in sentences. *Languages and Cognitive Processes,*, *5*, 105-113.

Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: how the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*, 685–712.

Vigliocco, G., & Nicol, J. (1998). Separating hierarchical relations and word order in language production: is proximity concord syntactic or linear? *Cognition*, *68* (1), B13-B29.

Von Helmholz, H. (1866). *Treatise on physiological optics* (Vol. 3).

Von Stechow, A. (1984a) Comparing semantic theories of comparison. *Journal of Semantics*, *3*, 1-77.

Von Stechow, A. (1984b) My reaction to Cresswell's, Hellan's, Hoeksema's and Seuren's Comments. *Journal of Semantics*, *3*, 183-199

Von Stechow, A., Krasikova, S., & Penka, D. (2004). The meaning of German um zu: necessary condition and enough/too. *Modal Verbs and Modality*. Universitat Tubingen.

Wagers, M., Lau, E., & Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *Journal of Memory and Language*, *61* (2), 206-237.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*, 392-393.

Warren, T., Vasisth, S., Hirotani, M., & Drenhaus, H. (2006). Licensor strength and locality effects in negative polarity licensing. *CUNY Human Sentence Processing Conference.*

Wason, P. (1961). Responses to affirmative and negative binary statements. *British Journal of Psychology*, *52*, 133-142.

Wason, P. (1965). The contexts of plausible denial. *Journal of Verbal Learning & Verbal Behavior*, *4* (1), 7-11.

Wason, P. (1981). Understanding and the limits of formal thinking. In H. P. Bouveresse (Ed.), *Meaning and Understanding.* Walter de Gruyter.

Wason, P., & Jones, S. (1963). Negatives: Denotation and connotation. *British Journal of Psychology*, *54* (4), 299-307.

Wason, P., & Reich, S. (1979). A verbal illusion. *Quarterly Journal of Experimental Psychology*, *31* (4), 591-597.

Wegner, D. M., Wenzlaff, R., Kerker, R. M., & & Beattie, A. E. (1981). Incrimination through innuendo: Can media questions become public answers? *Journal of Personality and Social Psychology*, *40*, 822– 832.

Wei, M. (1993). An analysis of word relatedness correlation measures. University of Western Ontario.

Wellwood, A. (2014). Measuring Predicates. PhD dissertation, University of Maryland, College Park.

Wellwood, A., Hacquard, V., & Pancheva, R. (2012). Measuring and comparing individuals and events. *Journal of Semantics*, *29* (2), 207-228.

Wellwood, A., Pancheva, R., Hacquard, V., & Phillips, C. (2012b). Deconstructing a comparative illusion. University of Maryland & University of Southern California.

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

Wharton, C. M., & Grafman, J. (1998). Deductive reasoning and the brain. *Trends in Cognitive Sciences*, *2* (2), 54-59.

Williams, E. (1977). Discourse and logical form. *Linguistic Inquiry*, 101-139.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of HLT-EMNLP.*

Wold, D. (1995). Antecedent-Contained Deletion in Comparative Constructions. Ms, MIT.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. *Proceedings of 32nd annual Meeting of the Association for Computational Linguistics.*

Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain & Language*, *108* (1), 40-55.

Zanuttini, R. (1991). *Syntactic Properties of Sentential Negation: A Comparative Study of Romance Languages.* University of Pennsylvania.

Zeijlstra, H. (2004). *Sentential Negation and Negative Concord.* University of Amsterdam.

Zwarts, F. (1995). Noveridical contexts. *Linguistic Analysis*, *25* (3-4), 286-312.

Zwicky, A. (2008, July 20). *The Astonishment Effect in negation*. Retrieved from http://languagelog.ldc.upenn.edu/nll/?p=382

9 APPENDIX A: EXPERIMENTAL STIMULI

9.1 Experiment 1

Conditions:

       a. Nonrepeatable, control,
       b. Nonrepeatable, illusion,
       c. Repeatable, control,
       d. Repeatable, illusion

(1)   a. More relatives went to my 21st birthday party than friends did because my parents invited everyone that they knew.
       b. More relatives went to my 21st birthday party than my friend did because my parents invited everyone that they knew.
       c. More relatives went to my childhood birthday parties than friends did because my parents invited everyone that they knew.
       d. More relatives went to my childhood birthday parties than my friend did because my parents invited everyone that they knew.

(2)   a. More quarterbacks were selected in the 2011 NFL draft than wide receivers were even though the quarterbacks didn't play as well.
       b. More quarterbacks were selected in the 2011 NFL draft than our wide receiver was even though the quarterbacks didn't play as well.
       c. More quarterbacks were interviewed after the 2011 NFL draft than wide receivers were even though the quarterbacks didn't play as well.
       d. More quarterbacks were interviewed after the 2011 NFL draft than our wide receiver was even though the quarterbacks didn't play as well.

(3)   a. This Friday, more convenience store clerks claimed jackpot lottery winnings than customers did because clerks got the tickets at discounted prices.
       b. This Friday, more convenience store clerks claimed jackpot lottery winnings than my customer did because clerks got the tickets at discounted prices.
       c. This year, more convenience store clerks bought jackpot lottery tickets than customers did because clerks got the tickets at discounted prices.
       d. This year, more convenience store clerks bought jackpot lottery tickets than my customer did because clerks got the tickets at discounted prices.

(4)   a. Last fall, more engineers relocated to San Francisco than accountants did in order to work in our corporate offices.
       b. Last fall, more engineers relocated to San Francisco than our accountant did in order to work in our corporate offices.

c. Last year, more engineers traveled to San Francisco than accountants did in order to work in our corporate offices.

d. Last year, more engineers traveled to San Francisco than our accountant did in order to work in our corporate offices.

(5)   a. Last spring more ducks hatched from eggs than chickens did due to the declining population of local predators.

b. Last spring more ducks hatched from eggs than my chicken did due to the declining population of local predators.

c. Last spring more ducks laid eggs here than chickens did due to the declining population of local predators.

d. Last spring more ducks laid eggs here than my chicken did due to the declining population of local predators.

(6)   a. More financial analysts graduated from Harvard's business school than managers did even though our company is close to Boston.

b. More financial analysts graduated from Harvard's business school than our manager did even though our company is close to Boston.

c. More financial analysts hired from Harvard's business school than managers did even though our company is close to Boston.

d. More financial analysts hired from Harvard's business school than our manager did even though our company is close to Boston.

(7)   a. More students at my school joined Facebook than teachers did when the site was first launched in 2004.

b. More students at my school joined Facebook than my teacher did when the site was first launched in 2004.

c. More students at my school used Facebook than teachers did when the site was first launched in 2004.

d. More students at my school used Facebook than my teacher did when the site was first launched in 2004.

(8)   a. More senior citizens had their appendix removed than teenagers did because doctors take extra precautions with the elderly.

b. More senior citizens had their appendix removed than our teenager did because doctors take extra precautions with the elderly.

c. More senior citizens had their moles removed than teenagers did because doctors take extra precautions with the elderly.

d. More senior citizens had their moles removed than our teenager did because doctors take extra precautions with the elderly.

(9)   a. More businesses were demolished for the new rail system than houses were but the city compensated residents for all disturbances.

b. More businesses were demolished for the new rail system than my house was but the city compensated residents for all disturbances.

c. More businesses were affected by the new rail system than houses were but the city compensated residents for all disturbances.

d. More businesses were affected by the new rail system than my house was but the city compensated residents for all disturbances.

(10)   a. More businesses in our town burned down in the fire than schools did and the local residents all bought homeowners insurance.
b. More businesses in our town burned down in the fire than our school did and the local residents all bought homeowners insurance.
c. More businesses in our town held drills after the fire than schools did and the local residents all bought homeowners insurance.
d. More businesses in our town held drills after the fire than our school did and the local residents all bought homeowners insurance.

(11)   a. More strawberry plants were ruined during the drought than geraniums were because geraniums generally need much less water.
b. More strawberry plants were ruined during the drought than my geranium was because geraniums generally need much less water.
c. More strawberry plants were watered during the drought than geraniums were because geraniums generally need much less water.
d. More strawberry plants were watered during the drought than my geranium was because geraniums generally need much less water.

(12)   a. This year more managers were fired from the company for poor performance than assistants were because the company was prioritizing team leadership skills.
b. This year more managers were fired from the company for poor performance than our assistant was because the company was prioritizing team leadership skills.
c. This year more managers were rewarded by the company for good performance than assistants were because the company was prioritizing team leadership skills.
d. This year more managers were rewarded by the company for good performance than our assistant was because the company was prioritizing team leadership skills.

(13)   a. More football players were formally inducted into the Hall of Fame this year than baseball players were despite the initial requests made by the owners.
b. More football players were formally inducted into the Hall of Fame this year than the baseball player was despite the initial requests made by the owners.
c. More football players were invited to visit the Hall of Fame this year than baseball players were despite the initial requests made by the owners.
d. More football players were invited to visit the Hall of Fame this year than the baseball player was despite the initial requests made by the owners.

(14)   a. Last season, more American tennis players won their final match than Canadian players did and this season the American players are undefeated.
b. Last season, more American tennis players won their final match than the Canadian player did and this season the American players are undefeated.
c. Last season, more American tennis players won their early matches than Canadian players did and this season the American players are undefeated.
d. Last season, more American tennis players won their early matches than the Canadian player did and this season the American players are undefeated.

(15)   a. After the dot-com bubble, more tech companies filed for bankruptcy than
         pharmaceutical companies did even though both industries were projected to thrive.
         b. After the dot-com bubble, more tech companies filed for bankruptcy than the
         pharmaceutical company did even though both industries were projected to thrive.
         c. During the dot-com bubble, more tech companies filed for patents than
         pharmaceutical companies did even though both industries were projected to thrive.
         d. During the dot-com bubble, more tech companies filed for patents than the
         pharmaceutical company did even though both industries were projected to thrive.

(16)   a. More married couples bought their first house in the suburbs than bachelors did since
         the school districts are so much better.
         b. More married couples bought their first house in the suburbs than the bachelor did
         since the school districts are so much better.
         c. More married couples looked for new houses in the suburbs than bachelors did since
         the school districts are so much better.
         d. More married couples looked for new houses in the suburbs than the bachelor did
         since the school districts are so much better.

(17)   a. In the election, more Republicans in the Texas legislature voted against Proposition
         17 than Democrats did and no one voted for environmental reform laws.
         b. In the election, more Republicans in the Texas legislature voted against Proposition
         17 than the Democrat did and no one voted for environmental reform laws.
         c. In the nineties, more Republicans in the Texas legislature voted against tax cuts than
         Democrats did and no one voted for environmental reform laws.
         d. In the nineties, more Republicans in the Texas legislature voted against tax cuts than
         the Democrat did and no one voted for environmental reform laws.

(18)   a. More actors were cast in the director's new movie than actresses were since the actors
         would work for less money.
         b. More actors were cast in the director's new movie than the actress was since the
         actors would work for less money.
         c. More actors were cast in the studio's new movies than actresses were since the actors
         would work for less money.
         d. More actors were cast in the studio's new movies than the actress was since the
         actors would work for less money.

(19)   a. More pop stars got their bellybuttons pierced before the photo shoot than
         supermodels did even though the talent agency advised against it.
         b. More pop stars got their bellybuttons pierced before the photo shoot than the
         supermodel did even though the talent agency advised against it.
         c. More pop stars got their nails painted before their photo shoots than supermodels did
         even though the talent agency advised against it.
         d. More pop stars got their nails painted before their photo shoots than the supermodel
         did even though the talent agency advised against it.

(20)   a. Last night more passengers in coach were bumped from the flight to Chicago than
         businessmen were but the airline got all passengers there eventually.

b. Last night more passengers in coach were bumped from the flight to Chicago than the businessman was but the airline got all passengers there eventually.
c. Last year more passengers in coach were bumped from the flights to Chicago than businessmen were but the airline got all passengers there eventually.
d. Last year more passengers in coach were bumped from the flights to Chicago than the businessman was but the airline got all passengers there eventually.

(21)   a. More guitar players went deaf from performing live shows than pop stars did since pop stars primarily work in the studio.
b. More guitar players went deaf from performing live shows than the pop star did since pop stars primarily work in the studio.
c. More guitar players sold merchandise by performing live shows than pop stars did since pop stars primarily work in the studio.
d. More guitar players sold merchandise by performing live shows than the pop star did since pop stars primarily work in the studio.

(22)   a. More botanists discovered the new species in the Amazon than entomologists did but they jointly published the findings in journals.
b. More botanists discovered the new species in the Amazon than the entomologist did but they jointly published the findings in journals.
c. More botanists photographed the new species in the Amazon than entomologists did but they jointly published the findings in journals.
d. More botanists photographed the new species in the Amazon than the entomologist did but they jointly published the findings in journals.

(23)   a. According to the insurance company, more people with high blood pressure got Alzheimers disease than diabetics did which is why they pay higher insurance premiums. b. According to the insurance company, more people with high blood pressure got Alzheimers disease than the diabetic did which is why they pay higher insurance premiums.
c. According to the insurance company, more people with high blood pressure got medical exams than diabetics did which is why they pay higher insurance premiums.
d. According to the insurance company, more people with high blood pressure got medical exams than the diabetic did which is why they pay higher insurance premiums.

(24)   a. More lawyers retired to Florida last year than judges did because of the temperate climate and beautiful beaches.
b. More lawyers retired to Florida last year than the judge did because of the temperate climate and beautiful beaches.
c. More lawyers vacationed in Florida last year than judges did because of the temperate climate and beautiful beaches.
d. More lawyers vacationed in Florida last year than the judge did because of the temperate climate and beautiful beaches.

(25)   a. As many aunts attended Jane's wedding as uncles did since Jane made sure everyone would be available.
b. As many aunts attended Jane's wedding as my uncle did since Jane made sure

everyone would be available.

c. As many aunts attended Jane's dinners as uncles did since Jane made sure everyone would be available.

d. As many aunts attended Jane's dinners as my uncle did since Jane made sure everyone would be available.

(26) a. As many talent scouts attended the 2011 championship game as coaches did since many of the players were top prospects.
b. As many talent scouts attended the 2011 championship game as the coach did since many of the players were top prospects.
c. As many talent scouts attended the 2011 home games as coaches did since many of the players were top prospects.
d. As many talent scouts attended the 2011 home games as the coach did since many of the players were top prospects.

(27) a. As many accountants at the firm went bald in their fifties as lawyers did in response to the stress and long hours.
b. As many accountants at the firm went bald in their fifties as my lawyer did in response to the stress and long hours.
c. As many accountants at the firm went golfing on their weekends as lawyers did in response to the stress and long hours.
d. As many accountants at the firm went golfing on their weekends as my lawyer did in response to the stress and long hours.

(28) a. This Sunday more bloggers broke the news about the celebrity divorce than newspapers did but the blogs were not very widely read.
b. This Sunday more bloggers broke the news about the celebrity divorce than our newspaper did but the blogs were not very widely read.
c. This week more bloggers discussed the details of the celebrity divorce than newspapers did but the blogs were not very widely read.
d. This week more bloggers discussed the details of the celebrity divorce than our newspaper did but the blogs were not very widely read.

(29) a. Last June, as many houses were painted professionally as apartments were because of the coupons advertised in the newspaper.
b. Last June, as many houses were painted professionally as my apartment was because of the coupons advertised in the newspaper.
c. Last spring, as many houses were cleaned professionally as apartments were because of the coupons advertised in the newspaper.
d. Last spring, as many houses were cleaned professionally as my apartment was because of the coupons advertised in the newspaper.

(30) a. In 2002 as many governors ran for president as senators did even though the public disapproved of their positions.
b. In 2002 as many governors ran for president as our senator did even though the public disapproved of their positions.
c. In 2002 as many governors praised the policy as senators did even though the public

disapproved of their positions.
d. In 2002 as many governors praised the policy as our senator did even though the public disapproved of their positions.

(31) a. As many country stars won a Grammy award last year as rock bands did but CD sales were best for rap music.
b. As many country stars won a Grammy award last year as my rock band did but CD sales were best for rap music.
c. As many country stars performed a love song last year as rock bands did but CD sales were best for rap music.
d. As many country stars performed a love song last year as my rock band did but CD sales were best for rap music.

(32) a. As many private universities banned Greek life as state schools did despite the concerns voiced by the university alumni.
b. As many private universities banned Greek life as our state school did despite the concerns voiced by the university alumni.
c. As many private universities sponsored sorority events as state schools did despite the concerns voiced by the university alumni.
d. As many private universities sponsored sorority events as our state school did despite the concerns voiced by the university alumni.

(33) a. As many fourth-graders had their tonsils out as third-graders did and they were all well-behaved during the procedure.
b. As many fourth-graders had their tonsils out as my third-grader did and they were all well-behaved during the procedure.
c. As many fourth-graders had their teeth cleaned as third-graders did and they were all well-behaved during the procedure.
d. As many fourth-graders had their teeth cleaned as my third-grader did and they were all well-behaved during the procedure.

(34) a. Last semester as many administrators at our school caught the chicken pox as teachers did which caused the office to be severely understaffed.
b. Last semester as many administrators at our school caught the chicken pox as our teacher did which caused the office to be severely understaffed.
c. Last semester as many administrators at our school caught the stomach flu as teachers did which caused the office to be severely understaffed.
d. Last semester as many administrators at our school caught the stomach flu as our teacher did which caused the office to be severely understaffed.

(35) a. As many high-schoolers dressed up on Halloween night this year as children did but the high-schoolers chose different types of costumes.
b. As many high-schoolers dressed up on Halloween night this year as my child did but the high-schoolers chose different types of costumes.
c. As many high-schoolers dressed up for costume parties this year as children did but the high-schoolers chose different types of costumes.

d. As many high-schoolers dressed up for costume parties this year as my child did but the high-schoolers chose different types of costumes.

(36)   a. As many dishwashers were fired on their first day as cooks were because the restaurant had high standards for employees.
b. As many dishwashers were fired on their first day as our cook was because the restaurant had high standards for employees.
c. As many dishwashers were reprimanded during their first week as cooks were because the restaurant had high standards for employees.
d. As many dishwashers were reprimanded during their first week as our cook was because the restaurant had high standards for employees.

(37)   a. As many actors brought dates to the movie's premiere as actresses did and one actor also brought all his friends.
b. As many actors brought dates to the movie's premiere as the actress did and one actor also brought all his friends.
c. As many actors brought dates to the producer's parties as actresses did and one actor also brought all his friends.
d. As many actors brought dates to the producer's parties as the actress did and one actor also brought all his friends.

(38)   a. As many taxi drivers joined the union last week as bus drivers did because of the way the union renegotiated salaries.
b. As many taxi drivers joined the union last week as the bus driver did because of the way the union renegotiated salaries.
c. As many taxi drivers criticized the union last year as bus drivers did because of the way the union renegotiated salaries.
d. As many taxi drivers criticized the union last year as the bus driver did because of the way the union renegotiated salaries.

(39)   a. During the Vietnam War, as many high school dropouts were drafted into the army as graduates were since the armed forces were so short staffed.
b. During the Vietnam War, as many high school dropouts were drafted into the army as the graduate was since the armed forces were so short staffed.
c. During the Vietnam War, as many high school dropouts were promoted in the army as graduates were since the armed forces were so short staffed.
d. During the Vietnam War as many high school dropouts were promoted in the army as the graduate was since the armed forces were so short staffed.

(40)   a. As many cats were adopted from the animal shelter last month as dogs were thanks to the donations by the local residents.
b. As many cats were adopted from the animal shelter last month as the dog was thanks to the donations by the local residents.
c. As many cats were groomed at the animal shelter last month as dogs were thanks to the donations by the local residents.
d. As many cats were groomed at the animal shelter last month as the dog was thanks to the donations by the local residents.

(41)   a. As many Middle Eastern governments banned the website as communist countries
        did resulting in widespread public outrage and numerous protests.
        b. As many Middle Eastern governments banned the website as the communist country
        did resulting in widespread public outrage and numerous protests.
        c. As many Middle Eastern governments censored the websites as communist countries
        did resulting in widespread public outrage and numerous protests.
        d. As many Middle Eastern governments censored the websites as the communist
        country did resulting in widespread public outrage and numerous protests.

(42)   a. Over winter break, as many skiers broke collar bones in Aspen as snowboarders did
        because of the record breaking winter this year.
        b. Over winter break, as many skiers broke collar bones in Aspen as the snowboarder
        did because of the record breaking winter this year.
        c. Over winter break, as many skiers booked private lessons in Aspen as snowboarders
        did because of the record breaking winter this year.
        d. Over winter break, as many skiers booked private lessons in Aspen as the
        snowboarder did because of the record breaking winter this year.

(43)   a. This spring as many physics students passed Chemistry 101 as chemistry majors did
        since chemistry is required for all science majors.
        b. This spring as many physics students passed Chemistry 101 as the chemistry major
        did since chemistry is required for all science majors.
        c. This spring as many physics students attended chemistry lectures as chemistry majors
        did since chemistry is required for all science majors.
        d. This spring as many physics students attended chemistry lectures as the chemisty
        major did since chemistry is required for all science majors.

(44)   a. As many girls in our class were born in September as boys were but it must have just
        been a coincidence.
        b. As many girls in our class were born in September as the boy was but it must have
        just been a coincidence.
        c. As many girls in our class were absent in September as boys were but it must have
        just been a coincidence.
        d. As many girls in our class were absent in September as the boy was but it must have
        just been a coincidence.

(45)   a. As many innocent politicians resigned from office during the scandal as guilty ones
        did due to all the attention from the media.
        b. As many innocent politicians resigned from office during the scandal as the guilty
        one did due to all the attention from the media.
        c. As many innocent politicians issued press releases during the scandal as guilty ones
        did due to all the attention from the media.
        d. As many innocent politicians issued press releases during the scandal as the guilty
        one did due to all the attention from the media.

(46)   a. This New Year's, as many motorcyclists died in accidents as drunk drivers did
        because of the rainy weather and road conditions.

b. This New Year's, as many motorcyclists died in accidents as the drunk driver did because of the rainy weather and road conditions.
c. This holiday season, as many motorcyclists got in accidents as drunk drivers did because of the rainy weather and road conditions.
d. This holiday season, as many motorcyclists got in accidents as the drunk driver did because of the rainy weather and road conditions.

(47)   a. Last year, as many freshmen were expelled from school as seniors were so the high school implemented tougher discipline policies.
b. Last year, as many freshmen were expelled from school as the senior was so the high school implemented tougher discipline policies.
c. Last year, as many freshmen were late to class as seniors were so the high school implemented tougher discipline policies.
d. Last year, as many freshmen were late to class as the senior was so the high school implemented tougher discipline policies.

(48)   a. As many stargazers saw the 2009 lunar eclipse as astronomers did after the new telescopes became much more affordable.
b. As many stargazers saw the 2009 lunar eclipse as the astronomer did after the new telescopes became much more affordable.
c. As many stargazers saw shooting stars in 2009 as astronomers did after the new telescopes became much more affordable.
d. As many stargazers saw shooting stars in 2009 as the astronomer did after the new telescopes became much more affordable.

9.2   Experiment 2

Conditions:

a. Nonrepeatable, control,
b. Nonrepeatable, illusion,
c. Repeatable, control,
d. Repeatable, illusion

(1)   a. More relatives went to my 21st birthday party than friends did because my parents invited everyone that they knew.
b. More relatives went to my 21st birthday party than my friends did because my parents invited everyone that they knew.
c. More relatives went to my childhood birthday parties than friends did because my parents invited everyone that they knew.
d. More relatives went to my childhood birthday parties than my friends did because my parents invited everyone that they knew.

(2)   a. More students at my school joined Facebook than professors did when the site was first launched in 2004.
b. More students at my school joined Facebook than my professors did when the site was first launched in 2004.
c. More students at my school used Facebook than professors did when the site was first launched in 2004.
d. More students at my school used Facebook than my professors did when the site was first launched in 2004.

(3)   a. This Friday, more convenience store clerks claimed jackpot lottery winnings than customers did because clerks got the tickets at discounted prices.
b. This Friday, more convenience store clerks claimed jackpot lottery winnings than my customers did because clerks got the tickets at discounted prices.
c. This year, more convenience store clerks bought jackpot lottery tickets than customers did because clerks got the tickets at discounted prices.
d. This year, more convenience store clerks bought jackpot lottery tickets than my customers did because clerks got the tickets at discounted prices.

(4)   a. More senior citizens had their appendix removed than teenagers did because doctors take extra precautions with the elderly.
b. More senior citizens had their appendix removed than my teenagers did because doctors take extra precautions with the elderly.
c. More senior citizens had their moles removed than teenagers did because doctors take extra precautions with the elderly.
d. More senior citizens had their moles removed than my teenagers did because doctors take extra precautions with the elderly.

(5)   a. Last spring more ducks hatched from eggs than chickens did due to the declining population of local predators.
b. Last spring more ducks hatched from eggs than my chickens did due to the declining population of local predators.
c. Last spring more ducks laid eggs here than chickens did due to the declining population of local predators.
d. Last spring more ducks laid eggs here than my chickens did due to the declining population of local predators.

(6)   a. More strawberry plants were ruined during the drought than geraniums were because geraniums tend to need very little water.
b. More strawberry plants were ruined during the drought than my geraniums were because geraniums tend to need very little water.
c. More strawberry plants were watered during the drought than geraniums were because geraniums tend to need very little water.
d. More strawberry plants were watered during the drought than my geraniums were because geraniums tend to need very little water.

(7)   a. More quarterbacks were selected in the 2011 NFL draft than wide receivers were even though the quarterbacks didn't play very well.

b. More quarterbacks were selected in the 2011 NFL draft than our wide receivers were even though the quarterbacks didn't play very well.
c. More quarterbacks were interviewed after the 2011 NFL draft than wide receivers were even though the quarterbacks didn't play very well.
d. More quarterbacks were interviewed after the 2011 NFL draft than our wide receivers were even though the quarterbacks didn't play very well.

(8) a. Last night more passengers in coach were bumped from the flight to Chicago than first-class passengers were but our airline got all passengers there eventually.
b. Last night more passengers in coach were bumped from the flight to Chicago than our first-class passengers were but our airline got all passengers there eventually.
c. Last year more passengers in coach were bumped from the flights to Chicago than first-class passengers were but the airline got all passengers there eventually.
d. Last year more passengers in coach were bumped from the flights to Chicago than our first-class passengers were but the airline got all passengers there eventually.

(9) a. Last fall, more engineers relocated to San Francisco than accountants did in order to work in our corporate offices.
b. Last fall, more engineers relocated to San Francisco than our accountants did in order to work in our corporate offices.
c. Last year, more engineers traveled to San Francisco than accountants did in order to work in our corporate offices.
d. Last year, more engineers traveled to San Francisco than our accountants did in order to work in our corporate offices.

(10) a. More businesses in our town burned down in the fire than schools did and the local residents all bought homeowners insurance.
b. More businesses in our town burned down in the fire than our schools did and the local residents all bought homeowners insurance.
c. More businesses in our town held drills after the fire than schools did and the local residents all bought homeowners insurance.
d. More businesses in our town held drills after the fire than our schools did and the local residents all bought homeowners insurance.

(11) a. More financial analysts graduated from Harvard's business school than managers did even though our company is close to Boston.
b. More financial analysts graduated from Harvard's business school than our managers did even though our company is close to Boston.
c. More financial analysts hired from Harvard's business school than managers did even though our company is close to Boston.
d. More financial analysts hired from Harvard's business school than our managers did even though our company is close to Boston.

(12) a. This year more managers were fired from the company for poor performance than assistants were because the company was prioritizing team leadership skills.
b. This year more managers were fired from the company for poor performance than our assistants were because the company was prioritizing team leadership skills.

c. This year more managers were rewarded by the company for good performance than assistants were because the company was prioritizing team leadership skills.

d. This year more managers were rewarded by the company for good performance than our assistants were because the company was prioritizing team leadership skills.

(13) a. As many aunts attended my sister's wedding as uncles did since Jane made sure everyone would be available.
b. As many aunts attended my sister's wedding as my uncles did since Jane made sure everyone would be available.
c. As many aunts attended my sister's dinners as uncles did since Jane made sure everyone would be available.
d. As many aunts attended my sister's dinners as my uncles did since Jane made sure everyone would be available.

(14) a. Last semester as many administrators at my school caught the chicken pox as teachers did which caused the office to be severely understaffed.
b. Last semester as many administrators at my school caught the chicken pox as my teachers did which caused the office to be severely understaffed.
c. Last semester as many administrators at my school caught the stomach flu as teachers did which caused the office to be severely understaffed.
d. Last semester as many administrators at my school caught the stomach flu as my teachers did which caused the office to be severely understaffed.

(15) a. As many accountants at the firm went bald in their fifties as lawyers did in response to the stress and long hours.
b. As many accountants at the firm went bald in their fifties as my lawyers did in response to the stress and long hours.
c. As many accountants at the firm went golfing on their weekends as lawyers did in response to the stress and long hours.
d. As many accountants at the firm went golfing on their weekends as my lawyers did in

(16) a. As many fourth-graders had their tonsils out as third-graders did and they were all well-behaved during the procedure.
b. As many fourth-graders had their tonsils out as my third-graders did and they were all well-behaved during the procedure.
c. As many fourth-graders had their teeth cleaned as third-graders did and they were all well-behaved during the procedure.
d. As many fourth-graders had their teeth cleaned as my third-graders did and they were all well-behaved during the procedure.

(17) a. Last June as many trucks were painted by the mechanics as cars were due to a promotion that included the service.
b. Last June as many trucks were painted by the mechanics as my cars were due to a promotion that included the service.
c. Last year as many trucks were cleaned by the mechanics as cars were due to a promotion that included the service.

d. Last year as many trucks were cleaned by the mechanics as my cars were due to a promotion that included the service.

(18)   a. As many high-schoolers dressed up on Halloween night this year as children did but the high-schoolers chose different types of costumes.
b. As many high-schoolers dressed up on Halloween night this year as my children did but the high-schoolers chose different types of costumes.
c. As many high-schoolers dressed up for costume parties this year as children did but the high-schoolers chose different types of costumes.
d. As many high-schoolers dressed up for costume parties this year as my children did but the high-schoolers chose different types of costumes.

(19)   a. As many talent scouts attended the 2011 championship game as coaches did since many of the players were top prospects.
b. As many talent scouts attended the 2011 championship game as our coaches did since many of the players were top prospects.
c. As many talent scouts attended the 2011 home games as coaches did since many of the players were top prospects.
d. As many talent scouts attended the 2011 home games as our coaches did since many of the players were top prospects.

(20)   a. As many private universities banned Greek life as state schools did despite the concerns voiced by the university alumni.
b. As many private universities banned Greek life as our state schools did despite the concerns voiced by the university alumni.
c. As many private universities sponsored sorority events as state schools did despite the concerns voiced by the university alumni.
d. As many private universities sponsored sorority events as our state schools did despite the concerns voiced by the university alumni.

(21)   a. As many houses were demolished for our town's new rail system as businesses were but the city compensated residents for all disturbances.
b. As many houses were demolished for our town's new rail system as our businesses were but the city compensated residents for all disturbances.
c. As many houses were affected by our town's new rail system as businesses were but the city compensated residents for all disturbances.
d. As many houses were affected by our town's new rail system as our businesses were but the city compensated residents for all disturbances.

(22)   a. As many country stars on our record label won a Grammy award last year as rock bands did but CD sales were best for rap music.
b. As many country stars on our record label won a Grammy award last year as our rock bands did but CD sales were best for rap music.
c. As many country stars on our record label performed a love song last year as rock bands did but CD sales were best for rap music.
d. As many country stars on our record label performed a love song last year as our rock bands did but CD sales were best for rap music.

(23)   a. In 2002 as many governors ran for president as senators did even though the public disapproved of their positions.
b. In 2002 as many governors ran for president as our senators did even though the public disapproved of their positions.
c. In 2002 as many governors praised the policy as senators did even though the public disapproved of their positions.
d. In 2002 as many governors praised the policy as our senators did even though the public disapproved of their positions.

(24)   a. As many dishwashers were fired on their first day as cooks were because the restaurant had high standards for employees.
b. As many dishwashers were fired on their first day as our cooks were because the restaurant had high standards for employees.
c. As many dishwashers were reprimanded during their first month as cooks were because the restaurant had high standards for employees.
d. As many dishwashers were reprimanded during their first month as our cooks were because the restaurant had high standards for employees.

9.3   Experiment 3

Conditions:

        a. Dependent plural, control,
b. Semantic plural, control,
c. Dependent plural, illusion,
d. Semantic plural, illusion

(1)   a. More crime movies have compelling titles than TV shows do.
b. More crime movies have violent scenes than TV shows do.
c. More crime movies have compelling titles than the TV show does.
d. More crime movies have violent scenes than the TV show does.

(2)   a. More biologists deserve Nobel prizes than astronomers do.
b. More biologists deserve research grants than astronomers do.
c. More biologists deserve Nobel prizes than the astronomer does.
d. More biologists deserve research grants than the astronomer does.

(3)   a. More violent criminals have life sentences than nonviolent criminals do.
b. More violent criminals have gang tattoos than nonviolent criminals do.
c. More violent criminals have life sentences than the nonviolent criminal does.
d. More violent criminals have gang tattoos than the nonviolent criminal does.

(4)   a. More linguistics professors have doctorates in psychology than neuroscientists do.
b. More linguistics professors have students in psychology than neuroscientists do.
c. More linguistics professors have doctorates in psychology than the neuroscientist

does.
d. More linguistics professors have students in psychology than the neuroscientist does.

(5)   a. More Germans have winter birthdays than Brazilians do.
        b. More Germans have winter jackets than Brazilians do.
        c. More Germans have winter birthdays than the Brazilian does.
        d. More Germans have winter jackets than the Brazilian does.

(6)   a. More commercial streets have appropriate speed limits than residential streets do.
        b. More commercial streets have brightly painted crosswalks than residential streets do.
        c. More commercial streets have appropriate speed limits than the residential street does.
        d. More commercial streets have brightly painted crosswalks than the residential street does.

(7)   a. More baby boomers have full-time jobs than recent graduates do.
        b. More baby boomers have stock investments than recent graduates do.
        c. More baby boomers have full-time jobs than the recent graduate does.
        d. More baby boomers have stock investments than the recent graduate does.

(8)   a. More rappers have contracts with Interscope Records than pop stars do.
        b. More rappers have contacts at Interscope Records than pop stars do.
        c. More rappers have contracts with Interscope Records than the pop star does.
        d. More rappers have contacts at Interscope Records than the pop star does.

(9)   a. More cruise ships have foreign captains than yachts do.
        b. More cruise ships have underage passengers than yachts do.
        c. More cruise ships have foreign captains than the yacht does.
        d. More cruise ships have underage passengers than the yacht does.

(10)   a. More teenagers have driver's licenses than college students do.
        b. More teenagers have homework assignments than college students do.
        c. More teenagers have driver's licenses than the college student does.
        d. More teenagers have homework assignments than the college student does.

(11)   a. More actresses want nose jobs than actors do.
        b. More actresses want lead roles than actors do.
        c. More actresses want nose jobs than the actor does.
        d. More actresses want lead roles than the actor does.

(12)   a. More bistros have wine bars than cafés do.
        b. More bistros have red wines than cafés do.
        c. More bistros have wine bars than the café does.
        d. More bistros have red wines than the café does.

(13)   a. More cats have striped tails than dogs do.
        b. More cats have mouse toys than dogs do.

c. More cats have striped tails than my dog does.
d. More cats have mouse toys than my dog does.

(14)   a. More hipsters have handlebar mustaches than grandfathers do.
        b. More hipsters have vinyl records than grandfathers do.
        c. More hipsters have handlebar mustaches than my grandfather does.
        d. More hipsters have vinyl records than my grandfather does.

(15)   a. More private schools have admissions applications than public schools do.
        b. More private schools have admissions requirements than public schools do.
        c. More private schools have admissions applications than my public school does.
        d. More private schools have admissions requirements than my public school does.

(16)   a. More wives have expensive hairdos than husbands do.
        b. More wives have beauty products than husbands do.
        c. More wives have expensive hairdos than my husband does.
        d. More wives have beauty products than my husband does.

(17)   a. More houses have swimming pools than apartment complexes do.
        b. More houses have modern appliances than apartment complexes do.
        c. More houses have swimming pools than my apartment complex does.
        d. More houses have modern applicances than my apartment complex does.

(18)   a. More desktops have DVD drives than laptops do.
        b. More desktops have computer viruses than laptops do.
        c. More desktops have DVD drives than my laptop does.
        d. More desktops have computer viruses than my laptop does.

(19)   a. More banks have wealthy CEOs than startups do.
        b. More banks have wealthy shareholders than startups do.
        c. More banks have wealthy CEOs than our startup does.
        d. More banks have wealthy shareholders than our startup does.

(20)   a. More high schools have school mascots than elementary schools do.
        b. More high schools have gym teachers than elementary schools do.
        c. More high schools have school mascots than the elementary school does.
        d. More high schools have gym teachers than the elementary school does.

(21)   a. More public libraries have science fiction sections than school libraries do.
        b. More public libraries have science fiction books than school libraries do.
        c. More public libraries have science fiction sections than our school library does.
        d. More public libraries have science fiction books than our school library does.

(22)   a. More small towns have Spanish names than cities do.
        b. More small towns have Spanish buildings than cities do.
        c. More small towns have Spanish names than our city does.
        d. More small towns have Spanish buildings than our city does.

(23)   a. More luxury cars have custom license plates than SUVs do.
         b. More luxury cars have foreign auto parts than SUVs do.
         c. More luxury cars have custom license plates than our SUV does.
         d. More luxury cars have foreign auto parts than our SUV does.

(24)   a. More backyards have grass lawns than front yards do.
         b. More backyards have lawn chairs than front yards do.
         c. More backyards have grass lawns than our front yard does.
         d. More backyards have lawn chairs than our front yard does.

9.4   Experiment 4

Conditions:

         a. Dependent plural, control,
         b. Semantic plural, control,
         c. Dependent plural, illusion,
         d. Semantic plural, illusion

(1)   a. The crime movies have compelling titles, and the TV show does too.
         b. The crime movies have violent scenes, and the TV show does too.
         c. More crime movies have compelling titles than the TV show does.
         d. More crime movies have violent scenes than the TV show does.

(2)   a. The biologists deserve Nobel prizes, and the astronomer does too.
         b. The biologists deserve research grants, and the astronomer does too.
         c. More biologists deserve Nobel prizes than the astronomer does.
         d. More biologists deserve research grants than the astronomer does.

(3)   a. The violent criminals have life sentences, and the nonviolent criminal does too.
         b. The violent criminals have gang tattoos, and the nonviolent criminal does too.
         c. More violent criminals have life sentences than the nonviolent criminal does.
         d. More violent criminals have gang tattoos than the nonviolent criminal does.

(4)   a. The linguistics professors have doctorates in psychology, and the neuroscientist does
         too.
         b. The linguistics professors have students in psychology, and the neuroscientist does
         too.
         c. More linguistics professors have doctorates in psychology than the neuroscientist
         does.
         d. More linguistics professors have students in psychology than the neuroscientist does.

(5)   a. The Germans have winter birthdays, and the Brazilian does too.
         b. The Germans have winter jackets, and the Brazilian does too.

268

c. More Germans have winter birthdays than the Brazilian does.
d. More Germans have winter jackets than the Brazilian does.

(6)   a. The commercial streets have appropriate speed limits, and the residential street does too.
b. The commercial streets have brightly painted crosswalks, and the residential street does too.
c. More commercial streets have appropriate speed limits than the residential street does.
d. More commercial streets have brightly painted crosswalks than the residential street does.

(7)   a. The baby boomers have full-time jobs, and the recent graduate does too.
b. The baby boomers have stock investments, and the recent graduate does too.
c. More baby boomers have full-time jobs than the recent graduate does.
d. More baby boomers have stock investments than the recent graduate does.

(8)   a. The rappers have contracts with Interscope Records, and the pop star does too.
b. The rappers have contacts at Interscope Records than pop star does too.
c. More rappers have contracts with Interscope Records than the pop star does.
d. More rappers have contacts at Interscope Records than the pop star does.

(9)   a. The cruise ships have foreign captains, and the yacht does too.
b. The cruise ships have underage passengers, and the yacht does too.
c. More cruise ships have foreign captains than the yacht does.
d. More cruise ships have underage passengers than the yacht does.

(10)   a. The teenagers have driver's licenses, and the college student does too.
b. The teenagers have homework assignments, and the college student does too.
c. More teenagers have driver's licenses than the college student does.
d. More teenagers have homework assignments than the college student does.

(11)   a. The actresses want nose jobs, and the actor does too.
b. The actresses want lead roles, and the actor does too.
c. More actresses want nose jobs than the actor does.
d. More actresses want lead roles than the actor does.

(12)   a. The bistros have wine bars, and the café does too.
b. The bistros have red wines, and the café does too.
c. More bistros have wine bars than the café does.
d. More bistros have red wines than the café does.

(13)   a. The cats have striped tails, and the dog does too.
b. The cats have mouse toys, and the dog does too.
c. More cats have striped tails than my dog does.
d. More cats have mouse toys than my dog does.

(14)   a. The hipsters have handlebar mustaches, and the grandfather does too.
        b. The hipsters have vinyl records, and the grandfather does too.
        c. More hipsters have handlebar mustaches than my grandfather does.
        d. More hipsters have vinyl records than my grandfather does.

(15)   a. The private schools have admissions applications, and the public school does too.
        b. The private schools have admissions requirements, and the public school does too.
        c. More private schools have admissions applications than my public school does.
        d. More private schools have admissions requirements than my public school does.

(16)   a. The wives have expensive hairdos, and the husband does too.
        b. The wives have beauty products, and the husband does too.
        c. More wives have expensive hairdos than my husband does.
        d. More wives have beauty products than my husband does.

(17)   a. The houses have swimming pools, and the apartment complex does too.
        b. The houses have modern appliances, and the apartment complex does too.
        c. More houses have swimming pools than my apartment complex does.
        d. More houses have modern appliances than my apartment complex does.

(18)   a. The desktops have DVD drives, and the laptop does too.
        b. The desktops have computer viruses, and the laptop does too.
        c. More desktops have DVD drives than my laptop does.
        d. More desktops have computer viruses than my laptop does.

(19)   a. The banks have wealthy CEOs, and the startup does too.
        b. The banks have wealthy shareholders, and the startup does too.
        c. More banks have wealthy CEOs than our startup does.
        d. More banks have wealthy shareholders than our startup does.

(20)   a. The high schools have school mascots, and the elementary school does too.
        b. The high schools have gym teachers, and the elementary school does too.
        c. More high schools have school mascots than the elementary school does.
        d. More high schools have gym teachers than the elementary school does.

(21)   a. The public libraries have science fiction sections, and the school library does too.
        b. The public libraries have science fiction books, and the school library does too.
        c. More public libraries have science fiction sections than our school library does.
        d. More public libraries have science fiction books than our school library does.

(22)   a. The small towns have Spanish names, and the city does too.
        b. The small towns have Spanish buildings, and the city does too.
        c. More small towns have Spanish names than our city does.
        d. More small towns have Spanish buildings than our city does.

(23)   a. The luxury cars have custom license plates, and the SUV does too.
        b. The luxury cars have foreign auto parts, and the SUV does too.

c. More luxury cars have custom license plates than our SUV does.
d. More luxury cars have foreign auto parts than our SUV does.

(24)   a. The backyards have grass lawns, and the front yard does too.
        b. The backyards have lawn chairs, and the front yard does too.
        c. More backyards have grass lawns than our front yard does.
        d. More backyards have lawn chairs than our front yard does.

9.5   Experiment 5

Conditions:

        a. Dependent plural, singular *than*-clause subject,
        b. Semantic plural, singular *than*-clause subject,
        c. Dependent plural, plural *than*-clause subject,
        d. Semantic plural, plural *than*-clause subject

(1)   a. More crime movies have compelling titles than the TV show does.
        b. More crime movies have violent scenes than the TV show does.
        c. More crime movies have compelling titles than the TV shows do.
        d. More crime movies have violent scenes than the TV shows do.

(2)   a. More biologists deserve Nobel prizes than the astronomer does.
        b. More biologists deserve research grants than the astronomer does.
        c. More biologists deserve Nobel prizes than the astronomers do.
        d. More biologists deserve research grants than the astronomers do.

(3)   a. More violent criminals have life sentences than the nonviolent criminal does.
        b. More violent criminals have gang tattoos than the nonviolent criminal does.
        c. More violent criminals have life sentences than the nonviolent criminals do.
        d. More violent criminals have gang tattoos than the nonviolent criminals do.

(4)   a. More linguistics professors have doctorates in psychology than the neuroscientist does.
        b. More linguistics professors have students in psychology than the neuroscientist does.
        c. More linguistics professors have doctorates in psychology than the neuroscientists do.
        d. More linguistics professors have students in psychology than the neuroscientists do.

(5)   a. More Germans have winter birthdays than the Brazilian does.
        b. More Germans have winter jackets than the Brazilian does
        c. More Germans have winter birthdays than the Brazilians do.
        d. More Germans have winter jackets than the Brazilians do.

(6) a. More commercial streets have appropriate speed limits than the residential street does.
b. More commercial streets have brightly painted crosswalks than the residential street does.
c. More commercial streets have appropriate speed limits than the residential streets do.
d. More commercial streets have brightly painted crosswalks than the residential streets do.

(7) a. More baby boomers have full-time jobs than the recent graduate does.
b. More baby boomers have stock investments than the recent graduate does.
c. More baby boomers have full-time jobs than the recent graduates do.
d. More baby boomers have stock investments than the recent graduates do.

(8) a. More rappers have contracts with Interscope Records than the pop star does.
b. More rappers have contacts at Interscope Records than the pop star does.
c. More rappers have contracts with Interscope Records than the pop stars do.
d. More rappers have contacts at Interscope Records than the pop stars do.

(9) a. More cruise ships have foreign captains than the yacht does.
b. More cruise ships have underage passengers than the yacht does.
c. More cruise ships have foreign captains than the yachts do.
d. More cruise ships have underage passengers than the yachts do.

(10) a. More teenagers have driver's licenses than the college student does.
b. More teenagers have homework assignments than the college student does.
c. More teenagers have driver's licenses than the college students do.
d. More teenagers have homework assignments than the college students do.

(11) a. More actresses want nose jobs than the actor does.
b. More actresses want lead roles than the actor does.
c. More actresses want nose jobs than the actors do.
d. More actresses want lead roles than the actors do.

(12) a. More bistros have wine bars than the café does.
b. More bistros have red wines than the café does.
c. More bistros have wine bars than the cafés do.
d. More bistros have red wines than the cafés do.

(13) a. More cats have striped tails than the dog does.
b. More cats have mouse toys than the dog does.
c. More cats have striped tails than the dogs do.
d. More cats have mouse toys than the dogs do.

(14) a. More hipsters have handlebar mustaches than the grandfather does.
b. More hipsters have vinyl records than the grandfather does.
c. More hipsters have handlebar mustaches than the grandfathers do.
d. More hipsters have vinyl records than the grandfathers do.

(15)   a. More private schools have admissions applications than the public school does.
         b. More private schools have admissions requirements than the public school does.
         c. More private schools have admissions applications than the public schools do.
         d. More private schools have admissions requirements than the public schools do.

(16)   a. More women have expensive hairdos than the man does.
         b. More women have beauty products than the man does.
         c. More women have expensive hairdos than the men do.
         d. More women have beauty products than the men do.

(17)   a. More houses have swimming pools than the apartment complex does.
         b. More houses have modern appliances than the apartment complex does.
         c. More houses have swimming pools than the apartment complexes do.
         d. More houses have modern appliances than the apartment complexes do.

(18)   a. More desktops have DVD drives than the laptop does.
         b. More desktops have computer viruses than the laptop does.
         c. More desktops have DVD drives than the laptops do.
         d. More desktops have computer viruses than the laptops do.

(19)   a. More banks have wealthy CEOs than the startup does.
         b. More banks have wealthy shareholders than the startup does.
         c. More banks have wealthy CEOs than the startups do.
         d. More banks have wealthy shareholders than the startups do.

(20)   a. More high schools have school mascots than the elementary school does.
         b. More high schools have gym teachers than the elementary school does.
         c. More high schools have school mascots than the elementary schools do.
         d. More high schools have gym teachers than the elementary schools do.

(21)   a. More public libraries have science fiction sections than the school library does.
         b. More public libraries have science fiction books than the school library does.
         c. More public libraries have science fiction sections than the school libraries do.
         d. More public libraries have science fiction books than the school libraries do.

(22)   a. More small towns have Spanish names than the city does.
         b. More small towns have Spanish buildings than the city does.
         c. More small towns have Spanish names than the cities do.
         d. More small towns have Spanish buildings than the cities do.

(23)   a. More luxury cars have custom license plates than the SUV does.
         b. More luxury cars have foreign auto parts than the SUV does.
         c. More luxury cars have custom license plates than the SUVs do.
         d. More luxury cars have foreign auto parts than the SUVs do.

(24)   a. More backyards have grass lawns than the front yards do.
         b. More backyards have lawn chairs than the front yards do.
         c. More backyards have grass lawns than the front yards do.

d. More backyards have lawn chairs than the front yards do.

9.6 Experiment 6

Conditions:

 a. Illusion, collective noun,
 b. Illusion, singular noun,
 c. Control, collective noun,
 d. Control, singular noun,

(1) a. More Millennials are unaffiliated with organized religion than the older generation is.
 b. More Millennials are unaffiliated with organized religion than the baby boomer is.
 c. The Millennials are unaffiliated with organized religion, and the older generation is too.
 d. The Millennials are unaffiliated with organized religion, and the baby boomer is too.

(2) a. More musicians at the competition are classically trained than the choir is.
 b. More musicians at the competition are classically trained than the choir is.
 c. The musicians at the competition are classically trained, and the choir is too.
 d. The musicians at the competition are classically trained, and the singer is too.

(3) a. More workers are disabled due to the accident at the plant than the management is.
 b. More workers are disabled due to the accident at the plant than the manager is.
 c. The workers are disabled due to the accident at the plant, and the management is too.
 d. The workers are disabled due to the accident at the plant, and the manager is too.

(4) a. More teachers at the school are in the union than the administrative staff is.
 b. More teachers at the school are in the union than the administrative assistant is.
 c. The teachers at the school are in the union, and the administrative staff is too.
 d. The teachers at the school are in the union, and the administrative assistant is too.

(5) a. More audience members at the playhouse are American than the cast is.
 b. More audience members at the playhouse are American than the actor is.
 c. The audience members at the playhouse are American, and the cast is too.
 d. The audience members at the playhouse are American, and the actor is too.

(6) a. More athletes are on scholarship at the university than the debate team is.
 b. More athletes are on scholarship at the university than the school valedictorian is.
 c. The athletes are on scholarship at the university, and the debate team is too.
 d. The athletes are on scholarship at the university, and the school valedictorian is too.

(7) a. More Westerners are allergic to peanuts than the rural tribe is.
 b. More Westerners are allergic to peanuts than the rural tribesman is.

c. The Westerners are allergic to peanuts, and the rural tribe is too.
d. The Westerners are allergic to peanuts, and the rural tribesman is too.

(8)   a. More firemen are capable of handling the emergency than the bomb squad is.
b. More firemen are capable of handling the emergency than the police chief is.
c. The firemen are capable of handling the emergency, and the bomb squad is too.
d. The firemen are capable of handling the emergency, and the police chief is too.

(9)   a. More suspected terrorists are under investigation by the FBI than the gang is.
b. More suspected terrorists are under investigation by the FBI than the gangster is.
c. The suspected terrorists are under investigation by the FBI, and the gang is too.
d. The suspected terrorists are under investigation by the FBI, and the gangster is too.

(10)   a. More priests are affiliated with the charity than the congregation is.
b. More priests are affiliated with the charity than the nun is.
c. The priests are affiliated with the charity, and the congregation is too.
d. The priests are affiliated with the charity, and the nun is too.

(11)   a. More bees are dying because of habitat changes than the ant colony is.
b. More bees are dying because of habitat changes than the boa constrictor is.
c. The bees are dying because of habitat changes, and the ant colony is too.
d. The bees are dying because of habitat changes, and the boa constrictor is too.

(12)   a. More military tanks are steel-reinforced than the Navy fleet is.
b. More military tanks are steel-reinforced than the Navy ship is.
c. The military tanks are steel-reinforced, and the Navy fleet is too.
d. The military tanks are steel-reinforced, and the Navy ship is too.

(13)   a. More business owners are incensed by the new city ordinance than the homeowners association is.
b. More business owners are incensed by the new city ordinance than the new homeowner is.
c. The business owners are incensed by the new city ordinance, and the homeowners association is too.
d. The business owners are incensed by the new city ordinance, and the new homeowner is too.

(14)   a. More passengers are nervous about the new airline policy than the flight crew is.
b. More passengers are nervous about the new airline policy than the flight attendant is.
c. The passengers are nervous about the new airline policy, and the flight crew is too.
d. The passengers are nervous about the new airline policy, and the flight attendant is too.

(15)   a. More foreign-born diplomats are familiar with the trade agreement than the American population is.
b. More foreign-born diplomats are familiar with the trade agreement than the American citizen is.
c. The foreign-born diplomats are familiar with the trade agreement, and the American

population is too.

d. The foreign-born diplomats are familiar with the trade agreement, and the American citizen is too.

(16)   a. More politicians are skeptical of the new policy than the public is.
b. More politicians are skeptical of the new policy than the voter is.
c. The politicians are skeptical of the new policy, and the public is too.
d. The politicians are skeptical of the new policy, and the voter is too.

(17)   a. More politicians are happy with the transportation workers' compensation than the union is.
b. More politicians are happy with the transportation workers' compensation than the worker is.
c. The politicians are happy with the transportation workers' compensation, and the union is too.
d. The politicians are happy with the transportation workers' compensation, and the worker is too.

(18)   a. More cheerleaders are excited about the home game than the team is.
b. More cheerleaders are excited about the home game than the quarterback is.
c. The cheerleaders are excited about the home game, and the team is too.
d. The cheerleaders are excited about the home game, and the quarterback is too.

(19)   a. More police officers are concerned about the recent criminal activity than the neighborhood is.
b. More police officers are concerned about the recent criminal activity than the neighbor is.
c. The police officers are concerned about the recent criminal activity, and the neighborhood is too.
d. The police officers are concerned about the recent criminal activity, and the neighbor is too.

(20)   a. More doctors are optimistic about the experimental drug than the medical panel is.
b. More doctors are optimistic about the experimental drug than the medical researcher is.
c. The doctors are optimistic about the experimental drug, and the medical panel is too.
d. The doctors are optimistic about the experimental drug, and the medical researcher is too.

(21)   a. More teachers are unhappy about the new curriculum than the class is.
b. More teachers are unhappy about the new curriculum than the student is.
c. The teachers are unhappy about the new curriculum, and the class is too.
d. The teachers are unhappy about the new curriculum, and the student is too.

(22)   a. More media outlets are biased against the defendant than the jury is.
b. More media outlets are biased against the defendant than the judge is.

c. The media outlets are biased against the defendant, and the jury is too.

d. The media outlets are biased against the defendant, and the judge is too.

(23)   a. More governors are opposed to the proposal than the congressional committee is.

b. More governors are opposed to the proposal than the local congressman is.

c. The governors are opposed to the proposal, and the congressional committee is too.

d. The governors are opposed to the proposal, and the local congressman is too.

(24)   a. More students are angry about the tuition increase than the faculty is.

b. More students are angry about the tuition increase than the professor is.

c. The students are angry about the tuition increase, and the faculty is too.

d. The students are angry about the tuition increase, and the professor is too.

9.7   Experiment 7a, Experiment 10

Conditions (Experiment 7a: presented with *to*-clause as shown; Experiment 10: presented without *to*-clause):

a. Negative degree quantifier, negative verb,

b. Negative degree quantifier, positive verb,

c. Positive degree quantifier, negative verb,

d. Positive degree quantifier, positive verb

(1)   a. In Maria's class, no test is too difficult (to fail)

b. In Maria's class, no test is too easy (to pass)

c. In Maria's class, no test is easy enough (to fail)

d. In Maria's class, no test is difficult enough (to pass)

(2)   a. In Laura's opinion, no relationship is too volatile (to fail)

b. In Laura's opinion, no relationship is too stable (to endure)

c. In Laura's opinion, no relationship is stable enough (to fail)

d. In Laura's opinion, no relationship is volatile enough (to endure)

(3)   a. According to Mark, no drug is too dangerous (to ban)

b. According to Mark, no drug is too safe (to legalize)

c. According to Mark, no drug is safe enough (to ban)

d. According to Mark, no drug is dangerous enough (to legalize)

(4)   a. Considering John's finances, no offer is too low (to reject)

b. Considering John's finances, no offer is too high (to accept)

c. Considering John's finances, no offer is high enough (to reject)

d. Considering John's finances, no offer is low enough (to accept)

(5)   a. In Gideon's opinion, no habit is too annoying (to discourage)

b. In Gideon's opinion, no habit is too practical (to encourage)

c. In Gideon's opinion, no habit is practical enough (to discourage)

d. In Gideon's opinion, no habit is annoying enough (to encourage)

(6)   a. According to the geologist, no data is too complicated (to misinterpret)

b. According to the geologist, no data is too straightforward (to interpret)

c. According to the geologist, no data is straightforward enough (to misinterpret)

d. According to the geologist, no data is complicated enough (to interpret)

(7)   a. According to the politician, no social program is too wasteful (to oppose)

b. According to the politician, no social program is too efficient (to support)

c. According to the politician, no social program is efficient enough (to oppose)

d. According to the politician, no social program is wasteful enough (to support)

(8)   a. During Kipp's presidential campaign, no rumor was too outlandish (to deny)

b. During Kipp's presidential campaign, no rumor was too realistic (to confirm)

c. During Kipp's presidential campaign, no rumor was realistic enough (to deny)

d. During Kipp's presidential campaign, no rumor was outlandish enough (to confirm)

(9)   a. For someone like Alex, no memory is too distant (to forget)

b. For someone like Alex, no memory is too recent (to recall)

c. For someone like Alex, no memory is recent enough (to forget)

d. For someone like Alex, no memory is distant enough (to recall)

(10)   a. When it comes to Sharon's friends, no event is too inconvenient (to skip)

b. When it comes to Sharon's friends, no event is too convenient (to attend)

c. When it comes to Sharon's friends, no event is convenient enough (to skip)

d. When it comes to Sharon's friends, no event is inconvenient enough (to attend)

(11)   a. In Sarah's opinion, no player is too slow (to lose)

b. In Sarah's opinion, no player is too fast (to win)

c. In Sarah's opinion, no player is fast enough (to lose)

d. In Sarah's opinion, no player is slow enough (to win)

(12)   a. According to the movie producer, no criticism is too trivial (to ignore)

b. According to the movie producer, no criticism is too serious (to address)

c. According to the movie producer, no criticism is serious enough (to ignore)

d. According to the movie producer, no criticism is trivial enough (to address)

(13)   a. Judging by Charlie's work, no error is too small (to overlook)

b. Judging by Charlie's work, no error is too big (to catch)

c. Judging by Charlie's work, no error is big enough (to overlook)

d. Judging by Charlie's work, no error is small enough (to catch)

(14)   a. When Lulu is talking, no sentence is too complex (to misunderstand)

b. When Lulu is talking, no sentence is too simple (to understand)

c. When Lulu is talking, no sentence is simple enough (to misunderstand)

d. When Lulu is talking, no sentence is complex enough (to understand)

(15)   a. For Hildy, no memento is too insignificant (to discard)
        b. For Hildy, no memento is too precious (to keep)
        c. For Hildy, no memento is precious enough (to discard)
        d. For Hildy, no memento is insignificant enough (to keep)

(16)   a. Frank believes that no advice is too stupid (to doubt)
        b. Frank believes that no advice is too sound (to trust)
        c. Frank believes that no advice is sound enough (to doubt)
        d. Frank believes that no advice is stupid enough (to trust)

9.8   Experiment 7b

Conditions:

    a. Negative nominal quantifier, negative degree quantifier
    b. Positive nominal quantifier, negative degree quantifier
    c. Negative nominal quantifier, positive degree quantifier
    d. Positive nominal quantifier, positive degree quantifier

(1)   a. In Maria's class, no test is too difficult to fail
       b. In Maria's class, every test is too difficult to fail.
       c. In Maria's class, no test is easy enough to fail.
       d. In Maria's class, every test is easy enough to fail.

(2)   a. In Laura's opinion, no relationship is too volatile to fail.
       b. In Laura's opinion, all relationships are too volatile to fail.
       c. In Laura's opinion, no relationship is stable enough to fail.
       d. In Laura's opinion, all relationships are stable enough to fail.

(3)   a. According to Mark, no drug is too dangerous to ban.
       b. According to Mark, every drug is too dangerous to ban.
       c. According to Mark, no drug is safe enough to ban.
       d. According to Mark, every drug is safe enough to ban.

(4)   a. When Jack inspects the uniforms, no stain is too subtle to miss.
       b. When Jack inspects the uniforms, all stains are too subtle to miss.
       c. When Jack inspects the uniforms, no stain is obvious enough to miss.
       d. When Jack inspects the uniforms, all stains are obvious enough to miss.

(5)   a. Considering John's finances, no salary offer is too low to reject.
       b. Considering John's finances, every salary offer is too low to reject.
       c. Considering John's finances, no salary offer is high enough to reject.
       d. Considering John's finances, every salary offer is high enough to reject.

(6)   a. Given Lily's relationship with her mother, no topic is too contentious to avoid.
        b. Given Lily's relationship with her mother, all topics are too contentious to avoid.
        c. Given Lily's relationship with her mother, no topic is benign enough to avoid.
        d. Given Lily's relationship with her mother, all topics are benign enough to avoid.

(7)   a. For someone like Alex, no memory is too distant to forget.
        b. For someone like Alex, every memory is too distant to forget.
        c. For someone like Alex, no memory is recent enough to forget.
        d. For someone like Alex, every memory is recent enough to forget.

(8)   a. With Liz's eating habits, no craving is too unhealthy to resist.
        b. With Liz's eating habits, all cravings are too unhealthy to resist.
        c. With Liz's eating habits, no craving is tempting enough to resist.
        d. With Liz's eating habits, all cravings are tempting enough to resist.

(9)   a. When it comes to Sharon's friends, no event is too inconvenient to skip.
        b. When it comes to Sharon's friends, every event is too inconvenient to skip.
        c. When it comes to Sharon's friends, no event is convenient enough to skip.
        d. When it comes to Sharon's friends, every event is convenient enough to skip.

(10)   a. According to the politician, no social program is too wasteful to oppose.
        b. According to the politician, all social programs are too wasteful to oppose.
        c. According to the politician, no social program is efficient enough to oppose.
        d. According to the politician, all social programs are efficient enough to oppose.

(11)   a. In Sarah's opinion, no player is too slow to lose.
        b. In Sarah's opinion, every player is too slow to lose.
        c. In Sarah's opinion, no player is fast enough to lose.
        d. In Sarah's opinion, every player is fast enough to lose.

(12)   a. Judging by Charlie's work, no error is too small to overlook.
        b. Judging by Charlie's work, all errors are too small to overlook.
        c. Judging by Charlie's work, no error is big enough to overlook.
        d. Judging by Charlie's work, all errors are big enough to overlook.

(13)   a. Judging by John's cooking style, no ingredient is too revolting to omit.
        b. Judging by John's cooking style, every ingredient is too revolting to omit.
        c. Judging by John's cooking style, no ingredient is tasty enough to omit.
        d. Judging by John's cooking style, every ingredient is tasty enough to omit.

(14)   a. For Hildy, no memento is too insignificant to discard.
        b. For Hildy, all mementos are too insignificant to discard.
        c. For Hildy, no memento is precious enough to discard.
        d. For Hildy, all mementos are precious enough to discard.

(15)   a. When Lulu is talking, no sentence is too complex to misunderstand.
        b. When Lulu is talking, every sentence is too complex to misunderstand.

c. When Lulu is talking, no sentence is simple enough to misunderstand.
d. When Lulu is talking, every sentence is simple enough to misunderstand.

(16)   a. Frank believes that no advice is too stupid to doubt.
        b. Frank believes that all advice is too stupid to doubt.
        c. Frank believes that no advice is sound enough to doubt.
        d. Frank believes that all advice is sound enough to doubt.

9.9   Experiment 8

Conditions:

        a. Negative nominal quantifier, negative verb, can
        b. Negative nominal quantifier, negative verb, should,
        c. Positive nominal quantifier, negative verb, can,
        d. Positive nominal quantifier, negative verb, should,
        e. Negative nominal quantifier, positive verb, can,
        f. Negative nominal quantifier, positive verb, should,
        g. Positive nominal quantifier, positive verb, can,
        h. Positive nominal quantifier, positive verb, should

(1)   a. In Maria's class, all tests can be failed.
        b. In Maria's class, all tests should be failed.
        c. In Maria's class, no tests can be failed.
        d. In Maria's class, no tests should be failed.
        e. In Maria's class, all tests can be passed.
        f. In Maria's class, all tests should be passed.
        g. In Maria's class, no tests can be passed.
        h. In Maria's class, no tests should be passed.

(2)   a. In Laura's opinion, no relationship can fail.
        b. In Laura's opinion, no relationship should fail.
        c. In Laura's opinion, all relationships can fail.
        d. In Laura's opinion, all relationships should fail.
        e. In Laura's opinion, no relationship can endure.
        f. In Laura's opinion, no relationship should endure.
        g. In Laura's opinion, all relationships can endure.
        h. In Laura's opinion, all relationships should endure.

(3)   a. According to Mark, no drug can be banned.
        b. According to Mark, no drug should be banned.
        c. According to Mark, all drugs can be banned.
        d. According to Mark, all drugs should be banned.
        e. According to Mark, no drug can be legalized.

f. According to Mark, no drug should be legalized.
g. According to Mark, all drugs can be legalized.
h. According to Mark, all drugs should be legalized.

(4)   a. Considering John's finances, no offer can be rejected.
       b. Considering John's finances, no offer should be rejected.
       c. Considering John's finances, all offers can be rejected.
       d. Considering John's finances, all offers should be rejected.
       e. Considering John's finances, no offer can be accepted.
       f. Considering John's finances, no offer should be accepted.
       g. Considering John's finances, all offers can be accepted.
       h. Considering John's finances, all offers should be accepted.

(5)   a. In Gideon's opinion, no habit can be discouraged.
       b. In Gideon's opinion, no habit should be discouraged.
       c. In Gideon's opinion, all habits can be discouraged.
       d. In Gideon's opinion, all habits should be discouraged.
       e. In Gideon's opinion, no habit can be encouraged.
       f. In Gideon's opinion, no habit should be encouraged.
       g. In Gideon's opinion, all habits can be encouraged.
       h. In Gideon's opinion, all habits should be encouraged.

(6)   a. According to the geologist, no data can be misinterpreted.
       b. According to the geologist, no data should be misinterpreted.
       c. According to the geologist, all data can be misinterpreted.
       d. According to the geologist, all data should be misinterpreted.
       e. According to the geologist, no data can be interpreted.
       f. According to the geologist, no data should be interpreted.
       g. According to the geologist, all data can be interpreted.
       h. According to the geologist, all data should be interpreted.

(7)   a. According to the politician, no social program can be opposed.
       b. According to the politician, no social program should be opposed.
       c. According to the politician, all social programs can be opposed.
       d. According to the politician, all social programs should be opposed.
       e. According to the politician, no social program can be supported.
       f. According to the politician, no social program should be supported.
       g. According to the politician, all social programs can be supported.
       h. According to the politician, all social programs should be supported.

(8)   a. During Kipp's presidential campaign, no rumor could have been denied.
       b. During Kipp's presidential campaign, no rumor should have been denied.
       c. During Kipp's presidential campaign, all rumors could have been denied.
       d. During Kipp's presidential campaign, all rumors should have been denied.
       e. During Kipp's presidential campaign, no rumor could have been confirmed.
       f. During Kipp's presidential campaign, no rumor should have been confirmed.

g. During Kipp's presidential campaign, all rumors could have been confirmed.
h. During Kipp's presidential campaign, all rumors should have been confirmed.

(9)   a. For someone like Alex, no memory can be forgotten.
b. For someone like Alex, no memory should be forgotten.
c. For someone like Alex, all memories can be forgotten.
d. For someone like Alex, all memories should be forgotten.
e. For someone like Alex, no memory can be recalled.
f. For someone like Alex, no memory should be recalled.
g. For someone like Alex, all memories can be recalled.
h. For someone like Alex, all memories should be recalled.

(10)   a. When it comes to Sharon's friends, no event can be skipped.
b. When it comes to Sharon's friends, no event should be skipped.
c. When it comes to Sharon's friends, all events can be skipped.
d. When it comes to Sharon's friends, all events should be skipped.
e. When it comes to Sharon's friends, no event can be attended.
f. When it comes to Sharon's friends, no event should be attended.
g. When it comes to Sharon's friends, all events can be attended.
h. When it comes to Sharon's friends, all events should be attended.

(11)   a. In Sarah's opinion, no player can lose.
b. In Sarah's opinion, no player should lose.
c. In Sarah's opinion, all players can lose.
d. In Sarah's opinion, all players should lose.
e. In Sarah's opinion, no player can lose.
f. In Sarah's opinion, no player should lose.
g. In Sarah's opinion, all players can lose.
h. In Sarah's opinion, all players should lose.

(12)   a. According to the movie producer, no criticism can be ignored.
b. According to the movie producer, no criticism should be ignored.
c. According to the movie producer, all criticism can be ignored.
d. According to the movie producer, all criticism should be ignored.
e. According to the movie producer, no criticism can be addressed.
f. According to the movie producer, no criticism should be addressed.
g. According to the movie producer, all criticism can be addressed.
h. According to the movie producer, all criticism should be addressed.

(13)   a. Judging by Charlie's work, no error can be overlooked.
b. Judging by Charlie's work, no error should be overlooked.
c. Judging by Charlie's work, all errors can be overlooked.
d. Judging by Charlie's work, all errors should be overlooked.
e. Judging by Charlie's work, no error can be caught.
f. Judging by Charlie's work, no error should be caught.
g. Judging by Charlie's work, all errors can be caught.
h. Judging by Charlie's work, all errors should be caught.

(14)   a. When Lulu is talking, no sentence can be misunderstood.
           b. When Lulu is talking, no sentence should be misunderstood.
           c. When Lulu is talking, all sentences can be misunderstood.
           d. When Lulu is talking, all sentences should be misunderstood.
           e. When Lulu is talking, no sentence can be understood.
           f. When Lulu is talking, no sentence should be understood.
           g. When Lulu is talking, all sentences can be understood.
           h. When Lulu is talking, all sentences should be understood.

(15)   a. For Hildy, no memento can be discarded.
           b. For Hildy, no memento should be discarded.
           c. For Hildy, all mementos can be discarded.
           d. For Hildy, all mementos should be discarded.
           e. For Hildy, no memento can be kept.
           f. For Hildy, no memento should be kept.
           g. For Hildy, all mementos can be kept.
           h. For Hildy, all mementos should be kept.

(16)   a. Frank believes that no advice can be doubted.
           b. Frank believes that no advice should be doubted.
           c. Frank believes that all advice can be doubted.
           d. Frank believes that all advice should be doubted.
           a. Frank believes that no advice can be trusted.
           b. Frank believes that no advice should be trusted.
           c. Frank believes that all advice can be trusted.
           d. Frank believes that all advice should be trusted.

(17)   a. Given Lily's relationship with her mother, no topic can be avoided.
           b. Given Lily's relationship with her mother, no topic should be avoided.
           c. Given Lily's relationship with her mother, all topics can be avoided.
           d. Given Lily's relationship with her mother, all topics should be avoided.
           e. Given Lily's relationship with her mother, no topic can be brought up.
           f. Given Lily's relationship with her mother, no topic should be brought up.
           g. Given Lily's relationship with her mother, all topics can be brought up.
           h. Given Lily's relationship with her mother, all topics should be brought up.

(18)   a. When Jack inspects the uniforms, no stain can be missed.
           b. When Jack inspects the uniforms, no stain should be missed.
           c. When Jack inspects the uniforms, all stains can be missed.
           d. When Jack inspects the uniforms, all stains should be missed.
           e. When Jack inspects the uniforms, no stain can be observed.
           f. When Jack inspects the uniforms, no stain should be observed.
           g. When Jack inspects the uniforms, all stains can be observed.
           h. When Jack inspects the uniforms, all stains should be observed.

(19)   a. With Liz's eating habits, no craving can be resisted.
           b. With Liz's eating habits, no craving should be resisted.

c. With Liz's eating habits, all cravings can be resisted.
d. With Liz's eating habits, all cravings should be resisted.
e. With Liz's eating habits, no craving can be indulged.
f. With Liz's eating habits, no craving should be indulged.
g. With Liz's eating habits, all cravings can be indulged.
h. With Liz's eating habits, all cravings should be indulged.

(20)   a. Judging by John's cooking style, no ingredient can be omitted.
        b. Judging by John's cooking style, no ingredient should be omitted.
        c. Judging by John's cooking style, all ingredients can be omitted.
        d. Judging by John's cooking style, all ingredients should be omitted.
        e. Judging by John's cooking style, no ingredient can be included.
        f. Judging by John's cooking style, no ingredient should be included.
        g. Judging by John's cooking style, all ingredients can be included.
        h. Judging by John's cooking style, all ingredients should be included.

9.10 Experiment 9

Conditions:

      a. internally anomalous
      b. internally consistent

(1)   a. In Maria's class, no test is too difficult to fail.
       b. In Maria's class, no test is too easy to fail.

(2)   a. In Laura's opinion, no relationship is too volatile to fail.
       b. In Laura's opinion, no relationship is too stable to fail.

(3)   a. According to Mark, no drug is too dangerous to ban.
       b. According to Mark, no drug is too safe to ban.

(4)   a. Gideon acts like no habit is too annoying to discourage.
       b. Gideon acts like no habit is too practical to discourage.

(5)   a. According to the politician, no social program is too wasteful to oppose.
       b. According to the politician, no social program is too efficient to oppose.

(6)   a. When it comes to Sharon's friends, no event is too inconvenient to skip.
       b. When it comes to Sharon's friends, no event is too convenient to skip.

(7)   a. In Sarah's opinion, no player is too slow to lose.
       b. In Sarah's opinion, no player is too fast to lose.

(8)   a. When Lulu is talking, no sentence is too complex to misunderstand.

b. When Lulu is talking, no sentence is too simple to misunderstand.

(9)   a. For Hildy, no memento is too insignificant to discard.
        b. For Hildy, no memento is too precious to discard.

(10)   a. Frank believes that no advice is too stupid to doubt.
        b. Frank believes that no advice is too sound to doubt.

9.11 Experiment 11

(1)   a. No pharmaceutical company knew that the drug was too dangerous to ban.
        b. The pharmaceutical company knew that no drug was too dangerous to ban.

(2)   a. No politician knew that the social program was too wasteful to oppose.
        b. The politician knew that no social program was too wasteful to oppose.

(3)   a. No expert explained that the childhood memory was too distant to forget.
        b. The expert explained that no childhood memory was too distant to forget.

(4)   a. No parent noticed that the extracurricular event was too inconvenient to skip.
        b. The parent noticed that no extracurricular event was too inconvenient to skip.

(5)   a. No coach was surprised that the runner was too slow to lose.
        b. The coach was surprised that no runner was too slow to lose.

(6)   a. No housecleaner knew that the memento was too insignificant to discard.
        b. The housecleaner knew that no memento was too insignificant to discard.

(7)   a. No job seeker realized that the advice was too silly to disregard.
        b. The job seeker realized that no advice was too silly to disregard.

(8)   a. No chef realized that the truffle was too expensive to omit.
        b. The chef realized that no truffle was too expensive to omit.

(9)   a. No manager was aware that the intern was too incompetent to fire.
        b. The manager was aware that no intern was too incompetent to fire.

(10)   a. No employee knew that the meeting was too unimportant to miss.
        b. The employee knew that no meeting was too unimportant to miss.

10 APPENDIX B: MIXED-EFFECTS MODELS

10.1 Experiment 1

10.1.1 Ratings

Random effects: (1 + Repeatable + Order | Item) + (1 + Illusion | Subject)

Acceptability	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	0.61	0.06	10.47			
Illusion (Y)	-0.87	0.08	-10.36	39.76	1	< .001
Repeatable (Y)	-0.09	0.06	-1.42	1.18	1	0.278
Order	0.02	0.03	0.55	0.91	1	0.339
Illusion:Repeatable	0.28	0.08	3.47	12.03	1	< .001
Illusion:Order	-0.08	0.04	-2.00	3.98	1	0.046

10.1.2 Reading times

Random effects: (1 | Item) + (1 | Subject)

RTs: *det*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	366.65	17.96	20.42			
Repeatable (Y)	27.25	12.73	2.14	4.58	1	0.032
Order	-28.73	6.38	-4.51	20.00	1	< .001

Random effects: (1 + Illusion * Repeatable | Item) + (1 + Order | Subject)

RTs: *noun*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	412.87	25.76	16.03			
Illusion (Y)	-2.67	15.37	-0.17	0.05	1	0.819
Repeatable (Y)	0.17	15.59	0.01	0.00	1	0.959
Order	-54.80	9.39	-5.84	19.69	1	< .001
Illusion:Repeatable	0.75	24.47	0.03	0.00	1	0.976
Illusion:Order	21.46	10.05	2.14	4.55	1	0.033

Random effects: (1 | Item) + (1 + Order | Subject)

RTs: *did*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	446.27	25.21	17.70			
Illusion (Y)	-4.39	16.49	-0.27	0.17	1	0.676
Repeatable (Y)	-15.87	16.48	-0.96	0.33	1	0.565
Order	-21.41	9.89	-2.17	11.09	1	< .001
Illusion:Repeatable	18.33	23.28	0.79	0.62	1	0.430
Illusion:Order	-16.02	11.67	-1.37	1.90	1	0.168

Random effects: (1 + Repeatable | Item) + (1 + Order | Subject)

RTs: *did* + 1	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	469.04	34.11	13.75			
Illusion (Y)	15.47	20.96	0.74	1.87	1	0.172
Repeatable (Y)	-4.61	24.30	-0.19	0.00	1	0.979
Order	-11.30	12.63	-0.90	13.77	1	< .001
Illusion:Repeatable	10.04	29.70	0.34	0.11	1	0.739
Illusion:Order	-46.53	14.87	-3.13	5.26	1	0.022

Random effects: (1 + Illusion + Order | Item) + (1 + Illusion + Repeatable + Order | Subject)

RTs: *did* + 2	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	408.90	25.93	15.77			
Illusion (Y)	63.47	22.55	2.81	11.10	1	< .001
Repeatable (Y)	9.41	19.37	0.49	0.51	1	0.477
Order	-45.09	11.88	-3.79	18.91	1	< .001
Illusion:Repeatable	0.80	26.77	0.03	0.00	1	0.976
Illusion:Order	-11.73	13.41	-0.88	0.76	1	0.382

Random effects: (1 + Illusion + Order | Item) + (1 + Order * Illusion | Subject)

RTs: *did* + 3	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	386.80	19.85	19.49			
Illusion (Y)	38.11	15.81	2.41	2.13	1	0.144
Repeatable (Y)	18.07	13.73	1.32	0.04	1	0.838
Order	-42.13	10.54	-4.00	16.57	1	< .001
Illusion:Repeatable	-40.02	19.41	-2.06	4.25	1	0.039
Illusion:Order	4.38	12.94	0.34	0.12	1	0.730

Random effects: (1 | Item) + (1 + Order | Subject)

RTs: *did* + 4	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	398.57	21.83	18.26			
Illusion (Y)	21.15	13.74	1.54	3.18	1	0.074
Repeatable (Y)	-6.39	13.73	-0.47	1.11	1	0.292
Order	-37.36	7.42	-5.03	25.26	1	< .001
Illusion:Repeatable	-7.67	19.43	-0.40	0.16	1	0.692
Illusion:Order	5.82	9.73	0.60	0.36	1	0.549

Random effects: (1 + Illusion + Order | Item) + (1 + Order | Subject)

RTs: *did* + 5	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	396.21	20.41	19.42			
Illusion (Y)	0.58	13.86	0.04	2.72	1	0.099
Repeatable (Y)	-6.11	12.93	-0.47	0.04	1	0.845
Order	-22.41	8.66	-2.59	16.70	1	< .001
Illusion:Repeatable	18.46	18.27	1.01	4.15	1	0.042
Illusion:Order	-8.72	9.15	-0.95	0.20	1	0.654

Random effects: (1 + Order | Item) + (1 + Order | Subject)

RTs: *did* + 6	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	373.27	17.23	21.66			
Illusion (Y)	4.59	9.26	0.50	3.84	1	0.050
Repeatable (Y)	-16.33	9.26	-1.76	1.53	1	0.216
Order	-26.29	7.08	-3.71	12.97	1	< .001
Illusion:Repeatable	16.46	13.08	1.26	1.59	1	0.208
Illusion:Order	0.66	6.56	0.10	0.01	1	0.920

Random effects: (1 + Order | Item) + (1 + Order | Subject)

RTs: *did* + 7	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	380.97	17.95	21.23			
Illusion (Y)	2.83	9.37	0.30	0.00	1	0.969
Repeatable (Y)	-2.37	9.39	-0.25	0.68	1	0.408
Order	-30.11	6.92	-4.35	17.66	1	< .001
Illusion:Repeatable	-6.21	13.27	-0.47	0.22	1	0.639
Illusion:Order	0.10	6.64	0.02	0.00	1	0.989

10.2 Experiment 2

10.2.1 Ratings

Random effects: (1 + Repeatable + Illusion * Order | Item) + (1 + Order | Subject)

Acceptability	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	0.45	0.07	6.81			
Illusion (Y)	-0.22	0.07	-2.97	7.95	1	0.005
Repeatable (Y)	-0.07	0.08	-0.94	0.29	1	0.589
Order	0.04	0.04	1.18	2.61	1	0.106
Illusion:Repeatable	0.08	0.07	1.03	1.05	1	0.305
Illusion:Order	0.02	0.04	0.46	0.21	1	0.649

10.2.2 Reading times

Random effects: (1 | Item) + (1 + Order | Subject)

RTs: det	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	440.92	27.19	16.22			
Repeatable (Y)	3.98	13.20	0.30	0.09	1	0.763
* Order	-27.85	10.61	-2.63	6.50	1	0.011

Random effects: (1 + Order | Item) + (1 + Order + Illusion * Repeatable | Subject)

RTs: *noun*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	497.98	36.88	13.50			
Illusion (Y)	-15.58	15.87	-0.98	1.52	1	0.218
Repeatable (Y)	-5.58	15.04	-0.37	0.06	1	0.806
*** Order	-50.14	10.70	-4.69	16.70	1	< .001
Illusion:Repeatable	6.41	22.88	0.28	0.08	1	0.777
Illusion:Order	10.09	9.53	1.06	1.12	1	0.289

Random effects: (1 | Item) + (1 + Order | Subject)

RTs: *did*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	494.49	33.05	14.96			
Illusion (Y)	-3.72	16.41	-0.23	0.79	1	0.373
Repeatable (Y)	1.86	16.53	0.11	1.94	1	0.164
** Order	-21.62	12.55	-1.72	9.76	1	0.002
Illusion:Repeatable	28.56	23.31	1.23	1.51	1	0.220
* Illusion:Order	-29.46	11.69	-2.52	6.36	1	0.012

Random effects: (1 | Item) + (1 + Illusion * Repeatable + Order | Subject)

RTs: *did+1*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	514.48	37.24	13.81			
Illusion (Y)	13.75	21.00	0.66	1.80	1	0.179
Repeatable (Y)	-17.62	20.38	-0.86	0.67	1	0.413
** Order	-19.02	13.88	-1.37	6.92	1	0.009
Illusion:Repeatable	12.68	29.47	0.43	0.19	1	0.665
* Illusion:Order	-29.31	12.62	-2.32	5.40	1	0.020

Random effects: (1 + Illusion | Item) + (1 + Repeatable + Illusion * Order | Subject)

RTs: *did+2*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	449.67	25.48	17.65			
Illusion (Y)	17.60	17.84	0.99	0.67	1	0.412
Repeatable (Y)	17.28	13.80	1.25	1.74	1	0.187
*** Order	-43.89	11.15	-3.94	18.17	1	< .001
Illusion:Repeatable	-7.36	18.40	-0.40	0.16	1	0.689
Illusion:Order	-5.47	11.24	-0.49	0.24	1	0.624

Random effects: (1 | Item) + (1 + Order | Subject)

RTs: *did+3*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	435.94	24.30	17.94			
Illusion (Y)	9.06	12.87	0.70	2.09	1	0.148
Repeatable (Y)	0.95	12.89	0.07	0.31	1	0.580
*** Order	-29.14	9.29	-3.14	10.87	1	< .001
Illusion:Repeatable	8.13	18.20	0.45	0.20	1	0.655
Illusion:Order	1.74	9.13	0.19	0.04	1	0.848

Random effects: (1 | Item) + (1 + Order | Subject)

RTs: *did+4*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	439.32	25.92	16.95			
Illusion (Y)	2.78	12.18	0.23	0.11	1	0.740
Repeatable (Y)	12.64	12.16	1.04	0.66	1	0.417
** Order	-33.43	9.94	-3.36	9.98	1	0.002
Illusion:Repeatable	-11.36	17.21	-0.66	0.44	1	0.509
Illusion:Order	7.14	8.64	0.83	0.69	1	0.407

Random effects: (1 + centeredOrder | Item) + (1 + Illusion * centeredOrder | Subject)

RTs: *did+5*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	436.72	24.82	17.60			
Illusion (Y)	24.66	12.37	1.99	2.39	1	0.122
Repeatable (Y)	1.21	11.13	0.11	1.77	1	0.184
** Order	-28.63	10.45	-2.74	9.63	1	0.002
Illusion:Repeatable	-22.96	15.76	-1.46	2.10	1	0.147
Illusion:Order	-1.01	11.22	-0.09	0.02	1	0.899

Random effects: (1 | Item) + (1 + Order | Subject)

RTs: *did+6*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	432.65	23.48	18.43			
Illusion (Y)	0.15	10.57	0.01	0.19	1	0.665
Repeatable (Y)	0.27	10.55	0.03	0.21	1	0.649
*** Order	-26.63	9.24	-2.88	11.56	1	< .001
Illusion:Repeatable	6.26	14.95	0.42	0.18	1	0.674
Illusion:Order	-7.92	7.49	-1.06	1.12	1	0.290

Random effects: (1 | Item) + (1 + Order | Subject)

RTs: *did+7*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	441.72	23.85	18.52			
Illusion (Y)	0.05	10.20	0.01	0.37	1	0.540
Repeatable (Y)	-12.15	10.22	-1.19	1.15	1	0.283
** Order	-20.09	8.54	-2.35	8.71	1	0.003
Illusion:Repeatable	8.82	14.44	0.61	0.38	1	0.540
Illusion:Order	-7.41	7.24	-1.02	1.05	1	0.305

10.3 Pooled reaction times, Experiments 1 & 2

Random effects: (1 + Illusion * Order | Item) + (1 + Illusion * Rating + Order | Subject) + (1 | Experiment)

RTs: *did*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	465.63	20.31	22.93			
Illusion (Y)	1.97	10.99	0.18	0.10	1	0.753
Order	-34.51	10.87	-3.18	16.99	1	< .001
Rating	-11.30	10.56	-1.07	1.27	1	0.259
Illusion:Order	-18.25	8.74	-2.09	3.95	1	0.047
Illusion:Rating	5.02	14.05	0.36	0.12	1	0.729

Random effects: (1 | Item) + (1 + Order | Subject) + (1 | Experiment)

RTs: *did+1*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	485.16	22.80	21.28			
Illusion (Y)	15.04	10.76	1.40	2.60	1	0.107
Order	-25.03	11.91	-2.10	13.73	1	< .001
Rating	-22.63	11.20	-2.02	1.26	1	0.261
Illusion:Order	-34.68	10.23	-3.39	11.50	1	< .001
Illusion:Rating	24.58	14.50	1.70	2.88	1	0.090

Random effects: (1 + Illusion * Order | Item) + (1 + Order * Rating | Subject) + (1 + Illusion | Experiment)

RTs: *did+2*	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	437.18	22.29	19.61			
Illusion (Y)	28.20	23.02	1.23	1.54	1	0.215
Order	-54.08	8.00	-6.76	36.21	1	< .001
Rating	-30.97	10.00	-3.10	2.87	1	0.090
Illusion:Order	-6.46	13.60	-0.48	1.33	1	0.248
Illusion:Rating	32.87	12.42	2.65	6.76	1	0.009

10.4 Experiment 3

Random effects: (1 + Illusion * Plurality | Item) + (1 + Illusion * Plurality | Subject)

Acceptability	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	0.75	0.08	9.18			
Illusion (Y)	-1.11	0.15	-7.25	27.46	1	< .001
Plurality (Semantic)	-0.09	0.09	-0.98	1.91	1	0.166
Illusion:Plurality	0.43	0.16	2.77	6.86	1	0.009

10.5 Experiment 4

Random effects: (1 + Illusion * Plurality | Item) + (1 + Illusion * Plurality | Subject)

Acceptability	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	0.67	0.10	6.70			
Illusion (Y)	-0.95	0.14	-6.69	24.31	1	< .001
Plurality (Semantic)	-0.08	0.07	-1.03	0.28	1	0.600
Illusion:Plurality	0.27	0.11	2.40	5.32	1	0.021

10.6 Experiment 5

Random effects: (1 + Than.clause * Direct.object | Item) + (1 + Than.clause * Direct.object | Subject)

Acceptability	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	0.49	0.10	4.69			
Than-clause (Singular)	-0.64	0.13	-4.93	30.94	1	< .001
Direct object (Singular)	-0.02	0.09	-0.22	5.41	1	0.020
Than-clause:Direct object	-0.26	0.14	-1.87	3.36	1	0.067

10.7 Experiment 6

Random effects: (1 + Illusion | Item) + (1 + Illusion | Subject)

Acceptability	Estimate	Std. Error	t-value	X^2	df	p-value
(Intercept)	0.69	0.08	9.01			
Illusion (Y)	-0.29	0.13	-2.17	22.36	1	< .001
Noun (Singular)	-0.08	0.06	-1.26	19.66	1	< .001
Gradable (N)	-0.02	0.07	-0.28	2.46	1	0.117
Illusion:Noun	-0.24	0.09	-2.71	7.30	1	0.007
Illusion:Gradable	-0.47	0.12	-3.85	11.49	1	< .001

10.8 Experiment 7a

Random effects: (1 + DQ + Verb | Item) + (1 + DQ * Verb | Subject)

Accuracy (Y)	Estimate	Std. Error	z-value	p-value	X^2	df	p-value
(Intercept)	2.47	0.77	3.20	0.001			
DQ (too)	-3.85	0.79	-4.90	< .001	12.97	1	< .001
Verb (positive)	-0.93	0.70	-1.33	0.183	7.68	1	0.006
DQ:Verb	3.05	0.84	3.64	< .001	14.67	1	< .001

10.9 Experiment 7b

Random effects: (1 + DQ + Determiner | Item) + (1 + DQ + Determiner | Subject)

Accuracy (Y)	Estimate	Std. Error	z-value	p-value	X^2	df	p-value
(Intercept)	3.31	0.79	4.20	< .001			
DQ (too)	-1.25	0.67	-1.87	0.062	23.83	1	< .001
Determiner (no)	-1.69	0.65	-2.61	0.009	25.01	1	< .001
DQ:Determiner	-1.59	0.70	-2.27	0.023	4.90	1	0.027

10.10 Experiment 8

Random effects: (1 | Item) + (1 + Veridical + Inverted DQ | Subject)

Accuracy (Y)	Estimate	Std. Error	z-value	p-value	X^2	df	p-value
(Intercept)	-1.81	0.61	-2.98	0.003			
Veridical	0.51	1.03	0.50	0.617	0.19	1	0.661
Inverted DQ	-2.57	1.27	-2.03	0.043	10.11	1	0.001
Inverted verb	0.92	0.88	1.04	0.298	1.26	1	0.262

Random effects: (1 | Item) + (1 + Veridical + Inverted DQ | Subject)

Percept (Veridical)	Estimate	Std. Error	z-value	p-value	X^2	df	p-value
(Intercept)	-10.60	2.32	-4.58	< .001			
Inverted DQ	-1.89	3.20	-0.59	0.555	0.36	1	0.551

10.11 Experiment 9

Random effects: (1 + internal | Item) + (1 + internal | Subject)

Accuracy (Y)	Estimate	Std. Error	z-value	p-value	X^2	df	p-value
(Intercept)	-1.49	0.61	-2.45	0.014			
Internal (Non-anomaly)	2.87	0.94	3.06	0.002	8.85	1	0.003

Random effects: (1 + internal | Item) + (1 + internal | Subject)

Percept (Veridical)	Estimate	Std. Error	z-value	p-value	X^2	df	p-value
(Intercept)	-2.08	0.61	-3.39	< .001			
Internal (Non-anomaly)	3.78	0.89	4.23	< .001	16.9 1	1	< .001

10.12 Experiment 10

Random effects: (1 + Verb | Item) + (1 | Subject)

Accuracy (Y)	Estimate	Std. Error	z-value	p-value	X^2	df	p-value
(Intercept)	1.41	0.27	5.18	< .001			
Verb (positive)	-0.05	0.32	-0.16	0.874	0.00	1	0.966
DQ (too)	-0.64	0.27	-2.34	0.019	97.94	1	< .001
Task (comprehension)	0.20	0.39	0.52	0.605	0.48	1	0.488
Verb:DQ	-1.09	0.39	-2.83	0.005	0.07	1	0.786
Verb:Task	-0.05	0.45	-0.11	0.909	29.38	1	< .001
DQ:Task	-2.10	0.45	-4.65	< .001	3.63	1	0.057
Verb:DQ:Task	2.95	0.62	4.78	< .001	23.23	1	< .001

Random effects: (1 | Item) + (1 + Similarity + Confidence | Subject)

Accuracy (Y)	Estimate	Std. Error	z-value	p-value	X^2	df	p-value
(Intercept)	-2.03	0.53	-3.80	< .001			
Similarity	-0.84	0.47	-1.78	0.075	0.31	1	0.580
Confidence	2.01	1.12	1.80	0.072	1.47	1	0.226
Similarity:Confidence	2.36	1.10	2.16	0.031	5.04	1	0.025

10.13 Experiment 11

Random effects: (1 + Licensing | Item) + (1 + Licensing | Subject)

Accuracy (Y)	Estimate	Std. Error	z-value	p-value	X^2	df	p-value
(Intercept)	0.67	0.49	1.36	0.174			
Licensing (non-licensing)	-1.80	0.41	-4.39	< .001	12.76	1	< .001