Language Processing at Its Trickiest: Grammatical Illusions and Heuristics of Judgment

Evelina Leivada [ID]

Department of English and German Studies, Universitat Rovira i Virgili, 43002 Tarragona, Spain;
evelina.leivada@urv.cat

Abstract: Humans are intuitively good at providing judgments about what forms part of their native language and what does not. Although such judgments are robust, consistent, and reliable, human cognition is demonstrably fallible to illusions of various types. Language is no exception. In the linguistic domain, several types of sentences have been shown to trick the parser into giving them a high acceptability judgment despite their ill-formedness. One example is the so-called comparative illusion ('More people have been to Tromsø than I have'). To this day, comparative illusions have been tested mainly with monolingual, neurotypical speakers of English. The present research aims to broaden our understanding of this phenomenon by putting it to test in two populations that differ in one crucial factor: the number of languages they speak. A timed acceptability judgment task was administered to monolingual speakers of Standard Greek and bi(dia)lectal speakers of Standard and Cypriot Greek. The results are not fully in line with any of the semantic re-analyses proposed for the illusion so far, hence a new proposal is offered about what interpretation induces the illusion, appreciating the influence of both grammatical processing and cognitive heuristics. Second, the results reveal an effect of developmental trajectory. This effect may be linked to an enhanced ability to spot the illusion in bi(dia)lectals, but several factors can be identified as possible culprits behind this result. After discussing each of them, it is argued that having two grammars may facilitate the setting of a higher processing threshold, something that would entail decreased fallibility to grammatical illusions.

Keywords: bilectalism; grammatical illusions; acceptability judgments; reaction times; parsing

1. Introduction

One of the key characteristics of the computational brain is its ability to encode, retrieve, and communicate information, while learning from experience the expected value range of certain variables (e.g., the distance between two places) and continuously updating memory resources accordingly (Gallistel and King 2009). Another important characteristic is its tendency toward radical contextualization which, among other things, entails a propensity to (i) adhere to Grice's four maxims of co-operation,[1] even when the communicative context lacks the conversational features that call for the application of these maxims, and (ii) contextualize incoming stimuli by drawing on those memory resources that are most readily accessible (Kahneman and Tversky 1982; Hilton 1995; Stanovich and

---

[1] According to these maxims, one's conversational contribution should be as informative as necessary for a successful exchange (Grice 1957): (i) the maxim of quantity suggests that the interlocutors be as informative as possible, providing the necessary information for a successful exchange, and no more/less; (ii) the maxim of quality suggests that the interlocutors do not provide information they believe to be false; (iii) the maxim of relation posits that every contribution should be relevant and pertinent to the discussion; and (iv) the maxim of manner asks for clarity, brevity, and absence of ambiguity in the conversational contributions.

West 2000). To illustrate how these tendencies may lead to judgmental errors, Kahneman and Tversky (1982) introduced the Linda problem (1).

(1)   Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Asking for judgments of probability, Kahneman and Tversky (1982) presented participants with different options: A. Linda is a bank teller, and B. Linda is a bank teller and is active in the feminist movement. The results showed that people perceived Linda as a very good fit for a feminist activist. Yet, there is an often unnoticed logical relation between A and B: the set of feminist bank tellers is included in the set of bank tellers; thus, the probability of B cannot be higher than the probability of A (see (Kahneman 2011, pp. 156–65) for a detailed discussion of the Linda problem). This error in judgment boils down to the human tendency toward contextualization, and more specifically, to an application of Grice's maxim of relation: (1) involves information about Linda that participants take to be *relevant*—even though they do not encounter this information in a setting that calls for application of conversational maxims, because they are biased towards contextualizing information this way. As Adler (1984, p. 175) puts it, "[a]ssuming that Linda is a feminist provides the only explanatory path to the contents of the sketch of her".

The appeal to such tendencies and processing heuristics has been explained in terms of the brain employing a dual-process theory of reasoning: System 1 is automatic, largely unconscious, and operates in a computationally undemanding way, while System 2 is responsible for effortful computations that are typical of analytic intelligence (Stanovich and West 2000 and references therein). Giving the incorrect response to the Linda problem is the result of a failure of System 2 to intervene, check, and override the wrong response provided automatically by System 1. It is the latter that employs judgment heuristics and biases of various types (Kahneman 2000).

The Linda problem brings forward the question of whether naïve participants are good intuitive statisticians. While different answers to this question exist in the literature (cf. Cosmides and Tooby 1996; Anderson 1998), it seems that certain domains of cognition have been linked more heavily than others to intuitive answers that demonstrate biased or faulty reasoning. Put differently, although the appeal to processing heuristics is a general property of cognition—and as such, relevant to all cognitive domains—certain domains have been associated to more robust judgments than others. Language is one such domain.

Intuitive judgments about what forms part of one's linguistic repertoire are highly reliable and consistent (Sprouse and Almeida 2012), and this is important because such a degree of accuracy and stability does not characterize all types of judgments that are related to some aspect of human perception. For example, in the viral 'The Dress' photograph, judgments of color perception were found to differ both across participants (i.e., the dress was seen as blue/black, blue/brown or white/gold) and across testing sessions (Lafer-Sousa et al. 2015). To continue the analogy with statistics, naïve participants may not always be intuitively good statisticians, but there is little debate over the fact that they are intuitively good linguists. This exceptionality of language was grasped even in the early work of Amos Tversky and Daniel Kahneman on judgment and decision making in the 1960s:

"Amos told the class about an ongoing program of research at the University of Michigan that sought to answer this question: Are people good intuitive statisticians? We already knew that people are good intuitive grammarians: at age four a child effortlessly conforms to the rules of grammar as she speaks, although she has no idea that such rules exist. Do people have a similar intuitive feel for the basic principles of statistics? Amos reported that the answer was a qualified yes. We had a lively debate in the seminar and ultimately concluded that a qualified no was a better answer".

(Kahneman 2011, p. 5)

Although acceptability judgments in linguistic tasks are robust and reliable, human cognition is fallible to illusions of all types. In the domain of language processing, several types of illusions have been shown to trick the cognitive parser into giving them a high acceptability judgment despite the fact that they are grammatically ill-formed (see Phillips et al. 2011 for discussion of various examples). One of the most powerful grammatical illusions is the so-called comparative illusion (2).

(2)   More people have been to Russia than I have. (Montalbetti 1984)

Native speakers of English give sentences like (2) a mean rating of 3.7 (Wellwood et al. 2018) or 4.3–4.63 (O'Connor 2015) on a 7-point Likert scale where 7 is the maximum rating of well-formedness. However, when speakers that gave (2) a high acceptability rating are asked to provide its meaning, they often stop perplexed. This happens because the well-formedness of (2) is an *illusion*. The meaning of (2) cannot be that the number of the people that have been to Russia exceeds the number of me, similar to how 'More people have been to Tromsø than penguins have' means that the number of people that have been to Tromsø exceeds the number of penguins that have been to Tromsø. In linguistic terms, (2) is ill-formed because the main clause subject calls for a comparison of cardinalities of sets of individuals, but in the absence of a bare plural in the embedded clause subject, there is no appropriate constituent to host the degree variable after Verb Phrase ellipsis is resolved (Phillips et al. 2011; O'Connor 2015). Pronouns, proper names, and definite descriptions are not appropriate hosts, only a bare plural would be a legit correlate in the embedded clause (Bresnan 1973; De Dios-Flores 2016; Wellwood et al. 2018), as (3) shows.

(3)   How many penguins/*I/*Mary/*the penguin/*the penguins have been to Tromsø? (adapted from Wellwood et al. 2018, p. 544).

Why is (2) acceptable though? It was argued earlier that naïve participants stop perplexed when asked to spell out the meaning they have assigned to the sentence upon accepting it as well-formed. Although this is sometimes the case, not everybody is unable to provide a meaning. Some people do it effortlessly by offering one of the following two interpretations (4a–b).

(4a)   More people than just me have been to Russia.
(4b)   People have been to Russia more times than I have.

A long line of literature (Wellwood et al. 2009, 2018; O'Connor et al. 2012; O'Connor 2015; De Dios-Flores 2016) has linked (4a) and (4b) with different semantic re-analyses: set comparison or additive interpretation for (4a) and event comparison for (4b). Starting off from the meaning of the former, according to Wellwood et al. (2018), the set comparison re-analysis has been related to an interpretation of the comparative *more* as the additive *more* (5):

(5)   Helen ate four ice creams last weekend.

    a.   Comparative: Today she ate *more* (= a new set of ice creams that is bigger than four, e.g., five, six, etc.).

    b.   Additive: Today she ate *more* (= at least one more, but not necessarily five more).

Both readings of (5) involve some kind of comparison of sets, which is why I use the term "set comparison re-analysis" for them. More specifically, in (5a) the set of ice creams that Helen ate last weekend is compared to the set of ice creams she ate today, while in (5b) the set of ice creams that Helen ate last weekend is compared to a proper superset that involves both the ice creams of last weekend and the one(s) she ate today. These are the possible interpretations behind (4a).

The meaning of (4b) is different, because here we are dealing with a comparison of events, not a comparison of sets. Thus, (4b) can be read as saying something about the number of times that Russia has been visited, similar to how (6b) says something about the times the event of walking through the turnstile occurred.

(6)   12,000 persons walked through the turnstile yesterday (Krifka 1990, p. 488).

     a.   Object-related reading: 12,000 different persons.

     b.   Event comparison reading: 12,000 events of walking (but the persons may be fewer than 12,000).

To sum up, according to the literature, (2) can be interpreted by naïve participants as denoting a set comparison (i.e., along the lines of (5)) or an event comparison (as in (6b)).

I propose a third possibility. Upon judging (2) as highly acceptable, naïve participants assign a meaning to it that is related to some aspect of comparison, without necessarily settling in an unambiguous way on either the set comparison or the event comparison interpretation. Under this scenario, participants may not be ready to put their finger on a single, full-fledged, and fully coherent interpretation, but they have parsed the sentence, such that their parser has reached a potentially ambiguous and vague interpretation. Let us call this the "shallow comparison re-analysis".

Crucially, the three re-analyses make very different empirical predictions, and these predictions will be relevant for the design of the present experiment. The set comparison scenario presupposes that the subject of the embedded clause can be included in the denotation of the matrix subject (Wellwood et al. 2018). For example, in (2), the pronoun *I* denotes an entity that belongs to the set of *people*. In (7), however, this relation does not hold. Under the predictions of the set comparison scenario, (7) will not induce the illusion under this re-analysis.

(7)   More boys have visited Tromsø than she has.

In addition, if *more* is substituted with *fewer*, the additive reading of the set comparison is not possible (Wellwood et al. 2018). The prediction is that (8) will not induce the illusion under this reading.

(8)   Fewer boys have visited Tromsø than he has.

The event comparison re-analysis makes a different prediction. The comparison of events in (6b) entails that if the persons are fewer than 12,000, each event of walking through the turnstile must be carried out multiple times by the same person. In other words, the event comparison scenario entails that the action denoted by the predicate must be repeatedly carried out by the same entity (Krifka 1990). In this sense, the repeatability of the predicate is crucial in obtaining the illusion under this meaning. Further, (2) features a repeatable predicate (i.e., one can visit Russia as many times as one likes), but (9) does not (i.e., people graduate high school once). The prediction thus is that (9) will not induce the illusion under the event comparison interpretation.

(9)   More people have graduated high school than I have.

Having presented the different interpretations that may justify a high acceptability rating, the second important question to be asked is what exactly makes the illusion go unnoticed. Why is the parser tricked in constructing an interpretation for (2) instead of noticing outright its incoherence? Again, different explanations exist in the literature. One of them is that (2) shows a blend of two perfectly coherent—at a local level—templates (Townsend and Bever 2001). Comparison of events is very frequent in language and so is comparison of individuals. By blending the two templates in one sentence, the parser sees two chunks that it frequently encounters, without any *marked* deviation from the patterns it anticipates (i.e., what was earlier described as the brain's tendency to learn from experience and update the expected range of values accordingly). Put another way, by the time the deviation kicks in when the embedded clause subject is reached in (2), the parser is already into the process of constructing a meaning and actively anticipates a range of values. The pronoun 'I' is sufficiently close to this expected range of values for the parser to be tricked into not noticing the deviation. Had the illusion in (2) been completed as in (10), the parser would immediately notice the marked distortion, the process of assigning meaning would have failed, the sentence would not have been accepted as well-formed, and no illusion would arise.

(10)　More people have visited Tromsø than *beautifully.

Another view is presented in Phillips et al. (2011), who suggest that the experimental results that show sensitivity to the semantics of the predicate (i.e., its repeatability) suggest that Townsend and Bever's explanation is not fully satisfactory in the following sense: The illusions seem robust only with repeatable predicates. Given that the blend of locally coherent templates is there regardless of the semantics of the predicate, what can account for this difference in robustness? O'Connor (2015) summarizes these two different views in the following way: "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 and Bever (2001) posits no grammatical analysis at all at the time the illusion is first judged acceptable" (p. 74). It should be noted, however, that these two explanations do not really answer the same question. The emphasis of the event comparison explanation is on *how* the meaning is construed (possible answer: by resorting to a grammar-based interpretation that facilitates the event comparison re-analysis) and not on *why* the parser fails to spot the illusion *in the first place* (possible answer: because it operates by making use of processing heuristics).

Last, a third explanation that integrates assumptions from both previous ones suggests that syntactic parsing occurs both algorithmically and heuristically such that processing is shallow, partial, or simply "good enough" (Ferreira et al. 2002; Ferreira 2003; Ferreira and Patson 2007). Under this account, grammatical illusions are the outcome of a partial-match strategy that is operative during processing (Reder and Kusbit 1991; Kamas et al. 1996). Upon receiving a linguistic stimulus, its parts are matched to stored knowledge, so that an output is produced. However, the parser matches the stimulus to stored information only up to a point. In other words, a processing threshold is set, and the stimulus is checked up to this threshold, hence the notion of partial matching. The notion of processing threshold can be understood as the tipping point of the process of mental engaging with the received stimulus, that suffices to elicit an interpretation that the parser deems good enough.

Do all people set the threshold at the same level? There is no answer to this question, because to this day, comparative illusions have been tested mainly with monolingual, neurotypical speakers of English. It was found that English speakers spot the comparative illusion to varying degrees (Wellwood et al. 2018 and references therein). The cross-linguistic robustness of the illusion has been experimentally confirmed with speakers of Danish, and anecdotally confirmed with speakers of Swedish, Faroese, German, Icelandic, Polish, and Swedish (Christensen 2016). Given that some variation is observed even among people that speak the same language and fit the same behavioral profile, one expects that people with different profiles will show even greater differences. In other words, it is possible that people with pronounced differences in cognitive aptitude—the latter being defined as the combined outcome of mental abilities such as speed of computation, attention orienting, task-switching efficiency, etc. (Kahneman 2011; Stanovich 2011)—will perform differently in tasks that involve demanding processing over tricky patterns of superficial well-formedness. This prediction is grounded on an observation made by Kahneman in the context of a different stimulus (11), albeit one that tricks the parser too.

(11)　A chocolate bar and a candy cost €1.10. The chocolate bar costs €1 more than the candy. How much does the candy cost? (adapted from Kahneman 2011, p. 44)

Presenting this as a tricky stimulus already provides a very strong cue that the obvious answer to the question is wrong. Despite the salience of this cue, an 'obvious', intuitive answer did come to your mind, and this was 10 cents. If this intuitive answer is suppressed and checked for accuracy, the correct answer is easy to find; the correct answer is 5 cents. How often is the intuitive answer suppressed though? Kahneman used puzzles like (11) to show how our 'lazy' System 2 is not always checking intuitive answers. In his words, "the results are shocking. More than 50% of students at Harvard, MIT, and Princeton, gave the intuitive—incorrect—answer. At less selective universities, the rate of

demonstrable failure to check was in excess of 80%" ([Kahneman 2011](#), p. 45). The observation that students from universities that differ in their selectivity perform differently in the chocolate-and-candy problem suggests that it is indeed possible that different people are fallible to illusions to different degrees. The reasons for this are largely unknown: the difference could be plausibly attributed both to variable socio-economic background, to enhanced cognitive abilities, or to both, with an unknown nature-to-nurture ratio. Previous research on comparative illusions has not addressed the topic from this perspective and the factors that may lead to robust differences—if any—are also unknown.

In this context, the present research aims to broaden our knowledge about language processing in comparative illusions, by putting this phenomenon to test through comparing two populations that differ in one crucial factor: the number of languages they use on a daily basis. The aim is to investigate all the factors that, according to the literature, play a role in inducing the illusion (i.e., the repeatability of the predicate, the ability of the embedded clause subject to be included in the matrix clause subject, the role of *fewer*) plus an additional factor: the impact of having a different developmental trajectory on the processing of grammatical illusions.

In a nutshell, the questions that the present research aims to address are the following:

Research question 1: Are the factors that have been linked to inducing the illusion at its strongest equally robust in languages other than English?

Research question 2: Are people with different developmental trajectories fallible to comparative illusions to different degrees?

2. The Present Experiment

2.1. Setting the Context: Brief Summary of Previous Experiments and Their Results

Experiments on comparative illusions have so far supported mainly the event comparison scenario. [O'Connor et al.](#) ([2012](#)) tested 24 university students in a self-paced reading, rating, and recall study followed by an acceptability judgment task (AJT) that asked participants to rate the stimuli on a 1–7 Likert scale (see also [O'Connor 2015](#)). They found that illusions were judged as significantly less acceptable than control sentences and that illusions with non-repeatable predicates were rated lower than those with repeatable predicates. At the same time, the ratings suggest that the difference in judgments across illusions with different types of predicates, although moderately significant, was not big: 4.30 vs. 4.63 for illusions with non-repeatable and repeatable predicates, respectively. This indicates that comparative illusions are rated fairly similarly regardless of whether the semantics of the predicate preclude the event comparison re-analysis.

[Wellwood et al.](#) ([2018](#)) described their results from different tasks as providing "support only for the event comparison hypothesis" (p. 557). In their first experiment, they tested 64 Amazon mechanical Turk workers, tapping into all the factors that may influence the re-analysis of the illusion. It was found that the use of *fewer* in the illusion condition did not result in a statistically significant decrease in acceptability ratings, compared to the use of *more*, as should have been the case under the predictions of the additive *more* reading: if participants legitimized the illusion through the additive interpretation, replacing *more* with *fewer* should lead to a substantial increase in spotting the illusion, and consequently, a substantial decrease in judgments of well-formedness. The crucial result of the first experiment reported in [Wellwood et al.](#) ([2018](#)) shows a strong effect of repeatability for all conditions (i.e., both illusions and controls together). However, when one focuses exclusively on the illusion condition, which is the critical one, the difference between repeatable and non-repeatable predicates is only weakly significant. Similar to what was observed above for [O'Connor et al.](#) ([2012](#)), if this result is taken to support *only* the event comparison hypothesis, which is what the authors suggest, it is hard to see why the illusion does not disappear completely when a non-repeatable predicate is involved. [Wellwood et al.](#) ([2018](#)) suggest the answer may have to do precisely with the notion of repeatability in the following way: *graduate high school* is not a repeatable predicate, given our pragmatic knowledge about high school graduations, but since this 'once-per-individual' requirement

is not enforced grammatically in English, Wellwood et al. (2018, p. 563) argue that it is possible that speakers perceive this predicate as permitting an event counting reading. This answer seems to bring along two problems: First, it does not take into account the well-demonstrated role of non-grammatical, pragmatic knowledge in illusion processing (Maillat and Oswald 2009) in the sense that grammar does not operate alone. The (un)acceptability of linguistic stimuli is not shaped *exclusively* by grammatical constraints, but also by pragmatic knowledge, contextual assumptions, and cognitive biases (Leivada and Westergaard 2020). Therefore, a grammatical enforcement of the 'once-per-individual' requirement is not necessary, because pragmatic knowledge already precludes the event comparison re-analysis. The second problem is less theoretical: if the *non*-repeatable predicates that Wellwood et al. (2018) used in the design of their study eventually support a repeatable interpretation, the fundamental distinction between repeatable and non-repeatable predicates in their stimuli is significantly weakened, with all the implications that this carries for their results.

In their second experiment, Wellwood et al. (2018) tested how the features of the embedded clause subject (i.e., the *than* clause in (2)) impact the acceptability of illusions. The results showed that comparatives with non-repeatable predicates were less acceptable than those with repeatable predicates (crucially, this result does not refer only to illusions but also to control sentences), while in the illusion condition, there was again a difference in the ratings depending on the repeatability of the predicate (non-repeatable 4.27, repeatable 4.97). The results of the other experiments reported in Wellwood et al. (2018) confirmed the ones presented above; hence, they are not reported in detail.

De Dios-Flores (2016) ran a two-alternative forced-choice task that presented 35 participants with illusions like (2) together with their two possible interpretations (i.e., the ones given in (4a–b)), asking them to choose their preferred one. There were three conditions: the ambiguous condition included a repeatable predicate and permitted subject-inclusion (i.e., the subject of the embedded clause could be included in the denotation of the matrix subject), the event comparison condition blocked the additive reading by not permitting subject-inclusion (as in (7)), and the additive condition blocked the event comparison reading by featuring a non-repeatable predicate (as in (9)). The results showed that the ambiguous condition was indeed treated as ambiguous by the participants, as the preferences were split between the event comparison interpretation (60%) and the additive interpretation (40%). The event comparison condition that should elicit a strong event comparison interpretation—given that the additive reading was blocked—gave results similar to those of the ambiguous condition (61% for the event-comparison and 39% for the additive interpretation). Last, the additive condition, which blocked the event comparison interpretation, revealed a preference for the additive reading (69%). De Dios-Flores (2016) suggests that her results are consistent with the findings of Wellwood et al. (2009) and O'Connor et al. (2012) in showing that the preferred interpretation is the one facilitated by an event comparison re-analysis. However, recall that the ambiguous condition did not reveal a strong preference for the event comparison reading (i.e., a 60/40 balance), while from the two unambiguous conditions the one that showed the strongest effect in favor of one interpretation vs. the other was the additive one, something that suggests that participants responded more to the subject-inclusion cue. Regardless of what one concludes from these results, the fact is that participants were forced to choose one of the two answers, yet under the assumptions of the shallow comparison re-analysis proposed above, the illusion may well be ambiguous *in all its forms*, that is, with/without a repeatable predicate and with/without the possibility of subject-inclusion.

In the only study that tests illusions in a language other than English, Christensen (2016) presents data from a series of experiments on comparative illusions (he uses the term 'dead ends') in Danish. The results confirm that speakers of different languages are systematically tricked by comparative illusions. Christensen (2016) ran a timed two-choice AJT to 32 participants, asking them to judge the stimuli as 'good' or 'bad'. He found no effect of repeatability (contra Wellwood et al. 2018) and no difference between 'more' and 'fewer'. The analysis of the reaction times revealed that participants were fast at getting it wrong, that is, they were faster at providing the wrong answer (i.e., falling for the illusion). In the second experiment reported in Christensen (2016), 60 participants were

asked to evaluate whether or not the stimuli they were presented with made sense. The results confirmed that giving the correct answer (i.e., not falling for the illusion) was associated with longer response times. Christensen (2016) also found that people consistently failed to agree on one single interpretation for the comparative illusions. In light of previous neuroimaging work that did not provide evidence for extra syntactic processing for the illusions (Christensen 2010), the conclusion was that the interpretations provided by the participants are drawn from a small set of mutually (partially) incompatible interpretations ((5)–(6b)), and that this is done by direct syntactic analysis, not by re-analysis.

2.2. The Present Experiment: Method, Task, and Participants

I used Ibex Farm (Drummond 2013) in order to run a timed AJT that collected two measures: (i) acceptability judgments on a 3-point Likert scale that featured the options 'correct', 'neither correct, nor wrong', and 'wrong', and (ii) reaction times. The task was run in Standard Greek with 60 participants that fall into two groups:[2] 30 monolingual speakers of Standard Greek from Greece (16 female, average age: 37.22, SD: 13.7) and 30 bi(dia)lectal speakers of Cypriot Greek and Standard Greek from Cyprus (23 female, average age: 30.2, SD: 7.6). All participants completed or were in the process of completing tertiary education. In addition, all participants stated having Greek as their native language.[3]

All participants provided informed consent prior to their involvement in the study, in accordance with the Declaration of Helsinki. The Norwegian Center of Research Data reviewed and approved the study protocol (reference number: 55775/3/LH). The Cyprus National Bioethics Committee also reviewed and approved the study protocol (reference number: EEBK EΠ 2018.01.196). Exclusion criteria for participants involved non-native knowledge of Greek or native knowledge of other languages (assessed on the basis of self-report), reception of speech-pathology treatment and presence of neurological/neurodegenerative diseases (assessed on the basis of self-report), and inappropriate behavior in grammatical fillers.

Participants received the sentences one by one and were asked to press a key, judging their well-formedness on a 3-point Likert. It was explained to them that the experiment aimed to determine how the sentences sound to them, regardless of what grammar books say. Participants did not have the option to skip a test item or go back and change an answer they gave in a previous question. Before the task started, a warm-up session familiarized them with placing three fingers on keys '1' (wrong), '2' (neither correct nor wrong), and '3' (correct) in order to select an answer without moving the hands, thus minimizing noise in the online measure. The order of presentation of the sentences was pseudorandomized across participants.

With respect to the design features of the task, the ratio of fillers to actual test structures is 2:1. Each sentence (i.e., test items, grammatical fillers, and ungrammatical fillers) is 21 syllables

---

[2] Having one language of testing across groups is necessary because comparative constructions are grammaticalized in different ways across languages. If these ways translate into even minimally different lengths, this would inevitably muddle the credibility of the online measure.

[3] Native language was determined on the basis of self-report in both groups, however not all bilectal participants drew the distinction between Standard and Cypriot Greek or viewed them as two different varieties. Quite often speakers of the Cypriot dialect downplay the differences between the two varieties, reducing Cypriot only to an 'accent'. Roughly put, Cypriot Greek can be viewed as the home-variety, while Standard Greek is the official language that is also used in education, although in practice a lot of mixing is attested. Cypriot Greek and Standard Greek differ in all levels of linguistic analysis, while variation in the former is largely a matter of register. In general, the current Cypriot Greek koiné is a pan-Cypriot variety without identifiably local features (Tsiplakou et al. 2016). At the acrolectal pole of the continuum, one finds a standard variety, Standard Cypriot Greek, which is very similar to Standard Greek, but still recognizably different from it (Arvaniti 2010). However, Greek Cypriots do have exposure to Standard Greek, not only through education, but also through contact with Standard Greek media and a range of other social and cultural ties. With respect to the stimuli of the present experiment, none of its grammatical or lexical features fall among the features where Standard Cypriot Greek diverges from Standard Greek, following Arvaniti's (2010) presentation of the differences between the two varieties. Therefore, the task of judging the acceptability of the tested sentences should not pose any difficulty to the bilectal group. For more information about the linguistic profile of Greek Cypriots see Rowe and Grohmann (2013).

long after phonetic transcription. Grammatical fillers involve comparative structures too, hence they are somewhat similar to the test structures (e.g., "More times I've been to England than to Germany"). Ungrammatical fillers contain agreement errors, aiming to distract participants from the tested phenomenon of comparatives (e.g., "The key to these cabinets there are on the marble table"). Consequently, the grammatical fillers are easy to detect and give a high rating (target response: correct); likewise, the ungrammatical ones are easy to detect and give a low rating (target response: wrong). The test items feature comparative illusions that fall into two conditions: condition 1 involves a repeatable predicate, while condition 2 involves a non-repeatable predicate. Each condition has five items and all participants encountered all items, so excluding fillers, 1200 data points were collected, 600 for each measure. Table 1 presents the test items and explains the feature they test (see Supplementary Materials, for the Greek version of test items and the fillers).

Table 1. Test items.

| Condition	Test Item	Tested Feature
Repeatable predicate	1. More people have been to London than I have.	+ subject-inclusion 1st person
	2. More people have been to Milan than you have.	+ subject-inclusion 2nd person
	3. More boys have been to Paris than he has.	+ subject-inclusion 3rd person, gender congruent
	4. More girls have been to Stockholm than he has.	− subject-inclusion 3rd person, gender incongruent
	5. Fewer people have been to Berlin than I have.	− subject-inclusion, *fewer* type
Non-repeatable predicate	6. More kids have finished high school than I have.	+ subject-inclusion 1st person
	7. More kids have finished school than you have.	+ subject-inclusion 2nd person
	8. More men have finished school than he has.	+ subject-inclusion 3rd person, gender congruent
	9. More men have finished high school than she has.	− subject-inclusion 3rd person, gender incongruent
	10. Fewer people have finished high school than I have.	− subject-inclusion, *fewer* type

3. Results

As often happens in experiments that collect reaction times, this measure did not show a normal distribution due to a skewed right tail. The standard logarithm ($RT\prime = \log_{10}(RT)$) was used to normalize the data and a classical ±3 SD filter was applied to detect outliers. Only four reaction times were removed (for consistency the acceptability judgments linked to them were also removed), which led to the elimination of 0.33% of the data.

With respect to the acceptability judgments, the results revealed that the illusory effect reported for English speakers also holds in Greek. Figure 1 shows that the predominant judgment is 'correct', meaning that the deviation was not spotted, and participants were largely tricked into accepting the illusions as well-formed sentences. At the same time, the results appear quite balanced, as Figure 1 shows that a big number of participants fully rejected the sentences as 'wrong'. By comparison, the 'neither' responses are relatively few. This suggests that the sentences were either parsed successfully such that an interpretation was reached, in which case they were given the judgment 'correct', or that parsing failed to yield any interpretation, such that the illusion was spotted, and rejected as 'wrong'.
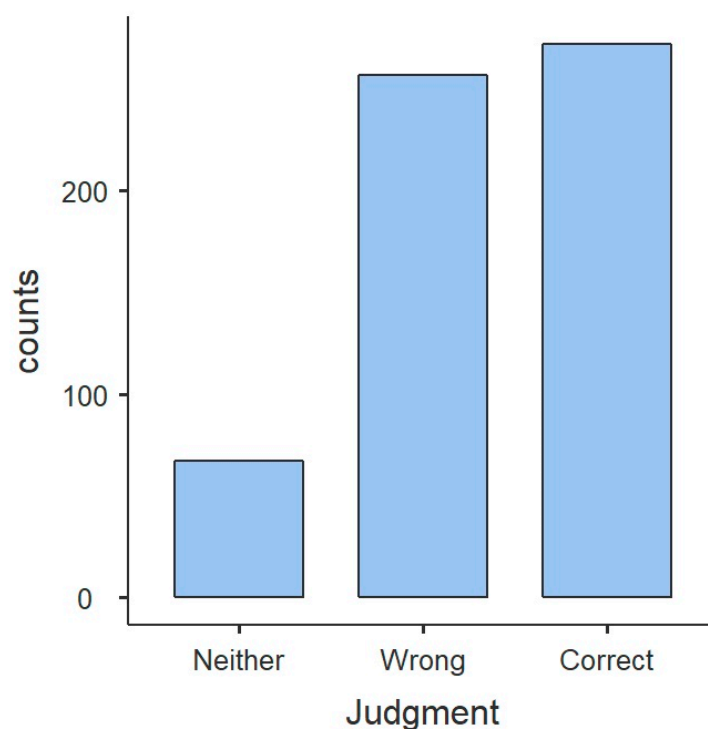
Figure 1. Acceptability judgments across participants and test items (596 data points). Correct/total: 45.6%, wrong/total: 43.1%, neither/total: 11.2%.

Upon splitting participants into two groups, monolinguals and bilectals, an interesting picture emerges. As shown in Figure 2, the two groups differ in their performance: the predominant answer for the monolinguals is 'correct', similar to what was observed in Figure 1. The predominant answer for the bilectals is 'wrong', meaning that the deviation was spotted and rejected more times. The two groups performed alike in terms of how often they chose the middle, 'neither' option. Using JAMOVI, a generalized linear model showed the difference in acceptability judgments between the two groups to be moderately significant, treating the dependent variable as multinomial ($\chi^2 = 2.4$, $p = 0.002$).
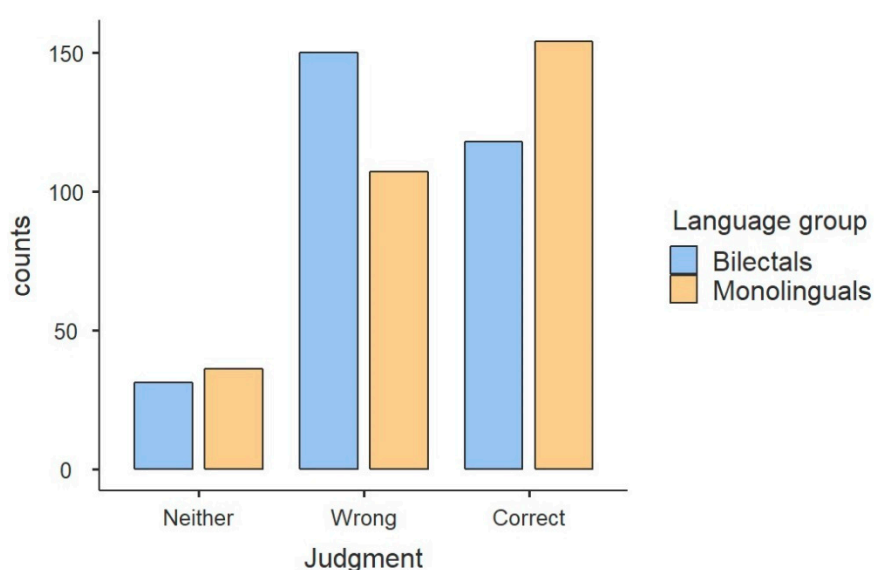
Figure 2. Acceptability judgments split for group. Bilectals, correct/total: 39.4%, wrong/total: 50.1%. neither/total: 10.3%. Monolinguals, correct/total: 51.8%, wrong/total: 36%. neither/total: 12.1%.

Turning to the online measure, the analysis of reaction times did not reveal evidence for a statistically significant difference between the two groups ($\chi^2 = 0.0279$, $p = 0.867$). Importantly, the interaction between judgment and reaction times was found to be highly significant both in the entire group of participants ($\chi^2 = 42.4$, $p < 0.001$) and after splitting for group (monolinguals: $\chi^2 = 16.8$, $p < 0.001$; bilectals: $\chi^2 = 30.3$, $p < 0.001$). As Figure 3 shows, the judgment 'correct' had the shortest decision time (agreeing with the results of Christensen (2016)), followed by 'wrong', and then 'neither'. This picture remains unaltered after splitting for group (Figure 4).
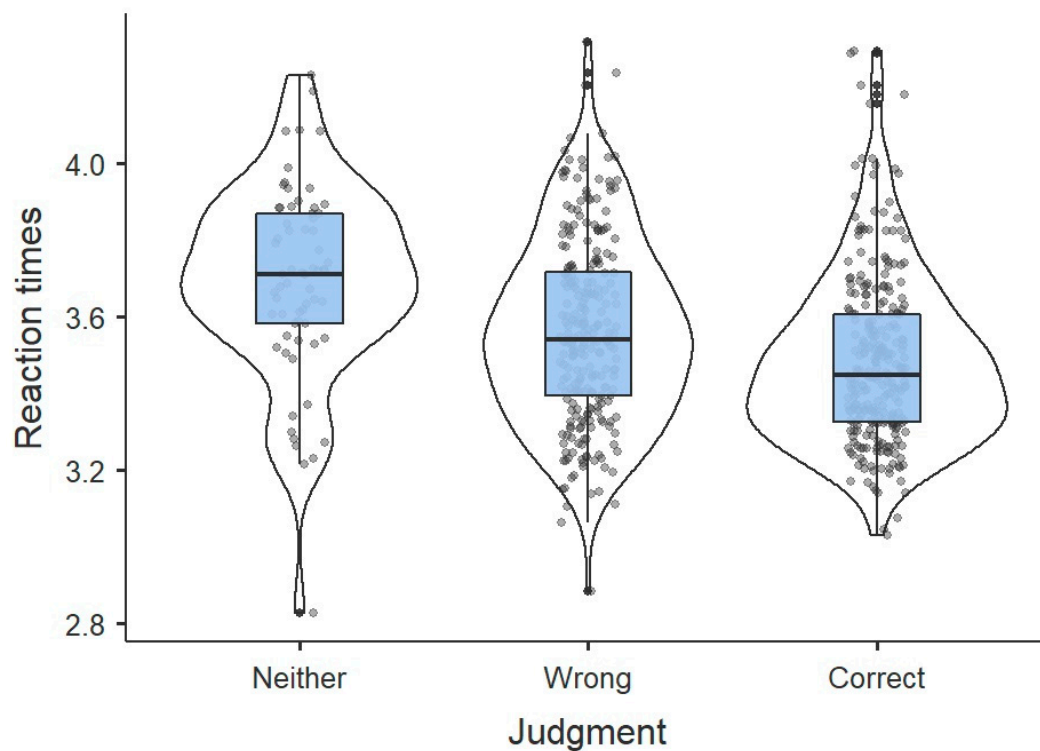
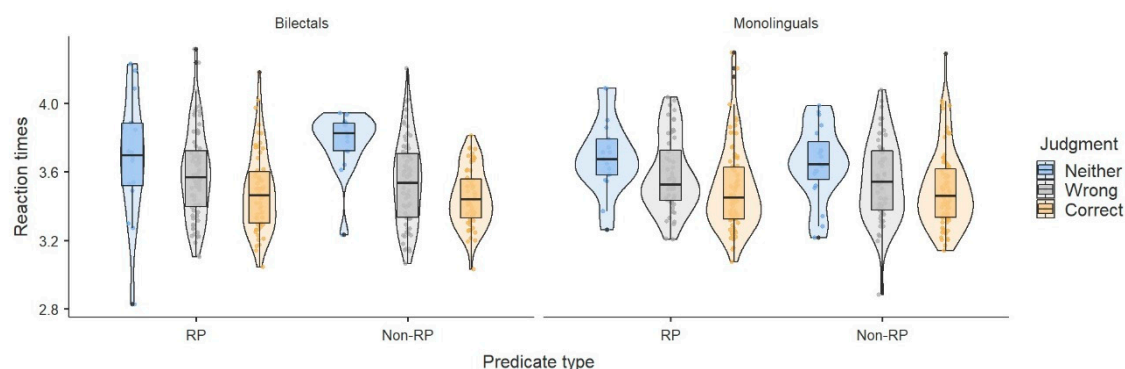**Figure 3.** Log-transformed reaction times across groups split for judgment.
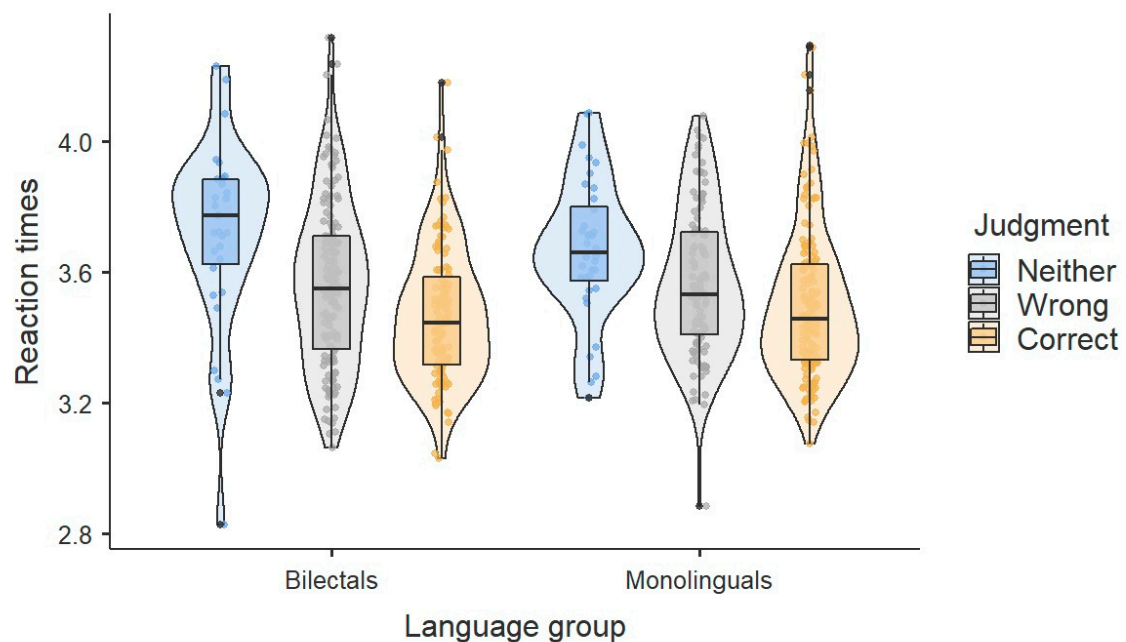
Figure 4. Log-transformed reaction times split for language group and judgment.

Contrary to what was expected on the basis of the results obtained from testing speakers of English (O'Connor et al. 2012; O'Connor 2015; De Dios-Flores 2016; Wellwood et al. 2018), repeatable predicates are not linked to a more robust illusion effect, that is, to higher acceptability judgments, corroborating the finding of Christensen (2016). When taking the two groups of participants together, the difference between the two conditions (i.e., repeatable vs. non-repeatable predicate, see Table 1) fails to reach statistical significance in relation to either measure (acceptability judgments: $\chi^2 = 0.176$, $p = 0.916$; reaction times: $\chi^2 = 1.06$, $p = 0.304$). Splitting for group, the impact of condition remains non-significant (monolinguals, acceptability judgments: $\chi^2 = 0.426$, $p = 0.808$; monolinguals, reaction times: $\chi^2 = 0.292$, $p = 0.589$; bilectals, acceptability judgments: $\chi^2 = 0.028$, $p = 0.986$; bilectals, reaction times: $\chi^2 = 0.811$, $p = 0.368$). As Figure 5 shows, bilectals have a remarkably uniform performance in terms of judgments across the two conditions. On the other hand, monolinguals show a very slight increase in their acceptability ratings in the repeatable condition, something that is in line with the results from English speakers. With respect to the reaction times, the pattern that was shown in Figures 3 and 4 remains the same after splitting for condition; in both conditions, the judgment 'neither' is accompanied by the longest reaction times (Figure 6).
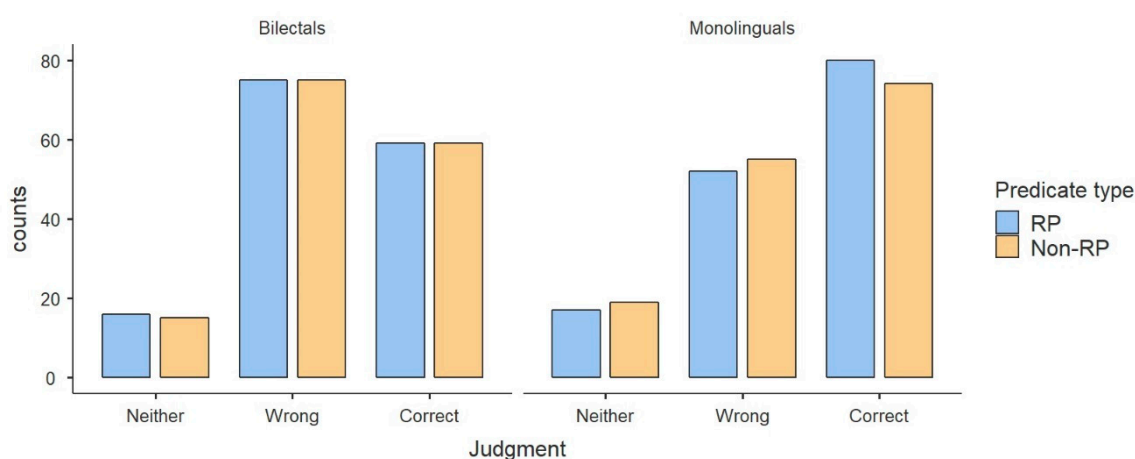
Figure 5. Acceptability judgments split for language group and type of predicate. RP = repeatable.

Figure 6. Log-transformed reaction times split for language group and type of predicate. RP = repeatable.

The type of test item (see 'tested feature' in Table 1) was not found to have a statistically significant effect on either the acceptability judgments or the reaction times, meaning that it was not the case that one type of illusion was accepted/rejected significantly more than others or that one type of illusion took significantly longer to process (effect of test item in both groups together, acceptability judgments: $\chi^2 = 13.6$, $p = 0.753$; both groups together, reaction times: $\chi^2 = 6.92$, $p = 0.646$; monolinguals, acceptability judgments: $\chi^2 = 10.0$, $p = 0.930$; monolinguals, reaction times: $\chi^2 = 7.78$, $p = 0.557$; bilectals, acceptability judgments: $\chi^2 = 13.3$, $p = 0.771$; bilectals, reaction times: $\chi^2 = 6.09$, $p = 0.731$).

Seeing the picture that Figures 3, 4 and 6 paint, one may wonder whether the aforementioned highly significant effect of judgment on reaction times is a true effect or whether it is induced by the 'neither' responses, which consistently show up as being associated with longer processing times. Put more clearly, the question is the following: If one views the 'neither' responses not as a real, spontaneous preference, but as a case of 'last resort' answer due to time pressure, will discarding them nullify the claimed effect? To answer this, the interaction between reaction times and acceptability judgments was re-analyzed after removing the 'neither' responses from the sample. The results affirm the robustness of the effect, which remains stable. The difference between the two remaining judgments, 'correct' and 'wrong', in terms of their associated decision times is statistically significant in monolinguals ($\chi^2 = 4.75$, $p = 0.029$), bilectals ($\chi^2 = 10.8$, $p = 0.001$), and when taking both groups together ($\chi^2 = 14.3$, $p < 0.001$).

4. Discussion

As mentioned in Section 1, the aim of this study is to address the following two questions:

Research question 1: Are the factors that have been linked to inducing the illusion at its strongest equally robust in languages other than English?

Research question 2: Are people with different developmental trajectories fallible to comparative illusions to different degrees?

Starting off from the first question, the answer is negative. If it is the case that speakers settle on either the event comparison or the set comparison re-analysis when they accept the illusion, how is it possible that they do either with a sentence that allows neither, by not permitting subject-inclusion and not involving a repeatable predicate? Let us demonstrate this problem with a concrete example. As Table 1 shows, there are sentences in the present study that do not involve a repeatable predicate. Among them, there are two that also disallow subject-inclusion, repeated below as (12) and (13):

(12)　More men have finished high school than she has.
(13)　Fewer people have finished high school than I have.

Thus, (12) does not permit an event comparison reading. As various studies on the acceptability of comparative illusions in English-speaking populations have repeatedly argued (O'Connor et al. 2012; O'Connor 2015; De Dios-Flores 2016; Wellwood et al. 2018), the event comparison interpretation is only possible if the semantic properties of the predicate express a repeatable action, and this is not the case here. Furthermore, (12) does not permit a set comparison reading, because the subject of the embedded clause cannot be included in the denotation of the matrix clause subject. Similarly, (13) does not permit an event comparison reading and the additive/set comparison reading is also ruled out. It is hard to even construct an interpretation of (13) that goes along the lines of (4a) and (5) for the set comparison reading, as shown in (14a) below. Equally strange is the event comparison reading, along the lines of (4b) and (6), given in (14b):

(14a) The set of people that have finished high school involves *fewer* members than the set of just me.
(14b) People have finished high school *fewer times* than I have finished high school.

Under these assumptions, we would expect a sharp decrease in the acceptability of these two test items, compared to the other test items given in Table 1, simply because no re-analysis—from the ones proposed in the literature—is possible for them. Recall that the effect of test item on acceptability judgments was not found to be significant. However, let us further analyze this finding by examining how (13) fares in comparison with the other test items. Figure 7 shows how many times a test item received the judgment 'correct', meaning that the illusion was not detected. The maximum possible number for each test item is 60. (13) received the judgment 'correct' 28 times. This is one time *more* than the test item marked in Figure 7 as 'RP, 3rd person gender congruent', which features both a repeatable predicate and permits subject inclusion. This finding is hard to explain under any of the two re-analyses of the illusion that were proposed for English.
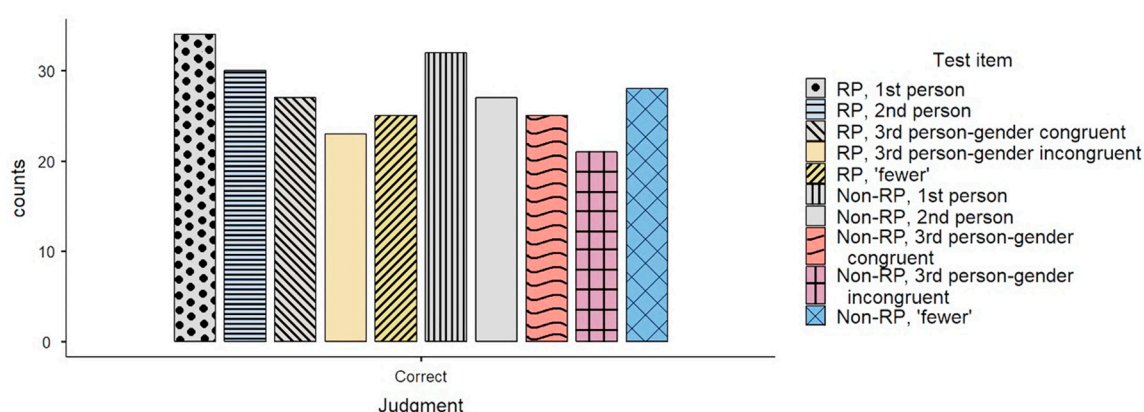
Figure 7. Counts of judgment 'correct' per test item.

Giving a negative answer to research question 1 does not undermine the credibility of the re-analyses proposed in the literature. In fact, I argue that drawing insights from these re-analyses can provide useful answers with respect to the *what*, *why*, and *how* questions that surround the phenomenon of comparative illusions.

Beginning with *why*, it was argued in Section 1 that the event comparison hypothesis focuses on how meaning is construed (its answer: by resorting to a grammar-based interpretation that facilitates the event comparison re-analysis; Wellwood et al. 2018), but it does not explain equally clearly why the parser fails to spot the illusion in the first place. Under the shallow comparison scenario that I propose here, the answer to *'why does the parser fail to spot the illusion in the first place?'* is because it encounters values that are pretty close to the predictions it forms on the basis of previous experience. To explain this further, the parser is oblivious to the manipulative context that induces the illusion, therefore it contextualizes a sentence like (13), by treating templates it has previously encountered (i.e., comparison of events and comparison of individuals) and knows to be meaningful as such. Put differently, given the tendency for radical contextualization, when the parser is dealing with a sentence that is largely syntactically correct, it seeks to assign a meaning to it, because it *presupposes* that the sentence *has* a meaning. For the same reason, people can easily assign structure to semantically anomalous sentences, such as Chomsky (1957) famous example "colorless green idea sleep furiously". If the parser resorts to a System 1 heuristic, it legitimizes this presupposition by constructing an interpretation (along the lines of the *what* question discussed below). If, on the other hand, System 2 intervenes and challenges this presupposition, the parser may spot the illusion. Both systems operate in tandem and this explains why there is variation in spotting the illusion even across individuals that fit the same behavioral profile (Wellwood et al. 2018). Overall, this answer to the *why* question fits theories that assume a complex interplay between heuristic processing and deep syntactic parsing (Karimi and Ferreira 2016). Evidence in favor of the deep grammar-based parsing account comes from Figure 7, which shows that people are sensitive to grammatical cues such as gender incongruency, but not to a degree that would lead to a statistically significant effect of test item.

Turning to the *how* question (i.e., *'how is the meaning of the illusion construed?'*), the explanation I propose partly agrees with the one presented in Phillips et al. (2011) and Wellwood et al. (2018): by positing grammar-based interpretations. Evidence for this claim comes from the finding that certain judgments are associated with longer reaction times, as shown in Figures 3, 4 and 6. The parser is actively engaged in grammatical processing, construing, evaluating, and discarding different interpretations, which is why the 'neither' and 'wrong' responses consistently take longer to form. In other words, it is not the case that most of the people who spotted the illusion did so immediately. This explanation parts company with the one by Phillips et al. (2011) and Wellwood et al. (2018) in accepting that deep grammatical processing occurs, but until a System 1 heuristic takes over. For some people, this happens early, and the illusion is accepted as well-formed straight away. For some people, the heuristic kicks in later, and the illusion may be spotted or not.

Finally, the speculative answer to the *what* question (i.e., *'what is the interpretation that the illusion finally gets?'*) is that there is not one. The acceptability of (13) is not predicted by either the set comparison or the event comparison re-analysis. Thus, (13) should not be acceptable, and yet it is. Under the shallow comparison scenario that I propose, (13) is acceptable because people do not settle on one full-fledged interpretation. They have parsed the sentence, such that their parser has constructed a potentially ambiguous and vague interpretation, and this is enough to render a high acceptability rating. This approach is in line with Christensen (2010) claim that comparative illusions are "semantic dead ends" in the sense that they cannot be assigned a full semantic interpretation. Perhaps the strongest argument in favor of the proposal I put forth here comes from a crucial observation made by Wellwood et al. (2018): "O'Connor (2015) results are suggestive here: she found that participants' reading times were slower following the ellipsis site for C[omparative] I[llusion]-type sentences as compared to controls, and that the reaction times correlated with acceptability ratings for controls but not for illusions. If shallow processing was behind high acceptability ratings for CIs, *we would expect to see no slowdown when participants eventually gave a high rating*" (p. 577; emphasis added). This is exactly what we see in the results of the present experiment: the highest rating (i.e., the response 'correct') shows no slowdown in reaction times, unlike the other two ratings (Figures 3, 4 and 6). Put more succinctly, slower reaction times do seem to go hand-in-hand with spotting the illusion.

With respect to the second research question and to whether people with different developmental trajectories are fallible to comparative illusions to different degrees, the answer seems to be a tentative yes. Given that variation in spotting the illusion was observed among speakers of the same language and even fit the same behavioral profile (Wellwood et al. 2018), it is perhaps unsurprising that such interspeaker differences show up as more pronounced when one compares individuals from different behavioral profiles/developmental trajectories. The obtained results provide tentative evidence for a significantly better ability of bilectals to spot and reject the illusions as ill-formed. In relation to why this happens, only speculations can be offered at this point, because many factors can be identified as possible culprits behind this finding.

First, the difference could be due to some grammar-specific particularities of Cypriot Greek. Although all participants were tested in Standard Greek, the bilectal speakers also use a second variety in their daily life: Cypriot Greek. More specifically, the issue could be that the tested illusions involve predicates in present perfect, which is arguably a novel development in Cypriot Greek, and absent from older forms of this variety (see Melissaropoulou et al. 2013 for an overview). Under this explanation, the fact that bilectals reject more than monolinguals the comparative illusions may have nothing to do with spotting the illusions themselves. Instead, it could be the result of cross-linguistic interference from their home-variety, in which this tense is absent and thus may sound alien to them. Overall, this seems to be a weak explanation, because the possibility of such interference is low for two reasons. First, recent studies demonstrate the existence of present perfect in contemporary Cypriot Greek (Melissaropoulou et al. 2013; Tsiplakou et al. 2016; Tsiplakou et al. 2019) and the tested participants are quite young (mean age in bilectals: 30.2, SD: 7.6). Second, it should not be overlooked that the language of testing is Standard Greek, a variety of Greek to which the bilectals definitely have exposure as it is one of the official languages of the state. In this sense, the use of present perfect should not cause any interference.

Another culprit behind this better performance of bilectals boils down to the fact that unlike their monolingual peers, they have at least two grammars (Standard and Cypriot Greek), which they use almost on a daily basis. To spell out the idea more clearly, if bilectals have frequent opportunities to code-switch, perhaps this constant engagement of top-down control processes, which has been explicitly linked to stimulus-driven code-switching (Blanco-Elorrieta and Pylkkänen 2018; Leivada et al. 2020), is responsible for their better performance in this task. It can be hypothesized that code-switching may facilitate the setting of a higher processing threshold through sharpened top-down control processes, something that would entail decreased fallibility to grammatical illusions. Setting a higher processing threshold could mean better matching to retrieval cues together with a superior

ability to attend selectively to various aspects of the stimuli, (dis)engaging attentional resources accordingly. In the literature, partial matching to retrieval cues has been already linked to agreement illusions (Phillips et al. 2011; Phillips 2013), and it is not implausible that it plays a role in comparative illusions as well. This interpretation of the results is in agreement with many behavioral studies that report an enhanced ability of bilinguals to make accurate acceptability judgments in the context of misleading sentences that feature semantic deviations (Galambos and Hakuta 1988; Bialystok 1988; Ricciardelli 1992; Cromdal 1999). Bilingual children have been also found to have a heightened pragmatic sensitivity (Groba et al. 2018). Viewing grammatical illusions as semantico-pragmatic constructs that arise within a manipulative context (Maillat and Oswald 2009), the enhanced ability of bilinguals to provide selective attention to specific aspects of a linguistic representation may be the key to explain what lead to the successful identification of these illusions as ill-formed sentences in one of the two tested groups.

A third hypothesis is that bilectals are anxious to prove that they can speak Standard Greek, hence they sometimes tend to hyper-correct and reject ill-formed sentences that are perfectly acceptable in Standard Greek. This (meta)linguistic insecurity is well-documented for both child and adult populations in Cyprus (Grohmann and Leivada 2012; Leivada et al. 2017) and has to do with the complex dynamics that underlie the standard-dialect continuum. If this hypothesis holds, the difference in the performance between monolinguals and bilectals does not have a cognitive background as the previous hypothesis posits, but a sociolinguistic one. To evaluate this hypothesis, it seems that it suffers from one major problem. If the source of the low acceptability ratings in bilectals was due to their tendency to hyper-correct, the grammatical fillers should also show this effect; they do not. In raw numbers, the performance of the two groups in the grammatical fillers amounts to 4/150 counts of 'wrong' for monolinguals and 3/150 for bilectals. In other words, both groups correctly accepted the grammatical fillers as 'correct'. To put these numbers in perspective, the equivalent number of 'wrong' judgments in the illusions (i.e., which means that the illusion was spotted and correctly rejected as wrong) is 109/300 for monolinguals vs. 150/300 for bilectals. It can thus be concluded that since the grammatical fillers do not show the slightest evidence of hyper-correction, this hypothesis lacks empirical support.

A fourth explanation could be the interaction of the two previous ones: setting a higher processing threshold makes bilectals less fallible to illusions, but at the same time, their performance is driven by linguistic insecurity, hence they also pay more attention to the stimuli. This is a valid hypothesis, but then one should account for the fact that the reaction times were not found to be significantly different between the two groups. If bilectals paid more attention and processed the stimuli more, the reaction times should have reflected this, because they are sensitive to the number of operations and alternative meanings that are computed (Bates et al. 1982). In order to evaluate the strength of this explanation, let us observe again the obtained reaction times. Indeed, the statistical analysis did not reveal a difference between the two groups ($\chi2 = 0.0279$, p = 0.867). In addition, the interaction between the type of judgment and reaction times was highly significant in both groups alike (monolinguals: $\chi^2 = 16.8$, $p < 0.001$; bilectals: $\chi^2 = 30.3$, $p < 0.001$). Recall, however, that after the 'neither' responses were removed from the sample, the difference between the two remaining judgments, 'correct' and 'wrong' in terms of their associated decision times, is *differentially* significant in the two groups (monolinguals: $\chi^2 = 4.75$, $p = 0.029$; bilectals: $\chi^2 = 10.8$, $p = 0.001$). Evidently, the result is only marginally significant in monolinguals (and if it is corrected for multiple comparisons this significance would disappear), while it is strong in bilectals. Put another way, there is evidence that the reaction times in bilectals *are* more sensitive to the given acceptability judgment in the sense that spotting illusion and rejecting it as 'wrong' is indeed associated with longer processing time. The 'wrong' responses in monolinguals are also linked with longer processing times (Figure 4), but the difference is not as strong as the one observed for bilectals. Overall, this explanation can be neither accepted nor rejected at present. On the one hand, it is possible that failing to find evidence for an overall group difference in reaction times is due to the small number of participants in each group and the big number of 'correct' responses in both

groups in the following sense: the judgment 'correct' is the one associated with faster reaction times in both groups and it is sufficiently plentiful in each group to mask an overall group difference. On the other hand, the fact remains that the two groups were not found to differ in terms of their overall reaction times. Crucially, removing the 'neither' responses from the sample is perhaps informative and helpful in order to discern subtler patterns of variation, but the results of such a move should be interpreted with caution. Excluding a portion of the obtained data for an analysis brings along the danger of data mining, by looking at too many possible associations at a subset level.

The last explanation that will be contemplated is that the difference between monolinguals and bilectals is an accidental one; a difference of the sort one can find any time two groups of people are compared. Given that there is no literature on the topic of grammatical illusions in Greek speaking populations, the obtained results cannot be interpreted in comparison with other studies from the same population. Until they are replicated, the possibility that this is an accidental difference cannot be discarded.

Overall, the results together with the above evaluation of each hypothesis seem to support the second explanation, lending support to the idea that the mental juggling involved in using two or more languages on a daily basis may play a role in establishing a higher processing threshold. However, more research is needed in order to evaluate the effect and the robustness of this result.

5. Outlook

The present experiment investigated comparative illusions: a special type of stimulus that tricks the cognitive parser in eliciting high acceptability judgments, despite its ill-formedness. Our knowledge of comparative illusions is fairly limited, given that the relevant literature involves experiments performed with monolingual, neurotypical speakers of English or Danish. In this experiment, a comparative approach was taken through testing two populations: monolingual speakers of Standard Greek and bilectal speakers of Standard and Cypriot Greek. A timed acceptability judgment task was run, featuring stimuli that are sensitive to different factors that have been argued to play a role in inducing the illusion, such as the ability to include the embedded clause subject in the denotation of the matrix clause subject and the repeatability of the predicate. Although the literature from English speakers suggests that the illusion exists as long as it is possible to construct an event counting interpretation, the results of this experiment did not fully confirm this claim. For this reason, a novel synthesis of factors was presented above in order to address the *why*, *what*, and *how* questions that surround the phenomenon of comparative illusions.

The answer to *'why does the parser fail to spot the illusion in the first place?'* goes through understanding that the parser is oblivious to the manipulative context that induces the illusion. Given the human bias for radical contextualization, the parser actively seeks to assign a meaning to comparative illusions because it presupposes that these sentences have a meaning. Upon encountering a sentence that does not show any *marked* deviation from templates that the parser recognizes as well-formed, one possible outcome is that the illusion fails to be spotted and a meaning is construed. The answer to *how* this happens boils down to two systems operating in tandem along the lines of a dual-process theory of reasoning: partly by computing grammar-based interpretations and partly by employing processing heuristics. Last, the answer to the *what* question (i.e., *'what is the interpretation that the illusion finally gets?'*) is that speakers do not settle on one full-fledged interpretation. Under the shallow comparison scenario that I propose, building on a long line of literature (Townsend and Bever 2001; Karimi and Ferreira 2016), the parser may be satisfied with an underspecified interpretation of comparative illusions. The present experiment provided novel evidence for this claim through the collected reaction times. As argued in the literature, if shallow processing is responsible for the high acceptability of comparative illusions, slower reaction times should go hand-in-hand with detecting the illusion (O'Connor 2015; Wellwood et al. 2018). This is indeed what my findings show: spotting the illusion is associated with longer reaction times. The same finding is reported in Christensen (2016) for speakers of Danish, thus suggesting a stable cross-linguistic pattern.

Last, the findings of the present study provide tentative support for the claim that people with different developmental trajectories are fallible to comparative illusions to different degrees. It was hypothesized that having a second language may facilitate setting a higher processing threshold, resulting in an enhanced ability to spot the illusion. At the same time, it was acknowledged that the tested sample of speakers is rather small and there is always a danger in drawing firm conclusions from small-sample studies. While an effect of developmental trajectory cannot be ruled out, this is only a hypothesis and the exact causes or dimensions of this effect have not been determined yet. It is likely that future research on bigger and better controlled samples will shed more light on the topic.

References

Adler, Jonathan E. 1984. Abstraction is uncooperative. *Journal for the Theory of Social Behaviour* 14: 165–81. [CrossRef]

Anderson, Judith L. 1998. Embracing uncertainty: The interface of Bayesian statistics and cognitive psychology. *Conservation Ecology* 2: 2. [CrossRef]

Arvaniti, Amalia. 2010. Linguistic practices in Cyprus and the emergence of Cypriot Standard Greek. *Mediterranean Language Review* 17: 15–45.

Bates, Elizabeth, Sandra McNew, Brian MacWhinney, Antonella Devescovi, and Stan Smith. 1982. Functional constraints on sentence processing: A cross linguistic study. *Cognition* 11: 245–99. [CrossRef]

Bialystok, Ellem. 1988. Levels of bilingualism and levels of linguistic awareness. *Developmental Psychology* 24: 560–67. [CrossRef]

Blanco-Elorrieta, Esti, and Liina Pylkkänen. 2018. Ecological validity in bilingualism research and the bilingual advantage. *Trends in Cognitive Sciences* 22: 1117–22. [CrossRef]

Bresnan, Joan. 1973. Syntax of the comparative clause construction in English. *Linguistic Inquiry* 4: 275–343.

Chomsky, Noam. 1957. *Syntactic Structures*. The Hague: Mouton.

Christensen, Ken Ramshøj. 2010. Syntactic reconstruction and reanalysis, semantic dead ends, and prefrontal cortex. *Brain and Cognition* 73: 41–50. [CrossRef]

Christensen, Ken Ramshøj. 2016. The dead ends of language: The (mis)interpretation of a grammatical illusion. In *Let Us have Articles Betwixt Us—Papers in Historical and Comparative Linguistics in Honour of Johanna L. Wood*. Edited by Sten Vikner, Henrik Jørgensen and Elly Van Gelderen. Aarhus: Aarhus University, pp. 129–60.

Cosmides, Leda, and John Tooby. 1996. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition* 58: 1–73. [CrossRef]

Cromdal, Jakob. 1999. Childhood bilingualism and metalinguistic skills: Analysis and control in young Swedish-English bilinguals. *Applied Psycholinguistics* 20: 1–20. [CrossRef]

De Dios-Flores, Iria. 2016. More people have presented in conferences than I have. Comparative illusions: When ungrammaticality goes unnoticed. In *On the Move: Glancing Backwards to Build a Future in English Studies*. Edited by Aitor Ibarrola-Armendariz and Jon Ortiz de Urbina Arruabarrena. Bilbao: Universidad de Deusto, Servicio de Publicaciones, pp. 219–28.

Drummond, Alex. 2013. Ibex Farm. Available online: http://spellout.net/ibexfarm (accessed on 18 July 2020).

Ferreira, Fernanda. 2003. The misinterpretation of noncanonical sentences. *Cognitive Psychology* 47: 164–203. [CrossRef]

Ferreira, Fernanda, and Nikole D. Patson. 2007. The 'good enough' approach to language comprehension. *Language and Linguistics Compass* 1: 71–83. [CrossRef]

Ferreira, Fernanda, Karl G. D. Bailey, and Vittoria Ferraro. 2002. Good-enough representations in language comprehension. *Current Directions in Psychological Science* 11: 11–15. [CrossRef]

Galambos, Sylvia Joseph, and Kenji Hakuta. 1988. Subject-specific and task-specific characteristics of metalinguistic awareness in bilingual children. *Applied Psycholinguistics* 9: 141–62. [CrossRef]

Gallistel, C. R., and Adam Philip King. 2009. *Memory and the Computational Brain*. Malden: Wiley-Blackwell.

Grice, H. Paul. 1957. Meaning. *The Philosophical Review* 62: 377–88. [CrossRef]

Groba, Agnes, Annick De Houwer, Jan Mehnert, Sonja Rossi, and Hellmuth Obrig. 2018. Bilingual and monolingual children process pragmatic cues differently when learning novel adjectives. *Bilingualism: Language and Cognition* 21: 384–402. [CrossRef]

Grohmann, Kleanthes K., and Evelina Leivada. 2012. Interface ingredients of dialect design: Bi-x, socio-syntax of development, and the grammar of Cypriot Greek. In *Towards a Biolinguistic Understanding of Grammar*. Edited by Anna Maria Di Sciullo. Amsterdam: John Benjamins, pp. 239–62.

Hilton, Denis J. 1995. The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin* 118: 248–71. [CrossRef]

Kahneman, Daniel. 2000. A psychological point of view: Violations of rational rules as a diagnostic of mental processes. *Behavioral & Brain Sciences* 23: 681–83.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kahneman, Daniel, and Amos Tversky. 1982. On the study of statistical intuitions. *Cognition* 11: 123–41. [CrossRef]

Kamas, Eleen N., Lynne M. Reder, and Michael S. Ayers. 1996. Partial matching in the Moses illusion: Response bias not sensitivity. *Memory & Cognition* 24: 687–99.

Karimi, Hossein, and Fernanda Ferreira. 2016. Good-enough linguistic representations and online cognitive equilibrium in language processing. *The Quarterly Journal of Experimental Psychology* 69: 1013–40. [CrossRef]

Krifka, Manfred. 1990. Four thousand ships passed through the lock: Object-induced measure functions on events. *Linguistics and Philosophy* 13: 487–520.

Lafer-Sousa, Rosa, Katherine L. Hermann, and Bevil R. Conway. 2015. Striking individual differences in color perception uncovered by "the dress" photograph. *Current Biology* 25: R545–46. [CrossRef]

Leivada, Evelina, and Marit Westergaard. 2020. Acceptable ungrammatical sentences, unacceptable grammatical sentences, and the role of the cognitive parser. *Frontiers in Psychology* 11: 364. [CrossRef]

Leivada, Evelina, Elena Papadopoulou, Maria Kambanaros, and Kleanthes K. Grohmann. 2017. The Influence of bilectalism and non-standardization on the perception of native grammatical variants. *Frontiers in Psychology* 8: 205. [CrossRef] [PubMed]

Leivada, Evelina, Marit Westergaard, Jon Andoni Duñabeitia, and Jason Rothman. 2020. On the phantom-like appearance of bilingualism effects on neurocognition: (How) should we proceed? *Bilingualism: Language and Cognition*, 1–14. [CrossRef]

Maillat, Didier, and Steve Oswald. 2009. Defining manipulative discourse: The pragmatics of cognitive illusions. *The International Review of Pragmatics* 1: 348–70. [CrossRef]

Melissaropoulou, Dimitra, Charalambos Themistocleous, Stavroula Tsiplakou, and Symeon Tsolakidis. 2013. The Present Perfect in Cypriot Greek revisited. In *Studies in Language Variation—European Perspectives IV*. Edited by Peter Auer, Javier Caro Reina and Göz Kaufmann. Amsterdam: John Benjamins, pp. 159–72.

Montalbetti, Mario. 1984. After Binding. Doctoral dissertation, MIT, Cambridge, MA, USA.

O'Connor, Ellen. 2015. Comparative Illusions at the Syntax-Semantics Interface. Ph.D. dissertation, University of Southern California, Los Angeles, CA, USA.

O'Connor, Ellen, Roumyana Pancheva, and Elsi Kaiser. 2012. Evidence for online repair of Escher sentences. In *Proceedings of Sinn und Bedeutung 17*. Edited by Emmanuel Chemla, Vincent Homer and Grégoire Winterstein. Paris: ENS, pp. 363–80.

Phillips, Colin. 2013. Parser-grammar relations. We don't understand everything twice. In *Language down the Garden Path. The Cognitive and Biological Basis for Linguistic Structures*. Edited by Montserrat Sanz, Itziar Laka and Michael K. Tanenhaus. Oxford: Oxford University Press, pp. 294–315.

Phillips, Colin, Matthew W. Wagers, and Ellen F. Lau. 2011. Grammatical illusions and selective fallibility in real-time language comprehension. In *Experiments at the Interfaces*. Edited by Jeffrey Runner. Bingley: Emerald Publications, vol. 37, pp. 147–80.

Reder, Lynne M., and Gail W. Kusbit. 1991. Locus of the Moses illusion: Imperfect encoding, retrieval, or match? *Journal of Memory & Language* 30: 385–406.

Ricciardelli, Lina A. 1992. Bilingualism and cognitive development in relation to threshold theory. *Journal of Psycholinguistic Research* 21: 301–16. [CrossRef]

Rowe, Charley, and Kleanthes K. Grohmann. 2013. Discrete bilectalism: Towards co-overt prestige and diglossic shift in Cyprus. *International Journal of the Sociology of Language* 224: 119–42. [CrossRef]

Sprouse, Jon, and Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax. *Journal of Linguistics* 48: 609–52. [CrossRef]

Stanovich, Keith E. 2011. *Rationality and the Reflective Mind*. Oxford: Oxford University Press.

Stanovich, Keith E., and Richard F. West. 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioral & Brain Sciences* 23: 645–65. [CrossRef]

Townsend, David, and Thomas G. Bever. 2001. *Sentence Comprehension: Integration of Habits and Rules*. Cambridge: MIT Press.

Tsiplakou, Stavroula, Spyros Armostis, and Dimitris Evripidou. 2016. Coherence 'in the mix'? Coherence in the face of language shift in Cypriot Greek. *Lingua* 17: 10–25. [CrossRef]

Tsiplakou, Stavroula, Spyros Armostis, Spyridoula Bella, Dimitris Michelioudakis, and Amalia Moser. 2019. He past perfect in Cypriot Greek: Innovation because—Or irrespective—Of contact? In *Selected Papers from the Ninth International Conference on Language Variation in Europe (ICLaVE 9), Malaga, June 2017*. Edited by Juan A. Villena-Ponsoda, Francisco Díaz-Montesinos, Antonio Ávila-Muñoz and Matilde Vida-Castro. Amsterdam: John Benjamins, pp. 233–43.

Wellwood, Alexis, Roumyana Pancheva, Valentine Hacquard, and Colin Phillips. 2009. The role of event comparison in comparative illusions. Poster presented at the CUNY Conference on Human Sentence Processing, University of California, Davis, CA, USA, March 26–28.

Wellwood, Alexis Roumyana Pancheva, Valentine Hacquard, and Colin Phillips. 2018. The anatomy of a comparative illusion. *Journal of Semantics* 35: 543–83. [CrossRef]