On the Need for Experimental Syntax

Ken Ramshøj Christensen
Aarhus University

Abstract
The use of expert intuition as a source of evidence in theoretical syntax has long been criticized. Here I review some of the main points of the debate. Using examples from research done by me and collaborators, I argue that an experimental approach is essential when studying subtle structural contrasts, in particular when doing comparative studies. The same applies to linguistic illusions where people are misled and interpret meaningless nonsense as meaningful. However, without expert intuition, experimental syntax would not get off the ground; it is based on expert intuition and syntactic analysis.

1. Introduction

Over the years, it has been debated whether the use of introspection is a reliable and valid source of data in theoretical syntax (Schütze, 1996). According to Gibson & Fedorenko (2010, 2013), the "standard" methodology in syntax, i.e. introspection in the form of expert intuitions about acceptability or grammaticality, is "weak". However, as Sprouse & Almeida (2017) note in their response to Branigan & Pickering (2017), the claim that this is the "standard" approach "is a caricature of linguistic methodology that, to our knowledge, has never been supported by evidence. Nonetheless, a charitable interpretation of this claim reveals two separate concerns", namely, "the routine use of small sample sizes" and "the susceptibility of [acceptability judgments] to investigator bias". First of all, the contrast in grammaticality or acceptability between two otherwise minimally different sentences may be due to semantic properties

Anne Mette Nyvad, Michaela Hejná, Anders Højen, Anna Bothe Jespersen & Mette Hjortshøj Sørensen (Eds.), ***A Sound Approach to Language Matters – In Honor of Ocke-Schwen Bohn*** (pp. 373-388). Dept. of English, School of Communication & Culture, Aarhus University.

of the individual lexical items (selection bias), rather than to the syntactic phenomenon in question. Multiple instances of the construction in question should be evaluated in order to make generalizations. Secondly, there is high a risk of confirmation bias on the part of the researcher seeking to support (or refute) some hypothesis, and indeed expert intuitions are normally not considered data in other branches of science. Asking a few colleagues or students may also bias the data, because they might be inclined to agree merely because they (more or less) subconsciously want to please the researcher. According to Gibson & Fedorenko (2010, p. 233), "the lack of validity of the standard linguistic methodology has led to many cases in the literature where questionable judgments have led to incorrect generalizations and unsound theorizing, especially in examples involving multiple clauses, where the judgments can be more subtle and possibly more susceptible to cognitive biases". The remedy, they argue, is to adopt a quantitative approach, e.g. by using corpus studies and experiments with multiple items and participants.

While Culicover & Jackendoff (2010) agree that grammaticality judgments should always be made on properly controlled data, they also argue that sometimes, subjective judgments are sufficient and just as good as experimental data. For one, corpus data may not always be very helpful. Certain sentence types, phrases, and words, which people nonetheless have clear intuitions about (a classic example is the parasitic gap), are very rare and may indeed not be found in a corpus, but very little (if anything) can be deduced about the grammatical status of such items from their non-occurrence in a corpus (Newmeyer, 2003).

Furthermore, as I shall argue in detail below, some intuitions are very robust and stable across subjects, including intuitions about grammatical illusions. This is true for language as well as for other cognitive domains, such as vision. Consider the diagrams in Figure 1. There is no need for a large sample of intuitions to ascertain that people consistently see a white triangle (which is not actually there) in the Kanizsa triangle, that the Necker cube is ambiguous (the lower left square is either the front or the back of the transparent box), or that the Devil's tuning fork is an impossible object (once you look closer):

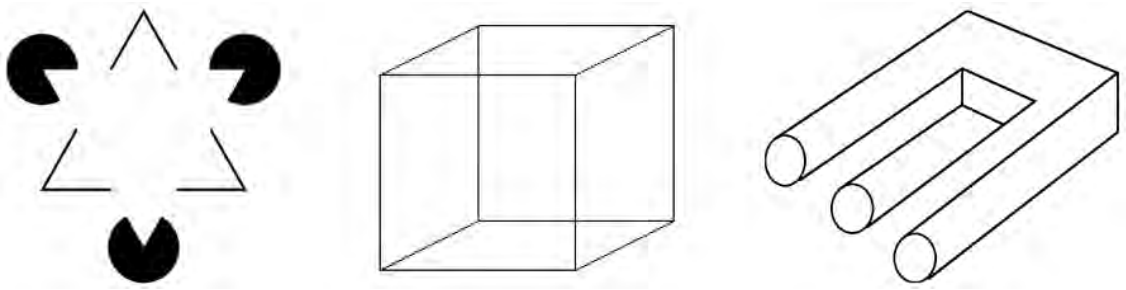

Figure 1. From left to right: the Kanizsa triangle, the Necker cube, and the Devil's tuning fork (also known as an impossible trident or a blivet)

In the same way, subjective judgments can form the basis for theory development, which may inspire experiments; just like optical illusions provide can be used to test visual theories, intuitions about sentences can be used to test grammatical theories (Townsend & Bever, 2001, p. 184). Indeed, "grammaticality judgments are the raw material for hypotheses about the structure of the language faculty. Without such judgments, the experimental enterprise cannot get off the ground" (Culicover & Jackendoff, 2010, p. 234). Along the same lines, Phillips (2009) argues that there is no crisis in theoretical linguistics. Before empirical claims become widely accepted generalizations, they are "scrutinized" by the linguistic community, and the standard methodology in theoretical syntax has not led to "unsound theorizing". In fact, "carefully constructed tests of well-known grammatical generalizations overwhelmingly corroborate the results of 'armchair linguistics'" (Phillips, 2009, p. 53) – in at least 95% of cases, according to Sprouse & Almeida's (2013) analysis of 1743 judgment pairs, but see Gibson et al. (2013). A replication rate of 95% is, to put it mildly, very impressive – far better than that of other sciences, including psychology (39%) and cancer biology (10%), as well as chemistry, physics, and medicine (Baker, 2016; Open Science Collaboration, 2015). Similarly, Featherston (2009, p. 131) argues that quantitative data and statistical analyses are indeed powerful tools, "but still just tools", which "produce a quantitative measure of how well some data supports our hypotheses". He suggests that "linguists use data and apply statistical tests, but do not forget that both the starting point and the end point of a study must be a grammatical analysis".

We should be methodologically tolerant because subjective introspection and experimental methods corroborate each other with an impressive level of convergence. A recent similar debate about whether syntactic priming is superior to and should replace acceptability judgments, or whether the two (and other) methodologies in fact supplement each

other can be found in the target paper by Branigan & Pickering (2017) and the many open peer commentaries, e.g. Adger (2017), Ambridge (2017), Hagoort (2017), Sprouse & Almeida (2017).

So, should we just "relax, lean back, and be a linguist" (Featherston, 2009)? Well, that depends. Although there may not be any real crisis (or at least, no more than in the sciences in general), there is still a serious issue. Gibson et al. (2013) argue that the 95% replication rate reported by Sprouse & Almeida (2013) is inflated due to the inclusion of theoretically irrelevant examples such as those in (1) below (where * means ungrammatical): everyone agrees about their acceptability and as such, they have no bearing on the falsification of hypotheses or on the choice between theories.

(1) a. *Was kissed John
b. John was kissed.

Like with the Kanizsa triangle in Figure 1, there is actually no need for an experiment or a survey to argue that (1)a is ungrammatical in English, whereas (1)b is completely well-formed; intuitions from the expert in the "armchair" will do. Furthermore, such examples "are not representative of the forefront of syntactic research because all current linguistic theories correctly predict [such] contrasts" (Gibson et al., 2013, p. 3).

However, even with an acceptable error rate of 5%, as is the norm in psychology and social science in general (reflected in the standard threshold of statistical significance, $p<0.05$), non-quantitative methods have no means of discovering what the errors are and correcting them. Behavioral, quantitative studies are required to test whether the subjective intuitions match reality. Furthermore, the more judgment pairs (intuitions) from a single speaker in a paper, the higher the risk of errors and an increasing uncertainty about what the data is. Assuming 5% error in a set of 1743 judgment pairs (Sprouse & Almeida, 2013; Sprouse, Schütze, & Almeida, 2013), 87 will be incorrect. That may not sound as a lot, but according to Gibson et al. (2013), in such large data set, there are $5.26*10^{148}$ possible ways of 5% being wrong (choosing 87 from 1743). A truly "unfathomable" number (Gibson et al., 2013, p. 233) – even when compared to the number of fundamental particles in the observable universe: 10^{80} (Mastin, 2018), or to the much smaller number of stars: 10^{22} (ESA, 2016). Even in a (short) book with a mere 100 example pairs, the number of ways of having 5% errors (5% 'wrong' subject/expert intuitions) is larger than 75 million. However, the findings reported by Sprouse & Almeida (2013) have been replicated by Mahowald, Graff, Hartman, & Gibson (2016) who also suggest an experimental method which makes it possible to make statistically valid

generalizations about acceptability from a very small sample. However, Mahowald et al. (2016, pp. 630-631) emphasize that their method requires clear (contrasts in) acceptability judgements from the researcher based on "informal investigation" and that "statistics should supplement, not replace, careful thought about syntax and semantics".

The message here is that with complex theories such as generative grammar, there is a need for a very high degree of reliability, and fine-grained syntactic contrasts of theoretical importance call for quantitative experimental methods. In this paper, I will illustrate the need for experimental syntax using work by myself and my collaborators.

2. Escapable islands

A syntactic island is a configuration that blocks extraction (Chomsky, 1986, 1995; Hofmeister & Sag, 2010; Rizzi, 1990; Ross, 1967; Sprouse & Hornstein, 2013). They are 'inescapable' (or at least difficult to escape from) in the sense that phrases cannot be moved out of them; they are 'marooned' (in somewhat the same sense that a pirate marooned on a deserted island cannot escape).

One famous example is the wh-island blocking extraction from a complement clause (Christensen, Kizach, & Nyvad, 2013a). As shown in (2), it is fully acceptable to have an embedded object question or an embedded adjunct question; the wh-element undergoes (short) movement to the left edge of the embedded clause. It is also possible to extract the question element from the embedded clause into the matrix clause, as shown in (3) where the wh-element undergoes long movement: from the base position to the edge of the embedded clause and then to the edge of the matrix clause. Crucially, long movement proceeds in short local incremental steps. When the two types of extraction (short and long movement) are combined, as in (4), problems arise because the long movement cannot take place in short local steps, as illustrated in Figure 2.

(2) I know [she can solve the problem in this way].
   a. I know [which problem$_1$ she can solve $\_\__1$ in this way].
   b. I know [how$_1$ she can solve this problem $\_\__1$].

(3) a. Which problem$_1$ do you think [$\_\__1$ she can solve $\_\__1$ in this way]?
   b. How$_1$ do you think [$\_\__1$ she can solve this problem $\_\__1$]?

(4) a. ?Which problem$_1$ do you wonder [how$_2$ she can solve $\_\__1$ $\_\__2$]?
   b. *How$_2$ do you wonder [which problem$_1$ she can solve $\_\__1$ $\_\__2$]?

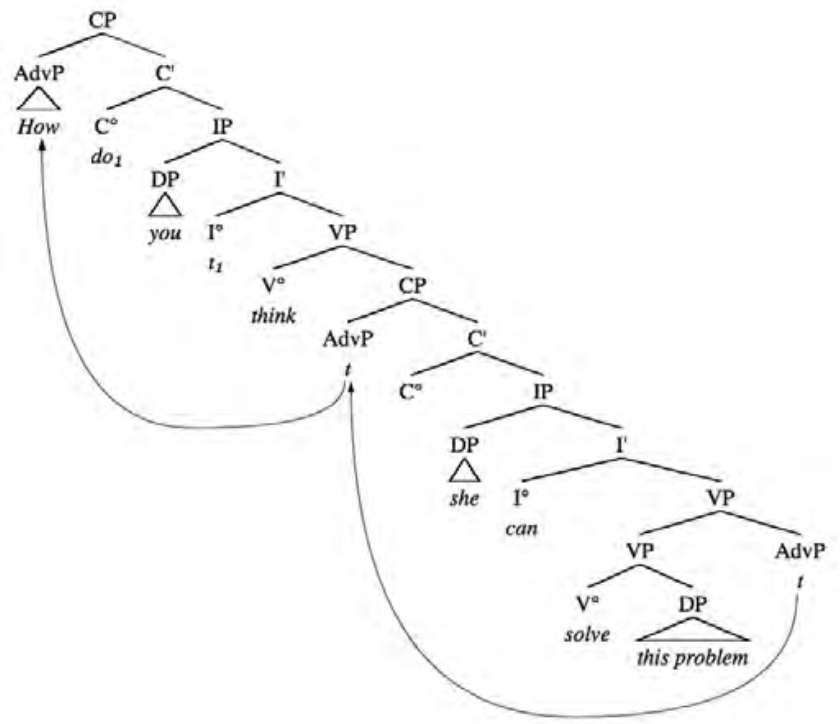

Figure 2.1. The syntactic structure of (3)b. Note that movement takes place in two successive (local) steps. Right: The syntactic structure of the ungrammatical (4)b. Here, long movement is not acceptable because it has to skip the position occupied by 'which problem'.

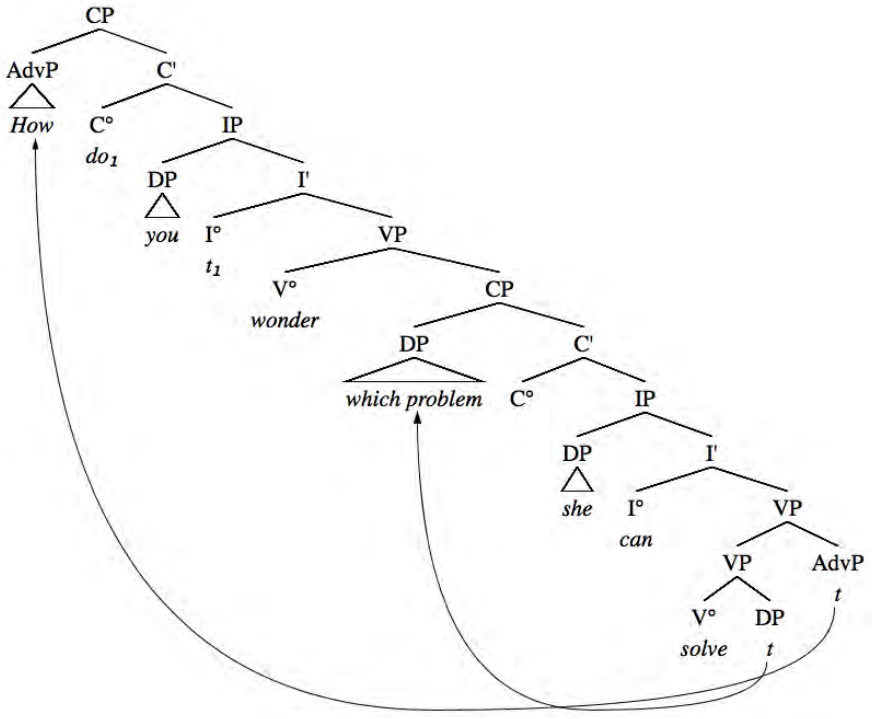

Figure 2.2. The syntactic structure of the ungrammatical (4)b. Here, long movement is not acceptable because it has to skip the position occupied by 'which problem'.

In (4)a, moving the embedded wh-object (*which problem*) across the wh-adjunct (*how*) is unacceptable (sometimes the diacritics say ?? or ?* indicating even lower levels of acceptability), if not fully ungrammatical. In (4)b, extracting the embedded wh-adjunct across the extracted wh-object is completely ungrammatical. The asymmetry in (4) is standardly assumed to be universal. It is indeed possible to find Danish examples that, at least to some speakers, match the asymmetry in (4), see e.g. Vikner (1995, p. 19). However, as shown by Christensen, Kizach & Nyvad (2013a; 2013b), it does not seem to hold in general for Danish.

In our studies, which involved three acceptability judgement experiments with multiple participants (60, 32, and 30), multiple different sentence tokens per condition (16, 12, and 16), and different scales of acceptability (two with a 5-point Likert-scale, one with a binary one), we found no statistically significant difference between the two island violations illustrated in (4). (Note that we replicated our initial results twice.) That is, the acceptability of the sentence pair in (5) is symmetric (they are equally acceptable), unlike the English structurally equivalent pair in (4) (asymmetric acceptability). People also found both significantly better than clearly ungrammatical control sentences.

(5) a. ?Hvad$_1$ ved hun godt [hvor$_2$ man kan leje __$_1$ __$_2$]?
What knows she well where one can rent?
"What does she know where you can rent?"

b. ?Hvor$_2$ ved hun godt [hvad$_1$ man kan leje __$_1$ __$_2$]?
Where knows she well what one can rent?
"Where does she know what you can rent?"

Extraction from a relative clause, as in (6)b below, is also assumed to be universally blocked due to the Complex Noun Phrase Constraint (Phillips, 2013; Ross, 1967):

(6) a. She wanted to meet the man [who recorded the conversation]?
b. *What$_1$ did she want to meet the man [who recorded __$_1$]?

Essentially, the problem is the same as illustrated in Figure 2: *who* in the relative clause blocks successive local movement of *what*. Though extractions from relative clauses have (famously) been reported to be acceptable in the Scandinavian languages (Engdahl, 1997; Engdahl & Ejerhed, 1982; Erteshik-Shir, 1973), such counter examples have been

argued to be merely 'apparent' counter examples; they do not involve extraction from relative clauses but from a different structure altogether, namely, small clauses (Kush & Lindahl, 2011; Kush, Omaki, & Hornstein, 2013). However, our experiment (Christensen & Nyvad, 2014), using examples such as (7), supports the idea that extractions from relative clauses are in fact grammatical in Danish, and that they are not merely apparent counter examples involving extractions from small clauses. This has subsequently also been shown for Swedish (Müller, 2015).

(7) a. Pia har engang set/mødt en pensionist [som/der havde sådan en hund].
Pia has once seen/met a pensioner COMP had such a dog
"Pia once met a pensioner who had such a dog."

b. Sådan en hund$_1$ har Pia engang set/mødt en pensionist [som/der havde __$_1$].
Such a dog has Pia once seen/met a pensioner COMP had
"Such a dog Pia once met a pensioner who had."

In our experiment (acceptability judgement on a 7-point Likert scale, 112 participants, 16 sentence tokens per condition) showed that the level of acceptability of extractions such as (7)b is highly dependent on the choice of matrix verb. The higher the frequency of usage of the main verb (measured as the number of occurrences in the online Danish corpus, KorpusDk), the more acceptable it is to have extraction from the relative clause inside the object. Consequently, the contrast in acceptability depends on lexical properties of the main verb, not on the construction as such. (It is also very easy to make a simple, fully acceptable sentence much less acceptable simply by using rare or less frequent words, compare *This man bought a new hat for his son* and *The gentleman purchased a novel bonnet for his offspring*.) This experiment also shows that it is important to include not only multiple participants but also multiple different tokens for each condition to avoid lexical confounds.

In short, Danish allows extraction from embedded questions, which are normally considered to be universally ungrammatical, and there is no argument–adjunct asymmetry in the extractions, also considered to be universal. Similarly, Danish allows extraction from relative clauses, also normally considered to be universally ungrammatical. These extraction patterns have serious implications for syntactic theory in general and for the syntactic theory of Danish in particular as they suggest a parametric difference between the two languages (Nyvad, Christensen, & Vikner, 2017;

Vikner, Christensen, & Nyvad, 2017) – "some islands have bridges that allow elements to escape, and this seems to be the case in the Scandinavian languages in particular" (Christensen & Nyvad, 2014, p. 42). Basically, the embedded CP layer in the tree in Figure 2 can be recursive in Danish but not in English, which also accounts for other independently observed phenomena (including stacked complementizers in Danish, e.g. *fordi at* 'because that'). But to see these effects and to avoid wrong generalizations, we need careful experiments and quantitative analyses. It is not clear how it could have been done without experimental syntax.

3. From the borderlands of understanding

In this section, I argue that that quantitative intuition data can be used to address otherwise counter-intuitive interpretations of so-called linguistic illusions. While it is intuitively true that language usually makes sense, that it is usually meaningful, it is not always true. During parsing (the incremental construction of a syntactic representation in language comprehension), we sometimes make intermediate, semantically anomalous interpretations. In (8), for example, we initially and temporarily interpret *where* as a modifier of the matrix verb *believe* (this is called 'early attachment'), even though it is a very unlikely and strange interpretation (*??Where did she believe? In the kitchen*). The extracted element is not compatible with the matrix verb. Subsequently, after encountering the rest of the sentence, we reanalyze it as modifying the embedded verb phrase (*buried the cat where?*). Despite the fact that sentences such as (8) are unambiguously grammatical, native speakers judge them as less than fully acceptable. Matrix verb incompatibility reduces acceptability (Christensen et al., 2013a; Fanselow & Frisch, 2006). (Here, it also seems difficult, though perhaps not impossible, to establish the systematic relationship between matrix verb incompatibility and reduced acceptability without an experimental approach.)

(8) Where did she believe that he had buried the cat?

Now, compare the two sentences in (9). In our experiment (Kizach, Nyvad, & Christensen, 2013) (60 participants, 16 different sentences, self-paced reading), we found that people initially attached *the pig in the pen* as the object of *noticed* and then reanalyzed it as the subject of *needed water* in (9)a. The matrix verb *notice* is compatible with either a nominal or a clausal object. In (9)b, on the other hand, people did not initially attach *the pig in the pen* as the object of *presumed*, because *presume* requires a clausal object.

(9) a. Alice noticed the pig in the pen needed water.
b. Alice presumed the pig in the pen needed water.

Crucially, though, such counter-intuitive interpretations are only made if they do not violate the syntactic structure. In other words, because the syntax of the verb dictates that the object must be a clause, we do not make strange semantic interpretations. If on the other hand, the syntax allows for it, we do make strange temporary interpretations that affect the overall acceptability.

We can even be systematically tricked by certain syntactic constructions, sometimes into believing that certain sentences that are meaningless are actually meaningful. Because people disagree on the interpretations as well as on the acceptability of such examples, these counter-intuitive findings are only accessible with an experimental quantitative approach. Compare (10) and (11). While there is no doubt that (10) is ambiguous between meaning either that she used the bag to hit him with, or that she hit the bag-carrying man, people disagree on the interpretation of (11), which in fact does not have one.

(10) She hit the man with the bag.

(11) More people have been to Paris than I have.

(11) is a so-called comparative illusion (Phillips, Wagers, & Lau, 2011) or dead end (Christensen, 2010, 2016); see also Townsend & Bever (2001, p. 184) and Saddy & Uriagereka (2004).[1] At first sight, (11) seems to be elliptical; something has been left out after *than I have*, like in (12) where *have been* is elided (i.e. is not repeated) between *than* and *Copenhagen*; (12) means *More people than have been to Copenhagen have been to Paris*, where the *than* phrase is reconstructed in the middle of the sentence.

(12) More people have been to Paris than to Copenhagen.

If the same procedure is applied to (11), the result is seriously anomalous or incongruous: **More people than I have been to Paris have been to Paris*.

[1] The earliest mentioning (but not analysis) of this illusory construction that I know of is Montalbetti (1984, p. 6): "To Hermann Schultze, my eternal gratitude for uttering the most amazing */?sentence I've ever heard: More people have been to Berlin than I have. (Some have taken this sentence to be a proof of the autonomy of syntax!)".

The sentence types in (10) and (11) are linguistic versions of the Necker cube and the Devil's tuning fork in Figure 1, respectively: The former is structurally ambiguous, the latter is globally incongruous or impossible.

I have investigated how people interpret sentences such as (11) in a series of studies, including an fMRI study (speeded acceptability, participants: n=19) (Christensen, 2010), an informal questionnaire (n=63) (Christensen, 2011), and an internet survey (multiple choice task, n=545) and two experiments (speeded acceptability, n=32 and 60) (Christensen, 2016). The results consistently showed that many people are tricked by the illusion and find sentences such as (11) meaningful. However, they do not agree on the interpretation. Interestingly, people seem to choose from a small set of mutually incompatible interpretations: 'Some people except me have been to Paris', 'More people than just me have been to Paris', or 'Some people have been to Paris more often than I have' – or they say that it is indeed meaningless. They do not find it ambiguous. This situation is very different from the one for (10), which people agree is meaningful and ambiguous.

Another type of illusion where people are systematically tricked is the so-called depth charge sentence (Kizach, Christensen, & Weed, 2016; Natsopoulos, 1985; Wason & Reich, 1979). Consider (13):

(13) No head injury is too trivial to be ignored.

Most people say that (13) means the same as (14), which is impossible. To ignore and to treat are definitely not the same, and in some contexts, they are opposites.

(14) No head injury is too trivial to be treated.

In our experiment, which included 19 participants and 150 sentences (moving window reading task), we manipulated three factors that together give rise to the depth charge effect: the number of negations ((13) has three: *no, trivial [=not important]*, *ignore [=not attend to]*), the plausibility of the relation between the subject and the verb (*head injury* and *be ignored*: not plausible), and the logic of the relation between the adjective and the verb (the more *trivial* the less we *ignore*: illogical). When a sentence is maximally complex (i.e., when there are multiple negations, the relation between subject and verb is implausible, and the relation between adjective and verb is illogical), the majority of the participants misunderstood the sentence to mean the same as (14), but were at the same time certain of their

answers. Given that people have strong opinions about the interpretation, some have argued that their interpretation is true. And who am I to tell them otherwise? However, how can (13) and (14) be synonymous? As the experiment shows, the interpretation differs systematically, plus manipulating the three factors, leads to predictable increases in error rates. Again, these findings would not be possible without experiments and quantitative data. (Our study also confirms the two previous studies of the phenomenon, again showing a high replication rate for linguistic studies.)

4. Conclusions

A sound approach that avoids "unsound theorizing" due to bias and secures a high degree of reliability and validity is experimental and quantitative. When people disagree significantly on the level of acceptability or grammaticality, or on the interpretation, an experimental quantitative approach is indeed required. Otherwise, it is difficult (if not impossible) to know what the data actually is or to detect whether or not the reported acceptability or interpretation is indeed real. This is particularly important with subtle distinctions of theoretical importance, such as the status of island violations, which are used to argue for universal properties of and constraints on human language. Likewise, when people disagree on acceptability and interpretation of linguistic illusions, we need experiments in order to determine the 'borderlands' of linguistic comprehension and to discover how linguistic processing interacts with general cognition. Finally, it should be kept in mind that none of the experimental findings discussed above (or elsewhere for that matter) would be possible without subjective intuition about acceptability and interpretation. Without expert intuition, the experimental enterprise would not get off the ground.

Acknowledgements

I would like to thank Ocke-Schwen Bohn for many interesting discussions about the validity of introspection and the nature of science. Thanks also to Anne Mette Nyvad, Sten Vikner, and Johannes Kizach, and to the anonymous reviewer for constructive criticism.

References

Adger, D. (2017). The limitations of structural priming are not the limits of linguistic theory. *Behavioral and Brain Sciences*, *40*, 18-19 (e283). http://dx.doi.org/10.1017/S0140525X17000310

Ambridge, B. (2017). Horses for courses: When acceptability judgments are more suitable than structural priming (and vice versa). *Behavioral and Brain Sciences*, *40*, 19 (e284). https://doi.org/10.1017/S0140525X17000322

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, *533*(7604), 452-454. https://doi.org/10.1038/533452a

Branigan, H. P., & Pickering, M. J. (2017). An experimental approach to linguistic representation. *Behavioral and Brain Sciences*, *40*. https://doi.org/10.1017/S0140525X16002028

Chomsky, N. (1986). *Barriers*. Cambridge, Mass: MIT Press.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, Mass.: MIT Press.

Christensen, K. R. (2010). Syntactic reconstruction and reanalysis, semantic dead ends, and prefrontal cortex. *Brain and Cognition*, *73*(1), 41-50. https://doi.org/10.1016/j.bandc.2010.02.001

Christensen, K. R. (2011). Flere folk har været i Paris end jeg har. In Hansen, Inger Schoonderbeek & Widell, Peter (Eds.), *13. Møde om Udforskningen af Dansk Sprog* (pp. 113-136). Aarhus: Nordic Department, Aarhus University. Retrieved from http://auinstallation42.cs.au.dk/fileadmin/projekter/Muds.dk/rapporter/MUDS_13.pdf

Christensen, K. R. (2016). The dead ends of language: The (mis)interpretation of a grammatical illusion. In S. Vikner, H. Jørgensen, & E. Van Gelderen (Eds.), *Let us have articles betwixt us – Papers in Historical and Comparative Linguistics in Honour of Johanna L. Woo*d (pp. 129-160). Aarhus: Dept. of English, School of Communication & Culture, Aarhus University. Retrieved from http://ebooks.au.dk/index.php/aul/catalog/download/119/107/466-1?inline=1

Christensen, K. R., Kizach, J., & Nyvad, A. M. (2013a). Escape from the Island: Grammaticality and (Reduced) Acceptability of wh-island Violations in Danish. *Journal of Psycholinguistic Research*, *42*(1), 51-70. https://doi.org/10.1007/s10936-012-9210-x

Christensen, K. R., Kizach, J., & Nyvad, A. M. (2013b). The processing of syntactic islands – An fMRI study. *Journal of Neurolinguistics*, *26*(2), 239-251. https://doi.org/10.1016/j.jneuroling.2012.08.002

Christensen, K. R., & Nyvad, A. M. (2014). On the nature of escapable relative islands. *Nordic Journal of Linguistics*, *37*(01), 29-45. https://doi.org/10.1017/S0332586514000055

Culicover, P. W., & Jackendoff, R. (2010). Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences*, *14*(6), 234-235.

Engdahl, E. (1997). Relative clause extraction in context. *Working Papers in Scandinavian Syntax*, *60*, 59-86.

Engdahl, E., & Ejerhed, E. (Eds.). (1982). *Readings on unbounded dependencies in Scandinavian languages*. Umeå; Stockholm, Sweden: Universitetet i Umeå; Almqvist & Wiksell International [distributor].

Erteshik-Shir, N. (1973). *On the nature of island constraints* (PhD dissertation). Massachusetts Institute of Technology.

ESA. (2016). How many stars are there in the Universe? Retrieved November 11, 2016, from http://www.esa.int/Our_Activities/Space_Science/Herschel/How_many_stars_are_there_in_the_Universe

Fanselow, G., & Frisch, S. (2006). Effects of processing difficulty on judgments of acceptability. In G. Fanselow, C. Fery, & M. Schlesewsky (Eds.), *Gradience in grammar: Generative perspectives* (pp. 291-316). Oxford: Oxford University Press.

Featherston, S. (2009). Relax, lean back, and be a linguist. *Zeitschrift Für Sprachwissenschaft*, *28*(1), 127-132. https://doi.org/10.1515/ZFSW.2009.014

Gibson, E., & Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, *14*(6), 233-234. https://doi.org/10.1016/j.tics.2010.03.005

Gibson, E., & Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. Language and Cognitive Processes, 28(1-2), 88-124. https://doi.org/10.1080/01690965.2010.515080

Gibson, E., Piantadosi, S. T., & Fedorenko, E. (2013). Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). *Language and Cognitive Processes*, *28*(3), 229-240. https://doi.org/10.1080/01690965.2012.704385

Hagoort, P. (2017). Don't forget the neurobiology: An experimental approach to linguistic representation. *Behavioral and Brain Sciences*, *40*, 26 (e292). https://doi.org/10.1017/S0140525X17000401

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, *86*(2), 366-415. https://doi.org/10.1353/lan.0.0223

Kizach, J., Christensen, K. R., & Weed, E. (2016). A Verbal Illusion: Now in Three Languages. *Journal of Psycholinguistic Research*, *45*(3), 753-768. https://doi.org/10.1007/s10936-015-9370-6

Kizach, J., Nyvad, A. M., & Christensen, K. R. (2013). Structure before Meaning: Sentence Processing, Plausibility, and Subcategorization. *PLoS ONE*, *8*(10), e76326. https://doi.org/10.1371/journal.pone.0076326

Kush, D., & Lindahl, F. (2011). On the escapability of islands in Scandinavian. In *85th Annual Meeting of the Linguistic Society of America*. Pittsburgh, Pennsylvania. Retrieved from http://ling.umd.edu/~kush/KushLindahl_LSA_ScandinavianExtraction.pdf

Kush, D., Omaki, A., & Hornstein, N. (2013). Microvariation in islands? In J. Sprouse & N. Hornstein (Eds.), *Experimental Syntax and Island Effects* (pp. 239-264). Cambridge: Cambridge University Press.

Mahowald, K., Graff, P., Hartman, J., & Gibson, E. (2016). SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language*, *92*(3), 619-635. https://doi.org/10.1353/lan.2016.0052

Mastin, L. (2018). The Universe By Numbers – The Physics of the Universe. Retrieved May 3, 2018, from https://www.physicsoftheuniverse.com/numbers.html

Montalbetti, M. M. (1984). *After binding : on the interpretation of pronouns*.

Massachusetts Institute of Technology, Dept. of Linguistics and Philosophy.
Müller, C. (2015). Against the Small Clause Hypothesis: Evidence from Swedish relative clause extractions. *Nordic Journal of Linguistics*, *38*(01), 67-92. https://doi.org/10.1017/S0332586515000062
Natsopoulos, D. (1985). A verbal illusion in two languages. *Journal of Psycholinguistic Research*, *14*(4), 385-397. https://doi.org/10.1007/BF01067882
Newmeyer, F. J. (2003). Grammar is grammar and usage is usage. *Language*, *79*(4), 682-707. https://doi.org/10.1353/lan.2003.0260
Nyvad, A. M., Christensen, K. R., & Vikner, S. (2017). CP-recursion in Danish: A cP/CP-analysis. *The Linguistic Review*, *34*(3), 449-477. https://doi.org/10.1515/tlr-2017-0008
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716
Phillips, C. (2009). Should we impeach armchair linguists? In S. Iwasaki, H. Hoji, P. M. Clancy, & S.-O. Sohn (Eds.), *Japanese/Korean Linguistics 17* (pp. 49-64). Stanford University: CSLI Publications. Retrieved from http://www.colinphillips.net/wp-content/uploads/2014/08/phillips2010_armchairlinguistics.pdf
Phillips, C. (2013). Some arguments and nonarguments for reductionist accounts of syntactic phenomena. *Language and Cognitive Processes*, *28*(1-2), 156-187. https://doi.org/10.1080/01690965.2010.530960
Phillips, C., Wagers, M. W., & Lau, E. F. (2011). 5 Grammatical Illusions and Selective Fallibility in Real-Time Language Comprehension. In J. T. Runner (Ed.), *Experiments at the Interfaces* (pp. 147-180). Bingley: Emerald Group Publishing.
Rizzi, L. (1990). *Relativized minimality*. Cambridge, Mass: MIT Press.
Ross, J. R. (1967). *Constraints on variables in syntax*. Massachusetts Institute of Technology.
Saddy, D., & Uriagereka, J. (2004). Measuring language. *International Journal of Bifurcation and Chaos*, *14*(02), 383-404.
Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago, Il: University of Chicago Press.
Sprouse, J., & Almeida, D. (2013). The empirical status of data in syntax: A reply to Gibson and Fedorenko. *Language and Cognitive Processes*, *28*(3), 222-228. https://doi.org/10.1080/01690965.2012.703782
Sprouse, J., & Almeida, D. (2017). Setting the empirical record straight: Acceptability judgments appear to be reliable, robust, and replicable. *Behavioral and Brain Sciences*, *40*, 43-44 (e311). https://doi.org/10.1017/S0140525X17000590
Sprouse, J., & Hornstein, N. (Eds.). (2013). *Experimental syntax and island effects*. Cambridge; New York: Cambridge University Press.
Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and

formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. *Lingua*, *134*, 219-248. https://doi.org/10.1016/j.lingua.2013.07.002

Townsend, D. J., & Bever, T. G. (2001). *Sentence Comprehension. The integration of habits and rules*. Cambridge, Mass: MIT Press.

Vikner, S. (1995). *Verb Movement and Expletive Subjects in the Germanic Languages*. Oxford University Press.

Vikner, S., Christensen, K. R., & Nyvad, A. M. (2017). V2 and cP/CP. In L. Bailey & M. Sheehan (Eds.), *Order and structure in syntax I: Word order and syntactic structure* (pp. 313-324). Berlin: Language Science Press. Retrieved from http://langsci-press.org/catalog/view/159/1125/971-1

Wason, P. C., & Reich, S. S. (1979). A verbal illusion. *Quarterly Journal of Experimental Psychology*, *31*(4), 591-597. https://doi.org/10.1080/14640747908400750