Forthcoming in *Proceedings of CLS 47*, 2011

# Two puzzles in experimental syntax and semantics

Yosef Grodzinsky
McGill University[1]

## 1 Two puzzles

Relating standard linguistic data to theory is hard; making experimental data bear on the theory is even harder. Two difficult problems are central, one relating to experimental design, and the other, to the interpretation of results. In this short note, I discuss these problems in the context of specific studies in experimental syntax and semantics.

I begin with an experimental design issue – THE DIMENSIONALITY-REDUCTION PROBLEM, arising in the midst of complex stimuli: theoretical linguists typically investigate the grammatical properties of complex linguistic objects through contrasts. As minimal pairs are hard to come by, multiple controls are introduced. This is done at no additional cost. Experiments, however, are usually subject to severe logistical, technological and analytic constraints that preclude the proliferation of controls. When we work with multi-dimensional contrasts, our task is to design compact tests that reduce dimensionality to a bare minimum, but still ascertain that the property that interests us is isolated. But when complex stimuli are compared along some dimension, and given a limited room for controls, how can other dimensions of the stimuli be made orthogonal to the test, namely, precluded from affecting the dependent variable? This problem is pronounced in the context of functional Magnetic Resonance Imaging (fMRI) language experiments, although it also arises when Reaction Time (RT) is the dependent measure. I describe the puzzle, sketch a solution – parametric designs – and illustrate it with 2 experiments, where syntactic or semantic properties of stimuli modulate brain responses.

I then move on to data analysis, and discuss THE MAPPING PROBLEM: most formal linguistic theories are designed to handle *categorical* variables, placed on a nominal scale, and carrying labels (e.g., ±well-formed, ±Grammatical) rather than numerical values. How can such variables be related to *ordinal* ones – numerical values on a scale, e.g., the time it takes to parse a sentence, or to changes in the intensity of some brain signal that their processing modulates? Experimental results presumably give us a special angle on linguistic knowledge and its implementation. But how do we incorporate these results into the theory? This problem was considered early on in the history of generative grammar (Miller & Chomsky, 1963), and in much subsequent work, mostly in the context of Reaction Time data. Interest in it has been recently renewed, as quantitative analyses of judgment behavior raised new questions (Bard, Robertson & Sorace, 1996, *passim*). Below, I consider how this problem arises for other types of experimental data (mostly errors by linguistically deficient individuals, whether aphasic patients suffering from focal brain-damage, or normally or pathologically developing children). Here, too, I illustrate with several concrete syntactic and semantic examples, and reflect on possible solutions, and on how they impact the way one should bring experimental results to bear on the theory.

[1] Yosef.grodzinsky@mcgill.ca. This work is supported by Canada Research Chairs, a SSHRC Standard Grant, and NIH Grant 00094. The continued funding of these agencies is gratefully acknowledged. Thanks also to McGill undergraduate and graduate students in a Special Topics class taught in Fall 2011, as well as to Bernhard Schwarz and Michael Wagner, whose help is much appreciated.

## 2 Dimensionality reduction in experimental design

The growth in abstractness and generality of linguistic theory has made its connection to empirical phenomena more complex and difficult to establish. If early on, plain minimal pairs sufficed as the bread-and butter of a linguist's daily toil, nowadays, facts are much more complex, and relating them to a theory is more subtle. This complexity makes the life of experimenters more difficult: if the linguist requires multiple contrasts in order to characterize a grammatical property, then it is difficult to see how an experiment that looks for reflections of that property in other domains can make do with less. This creates experimental complications, as some testing methods have limitations, and impose their own constraints on the type and quantity of material that can be used in an experiment. The conflict between the need for multiple contrasts (or dimensions) in order to home in on a property of interest, and limitations imposed by the technology that the experimenter uses, forces the invention of new test paradigms. In this section, I discuss THE DIMENSIONALITY-REDUCTION PROBLEM as it arises in the context of fMRI experiments on language. The need for dimensionality reduction arises when the property of interest can only be isolated by multiple controls, whose introduction would render the experiment unmanageable. Below, I describe how this problem arises, discuss some examples, and consider a solution that is intimately connected to the Mapping Problem discussed in section **3**, because it helps us connect continuous variables such as brain activity to categorical linguistic variables that are of interest to us. I then show how this solution is implemented in experiments that measure localized brain activity during tasks that require syntactic and semantic analysis, whose results I report and proceed to briefly ponder their theoretical significance.

### 2.1 The BOLD response in fMRI, and how linguistic stimuli modulate it

Functional Magnetic Resonance Imaging is a sensitive imaging technology that can record local magnetic changes induced by an object immersed in a magnetic field. In our case, this object is a human brain whose owner is exposed to different stimuli in a systematic fashion. Changes in stimulation induce changes in neuronal activity in loci entrusted with the relevant function (as determined by the stimulus contrast); increased neural activity results in increased flow of oxygenated blood to the active regions. The ratio between oxygenated and deoxygenated blood changes locally as a consequence. Oxygen is transported to cells by the large hemoglobin molecule, whose magnetic properties when oxygenated or deoxygenated are different, in fact opposite. Changes in the concentration of the two forms of hemoglobin are the response that the MRI instrument detects, known as the Blood-Oxygen-Level-Dependent (BOLD) response. Functional MR imaging-based experiments exploit this property, and present different input types to the brain, each of which is expected to tax a given region differentially. This manipulation is expected to increase neuronal activity in the regions entrusted with the function at issue, which in turn induces a change in $\frac{deoxyHb}{OxyHb}$ ratio, thereby modulating the regional BOLD signal. Comparing the BOLD response recorded in a region for 2 or more different stimuli therefore provides an indication of the degree to which the contrast between these stimuli modulates this region, i.e., its computation depends on neurons there. Measuring the BOLD response during the administration of sets of contrastive stimuli, and analyzing subsequent local BOLD contrasts, then, is at the heart of fMRI experimentation (Huettel et al., 2009). The instrument detects these changes at a fairly

high spatial resolution (though at a lower temporal resolution). The imaged brain is divided into voxels, small cubic volume units, each with a unique address in brain space. Signal intensity is sampled at each voxel repeatedly. Analyses compare intensities obtained for each condition within each voxel (or voxel cluster)[2]. When a lexical or syntactic manipulation is reported to have activated a particular brain region in fMRI, what the authors are saying is that the difference in the BOLD signal intensity between 2 conditions in a sufficiently large cluster of voxels located in this region differentiated the 2 types of stimuli that formed the experimental manipulation. fMRI can thus differentiate and localize the computation of linguistic properties and relations, and to test linguistic generalizations in terms of BOLD *signal location* and *signal intensity*. Typically, then, analyses "subtract" mean intensities in each voxel from one another, to test if given the variability observed, we can conclude that one effect mean is higher than the other.

## 2.2 Restricted stimulus sets and multiple comparisons

Linguists contrast grammatical and ungrammatical strings in order to establish the boundary of the domain of application of a grammar. But neurolinguists, interested in neural mechanisms that support grammar, are forced to operate somewhat differently. To illustrate, the primary data for syntactic theory are typically <string, label> pairs – sentences paired with judgments elicited from people, e.g., marked with an asterisk for ungrammaticality or left unchecked as they are well formed. The label thus typically takes 2 values (where the relevant category, or dimension, is marked ±). In experimentation, however, the format of data is different. Experimental tasks take linguistic objects and pair them up with quantitative measures (success, time, BOLD signal intensity at a particular brain voxel). The results of such experiments effectively add a dimension to the primary data, resulting in triplets <string, label, quantity> (e.g., <*John loves Mary*, +Grammatical, *n*>), where *n*, the quantity at issue – the dependent measure – is determined by the task, the paradigm, etc.

A concrete attempt to convert a linguistic contrast into an fMRI test, might help: it has been known since at least May (1977) that in some instances, multiple quantifiers in a sentence scope over one another in a manner that obeys island constraints – a typical diagnostic for syntactic movement. Sentence (1a) allows Wh-movement (1b); it is also ambiguous, allowing both surface (1c) and inverse scope (1d) of the 2 quantifiers. By comparison, (2) shows how an island constraint blocks both Wh-movement (2b) and inverse scope (2d). Only surface scope is allowed, leaving (2a) unambiguous – only one nurse told the story in (2a). This familiar correlation has been taken to support the conclusion that quantifiers move abstractly (Quantifier Raising at LF):

(1) a. Some nurse wants to dance with every patient
    b. [Which patient]$_i$ does [some nurse] want to dance with $t_i$?
    c. [Some nurse] wants to dance with [every patient]
    d. [Every patient]$_i$ [some nurse] wants to dance with $t_i$

(2) a. Some nurse who danced with every patient fell.
    b. *[Which patient] did some nurse who danced *t* fell?
    c. [Some nurse] who danced with [every patient] fell
    d. *[every patient]$_i$ [Some nurse] who danced with $t_i$ fell

[2] Alternatively, one can try to identify patterns of voxel activation all over the brain, without alluding to the particular time-series observed in a single voxel.

Expressed in experimental terms (i.e., operationalized), this paradigm translates into a 2*x*2*x*2 design, with *Island* (+, –), *Movement* (+, –) and *QR* (+, –) as factors with 2 levels (values) each. This results in 8 unique treatment combinations that correspond to the 8 examples in (1)-(2), organized in Table 1:

| | | ***Movement*** | | ***QR*** | |
|---|---|---|---|---|---|
| | | – | + | – | + |
| ***Island*** | – | (1a) | (1b) | (1c)=(1a) | (1d) |
| | + | (2a) | (2b) | (2c) | (2d) |

Table 1: experimental design for a Movement/QR test

The experiment seeks to test for *Movement* by *Island* as well as *QR* by *Island* interactions, an experimentalist's way of saying that movement (overt or covert) outside of an island affects the dependent variable – Grammaticality[3]. Linguistically, we can observe this interaction with no quantitative analysis: movement out of an island – both overt, of an Wh-phrase and covert, of a QNP – changes the sign of the dependent variable, Grammaticality, from "+" to "–", findings which point to the conclusion that overt and covert movement correlate, and give rise to the theoretical claim that QR is a species of syntactic movement.

In standard experiments, by contrast, the dependent variable is ordinal, in this case even continuous. Instead of a switch in grammatical status, we expect a change in the value of the continuous variable, detectable through statistical analyses.

Can we use fMRI to learn something about the relation between covert and overt movement? Perhaps, but it is difficult. To see that consider an attempt to design an fMRI test of the paradigm in (1)-(2). The first thought would be to use a grammaticality judgment task, and measure BOLD response as participants perform it. Evidence from fMRI that converges on the linguistic facts would be obtained when signal intensities would produce a *Movement* by *Island* interaction for both questions and wide scope readings. Crucially, this interaction would be recorded in the same voxel clusters, namely, in the same brain locus (or loci).

How would that work? The most common practice in the field is the "subtraction method": it begins by calculating the mean signal intensity $\bar{I}$ of the *Movement* conditions for each voxel, and then effectively "subtracting" one from the other, i.e., $\bar{I}_{+\text{Mov}} - \bar{I}_{-\text{Mov}}$, followed by testing whether the difference is large enough (given statistical assumptions) to support the conclusion that the 2 means are different. The same procedure is then repeated for the *Island* factor, and finally, an interaction effect is calculated for both the Wh-movement and the QR condition. If this effect is statistically significant, and moreover, if it will be found in the same brain locus (or loci) for both overt and covert movement, we may have obtained neurolinguistic evidence in support of QR. The linguistic test seems to have been successfully converted into an fMRI experiment.

But upon reflection, problems with this experiment arise. Primarily, it is clear that the background assumptions required for the judgment of the Wh-movement and QR cells in Table 1, as well as the tasks performed, are very different: in the former, participants are requested to plainly detect violations of grammaticality; but in the latter,

[3] There may be other expected effects under the QR view, which I now ignore (see Fox, 2003).

they are supposed to reflect on the possibility of interpretations that follow from surface scope and inverse scope orderings of the quantifiers. The tasks are different from one another, and may require different cognitive resources. In addition, there are non-syntactic features of stimuli: string length as measured by utterance time duration, number of words or syllables, phonological features, the presence of multiple quantifiers vs. R-expressions, and the like. The linguistic experiment is (mostly) indifferent to both the task, and stimulus properties: as long as the question is clear, controls can be added at will. The MR imaging instrument, however, may be quite sensitive to these dimensions of the stimuli. Direct comparisons between signal intensities of the Movement and QR conditions may thus be modulated not by the contrast of interest, but rather, by any of the above mentioned (and potentially other) confounding factors. A direct implementation of (1)-(2) in fMRI thus runs into difficulties due to the multiplicity of dimensions involved.

So, is the QR hypothesis testable in fMRI? On the design side, the answer is: perhaps, provided that special measures are taken – that a large number of controls be introduced to get around the problems: a control for task type, one for length, for the presence of multiple quantifiers vs. R-expressions, etc. This may be feasible, however each control condition would increase the size of the experiment, which would end up being cumbersome and long, perhaps too long. There are limitations to the duration of experiments: participants become fatigued and lose concentration, and may become claustrophobic and ask to leave the magnet prior to the completion of their test session.

On the analysis side, problems arise as well: a larger experiment would require additional statistical analyses, which might undermine the reliability of the results, due to too many comparisons that increase the chance of analytic error.[4] Increasing the dimensionality may not be the way to go. This difficulty may be the reason, in fact, that a test of the QR hypothesis (and of related issues) has not yet been carried out in fMRI.

The proliferation of necessary controls is indeed a problem that has plagued the field, making research through the subtraction method difficult, as it can only be effective with minimal pairs, contrasting in one feature, but identical across all syntactic and non-syntactic features. Consider Ben-Shachar *et al*. (2004), in which we tested the claim that syntactic movement activates Broca's region using a contrast between the Hebrew declarative (3a) and left-dislocated (3b) sentences (translated into English below):

(3) a. Danny gave the red book to the professor from Oxford.
b. [To the professor from Oxford]$_i$ Danny gave the red book $t_i$.

Stimuli appeared to consist of truly minimal pairs, as they contained the same words, meaning that they were controlled for size, length, phonological shape, lexical content, and the like. The "subtraction method" obtained a higher BOLD response for (3b) than for (3a) in a relatively large cluster of contiguous voxels in Broca's region. This was considered a welcome result, consistent with findings from aphasia, as well as from other imaging studies. It seemed to fortify the conclusion that mechanisms for the analysis of syntactic movement in comprehension are supported by neural machinery in

[4] It is also important to note that an interaction effect between signal intensities of the Movement and Island conditions for both Wh-movement and QR would not suffice to support a conclusion that the two relations share neural representation. Recall that in fMRI, both signal intensity and signal location are the relevant dimensions. For this conclusion to be supported, the interaction for both must therefore be found in the same brain loci.

this brain region (Grodzinsky, 1986, 2000). But is movement the sole dimension here? A reader may quickly realize that the activation contrast may be due to other factors: syntactic movement here serves semantic focus, making (3b) mean that *the professor from Oxford*, and no one else among the characters under consideration, received the red book [5]. The same meaning can be obtained for (3a) if this NP is stressed. Still, (3a-b) contrast in either stress or meaning, in addition to overt movement. We can introduce more controls, but that, as we have seen, might render the experiment unfeasible. Teasing apart multiple factors, then, is exceedingly difficult if done through direct contrasts.

In other words, the subtraction logic would work for tests in which every pair of conditions contrasts in exactly one property, the *Property Of Interest* (POI henceforth). We expect the net effect of the POI on the dependent variable (BOLD response in a designated cluster) to be solely determined by this contrast. Voxels in which a significant signal intensity difference is found are taken as involved in the processing of the POI. But if the contrast involves stimuli that differ on multiple dimensions as in the cases just discussed, its net effect cannot be attributed to a single property. We must reduce this problem from a multi- into a uni-dimensional one.

### 2.3 Reducing dimensionality – going parametric

Parametric experimental designs come to our rescue. They enable us to reduce the dimensionality of the contrast to the desirable single dimension, although at a cost of additional (yet reasonable) assumptions. The basic idea of parametric designs is the abandonment of direct comparisons for indirect ones: since comparing 2 conditions to one another is problematic, a parameter is chosen, such that when nested in each condition, its values are predicted to affect the cognitive/linguistic POI in a manner that differentiates between the conditions. This enables within-condition comparisons, namely comparisons between sub-conditions, which differ from one another only in the value of the nested parameter. Regions supporting the POI would be discovered thus: if conditions are indeed different in terms of the POI, then the parametric manipulation would modulate the BOLD response differentially: in the +POI regions, changes in parametric values in condition A (but not B) would strongly modulate the BOLD response; in –POI regions, it would be changes in parametric values in condition B (but not A). Crucially, these changes would be detected when the A sub-conditions are compared to one another, and likewise, the B sub-conditions. Big parametric effect would indicate voxel involvement, and the parametric effects for A and B can be calculated and compared. The neural basis of the POI would be discovered through the calculation of relative parametric effects, and without direct comparisons between conditions that differ on multiple dimensions. An example might help.

Cognitive psychologists refer to Working Memory (WM) as the active temporary store used in rapid cognitive tasks (Baddeley, 1997). The need to store items temporarily taxes WM increasingly as storage duration goes up. Increased cognitive effort (load) results in increased neural activity in the areas entrusted with WM. Cognitive psychologists further suppose that WM comes in several varieties: one for verbal tasks, another for visual tasks, etc. Distinct types of WM should be supported by distinct neural

[5] We could actually try to control for this confound by imposing emphatic stress on *the professor from Oxford* in (3a) so it express the same meaning as (3b). But that would introduce another potential confound: stress.

machinery. Thus a parametric manipulation – increasing WM load – on several WM types should result in differential effects: if distinct brain regions are entrusted with each WM type, then increased storage duration should modulate the BOLD response differentially. Pursuing this line of reasoning, Braver et al. (1997) sought to demonstrate that a load increase in verbal WM selectively modulates a certain brain area. Verbal WM is said to store "verbal" material temporarily, and to be distinct from other sorts of WM. They used the well-known *n*-back paradigm: participants are presented with a sequence of single items; for each item s/he has to indicate (through a yes/no button push) whether it is identical to a specified one (e.g., the letter "*m*"), or to an item that may be the previous one, the *2*-back one, or the *3*-back one. A BOLD response curve can now be charted within condition for each brain region of interest, and different WMs can be subsequently teased apart without a problematic direct comparison ("subtraction").

Braver et al. showed letter sequences in the "verbal" condition, and sequences of faces in the contrasting visual condition. The POI was therefore WM type, and item distance (red arrows) was the nested parameter, resulting in 6 sub-conditions (Figure 1A). The "verbal" WM condition is shown in Figure 1B and the visual WM in Figure 1C, see Smith & Jonides, 1999 for a review and a model):

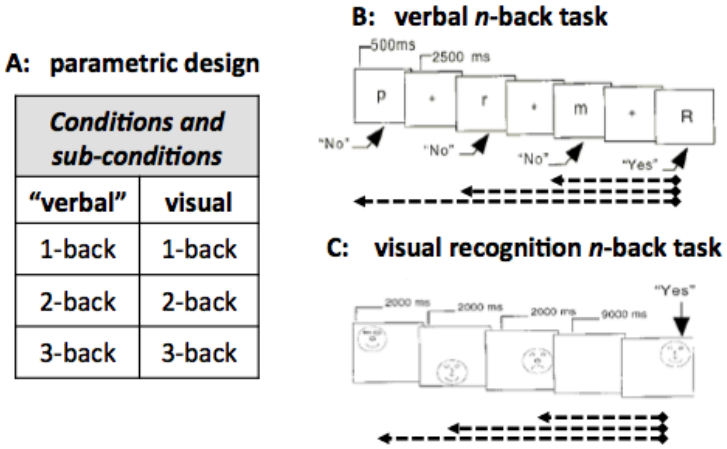


Figure 1: the parametric *n*-back task with item-distance as the parameter

Braver et al. succeeded in teasing the two WMs apart: they found that the intensity curve in Broca's area for $0 < n < 3$ in the verbal WM task was quite different from that obtained for the visual WM task. Their results, then, were obtained through an indirect comparison between conditions, and without "subtraction": they compared the slopes of the signal intensity curve obtained for each condition as a function of the change in value of the nested parameter (i.e., comparison of 2 linear effects). In Broca's region, then slope for "verbal" WM (the linear effect of the *n*-back parameter) was much steeper than that of the visual one, leading to the conclusion that "verbal" WM is distinct from its visual counterpart, and is further supported by neural tissue in Broca's region.

## 2.4 Application 1: parametric studies of syntactic movement

The study of syntactic modularity in the brain also faced the problem of multi-dimensionality. Suggestions that syntactic movement was supported by Broca's region (Grodzinsky, 1986; 2000) were based on evidence from aphasia as well as functional

imaging (notably Ben Shachar et al., 2003; 2004; Friederici et al., 2006; Grodzinky & Friederici, 2006). The latter studies were done through direct contrasts, and thus suffered from the multi-dimensionality problem. Andrea Santi and I parameterized the experimental paradigm, so that comparisons would be done indirectly. We had the *n*-back results in mind, and while "verbal" WM as conceived by Braver et al. could not be linked to syntax directly, we considered the possibility that that elevated signal intensity for movement found in fMRI in health reflects WM recruited to maintain trace/antecedent dependency, not syntactic effort (and likewise, the syntactic movement deficit in Broca's aphasia). To distinguish the movement and the WM accounts, and test for movement specificity in Broca's region, we looked for an intra-sentential dependency *other than* movement requiring a temporary store. Binding was a natural candidate, and we emulated the *n*-back design in an experiment that compared sentences with movement with those containing binding. The POI was ±movement, with both sides of the contrast containing a dependency relation, in which dependency distance was increased, similar to *n*-back. A distance parameter *d* increased the size of the dependency relation, where the values of *d* were obtained by increasing the number of *intervener NPs* between the two co-dependent elements in the same manner across conditions (Santi & Grodzinsky, 2007):

| ***Conditions and sub-conditions*** | |
|---|---|
| Movement | Binding |
| 1-back | 1-back |
| 2-back | 2-back |
| 3-back | 3-back |

Table 2: a skeletal parametric design for the Movement/Binding study

(4) **<u>Movement</u>**[6]

a. *1-back* The man and the mother of Jim love the woman who *Kate* burnt ◂
…NP... NP…NP…*The woman* …***NP***…*t*

b. *2-back* The mother of Jim loves the woman who *the man* and *Kate* burnt ◂
…NP... NP…*The woman*…***NP***… ***NP***…*t*

c. *3-back* Kate loves the woman who *the man* and *the mother* of *Jim* pinched ◂
...NP…*The woman*…***NP***…***NP***...***NP***… *t*

(5) **<u>Binding</u>**

a. *0-back*[7] The sister of Kim assumes that Anne loves the man who burnt himself
…NP... NP…NP… *the man* … *himself*

b. *1-back* The sister of Kim assumes that the man who loves *Anne* burnt himself
...NP...NP… *the man* …***NP***… *himself*

c. *2-back* Anne assumes that the man who loves *the sister* of *Kim* pinched himself
…NP... *the man*… ***NP***… ***NP***… *himself*

---

[6] Stimuli are cumbersome, but they are all of the same size, containing no confounding variables.

[7] Note that in (4), $1<n<3$, whereas in (5), $0<n<2$. This difference is immaterial: absolute values don't matter, but rather, rate of change in *n*. The 2 conditions are equal in this respect.

We presented these sentence types[8], measured subsequent changes in BOLD response, and compared the rates of signal change in both conditions. We found voxel clusters where signal intensity change rate as *d* went up (linear effect) in the Movement condition (4a-c) was significantly higher than its Binding analogue (5a-c). This comparison discovered the net effect of the POI on BOLD response. Movement localized in the anterior part of Broca's region (Brodmann's Area 45), where change in *d* in the Binding condition had virtually no effect. An isolable neural component for movement was thus discovered with no direct comparisons, as the parametric design reduced contrast dimensionality. Related studies have been carried out: in German. Scrambling and Movement were compared indirectly through a distance parameter (Makuuchi et al., 2010); in English, a hierarchically defined *d* was deployed (Santi et al., 2010).

## 2.5 Application 2: parametric studies of quantifier calculation

The successful deployment of a parameter in syntactic stimuli motivated new attempts. We moved to study quantification and its relation to the neurocognitive analysis of numerical quantities in fMRI. Leads from earlier studies suggested the existence of neural structures that distinguish between numerical quantities and number words (Cohen & Dehaene, 2001); other studies distinguished between quantifier types (McMillan et al., 2005; Troiani et al., 2008). However, these were "subtraction" studies, and so doubts linger on: the quantifiers compared differed not only in their semantic properties, but also phonologically, morphologically and syntactically (cf. *no*, *many*, *exactly two, less than three*). In other words, it is not clear that sentences with quantifiers can be compared through subtraction without confounds. In Heim et al. (forthcoming), we introduced a new *Parametric Proportion Paradigm* (PPP) to investigate proportional quantifiers.

The PPP exploits the fact that quantifiers that pertain to a part of a whole are consistent with more than one scenario about proportions: as long as the proportion of the relevant property of their restrictor in a scenario satisfies they denotation, the sentence is true. Thus (6a,b) are true in several scenarios in Figure 2:

(6) a. Most circles are blue.
b. Many circles are blue.

Figure 2: example visual stimuli for the PPP

We used such scenarios to investigate the fine neural structure of numerical analysis, and its relation to that dedicated to semantic analysis. Participants were inside the MRI machine, and were asked to verify auditory sentences with proportional quantifiers against visual scenarios, under time pressure. Each sentence was followed by a scenario containing 50 quasi-randomly positioned circles[9] with varying radii, divided into 2

[8] The task was grammaticality judgment, which means that (4)-(5) represent about half our stimuli. Ungrammatical sentences were used, but the BOLD response to these was not analyzed.

[9] 50 was chosen to avoid subitizing.

contiguous clusters of blue and yellow. The truth or falsity of each sentence depended on the quantifier, and on blue/yellow proportion (Figures 2-3). Judgments became more difficult as blue/yellow proportion approached 1, which resulted in elevated RTs.

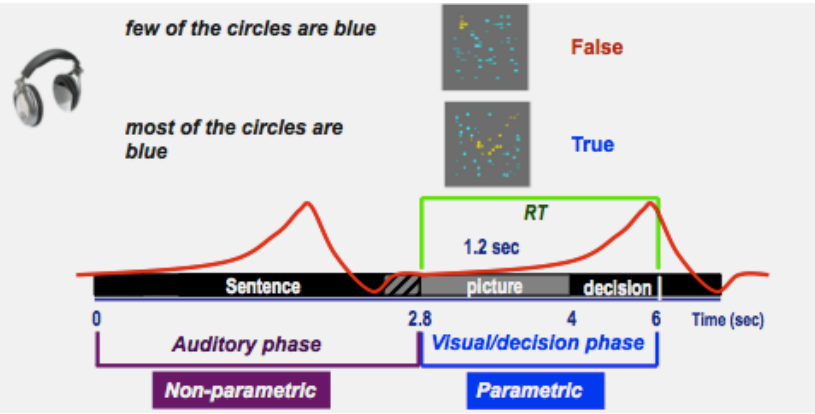


Figure 3: the course of a PPP trial

Proportion and RT were the parameters. Instead of comparing BOLD responses for the different quantifiers to one another, we plotted the way each proportion (5/50, 10/50,…,45/50 blue) modulated the BOLD response for each individual quantifier; we did the same with RT. We could then compare the curves. Once again, nested parameters reduced the multiple differences between conditions to the POI.

We used 6 German quantifiers, divided along two axes (Table 3)[10]: a. Polarity – negative and positive quantifiers (rows). Thus *less-than-half* contains a negative element that its positive cousin *more-than-half* is lacking. This contrast has special linguistic significance in German and Dutch, allowing scope splitting in certain contexts (e.g., Jacobs 1980; Penka 2007; Rullmann 1995; de Swart 2000). b. Degree vs. Proportion quantifiers (columns): the former denote a comparison of the numerosity of their restrictor to the numerosity of some quantity, large to a contextually determined degree $d_C$ (Partee, 1988). The latter denote a proportion expressed through set relations (Hackl, 2000; 2009). Resulting entries are roughly these (irrelevant details are suppressed):

| | *Fixed standard of Comparison* | *Contextual standard of Comparison* |
|---|---|---|
| ***NEG*** | ⟦Less-than half⟧(C)(B) = true iff \|C∩B\|<\|C–B\| | ⟦few⟧=$\lambda d \lambda x \in D.Num(x) < d_C$ & >1<br>⟦fewest⟧=$\lambda d \lambda x \in D.Num(x) << d_C$ & >1 |
| ***POS*** | ⟦More-than half⟧(C)(B) = true iff \|C∩B\|>\|C–B\|<br>⟦Most⟧(C)(B) = true iff \|C∩B\|>\|C–B\| | ⟦many⟧=$\lambda d \lambda x \in D.Num(x) > d_C$ & x>1 |

Table 3: partial design of the Heim et al., experiment

This rich design localized and separated different calculation types. It also enabled comparisons between quantifier types. In particular, it uncovered a Polarity brain locus, where negative quantifiers produced higher signal intensity than positive ones[11].

[10] *viele, wenige, mehr-als-die-Hälfte, weniger-als-die-Hälfte, die meisten, die wenigsten.*

[11] Only 4 quantifiers entered this analysis: *viele, wenige, mehr-als-die-Hälfte, weniger-als-die-Hälfte.*

## 2.6 Hints for a movement analysis for split negative quantifiers?

German and Dutch negative indefinites allow for split scope – (7) is 3-way ambiguous:

(7) Du musst keine Krawatte anziehen
You must no tie wear

a. 'It is not required that you wear a tie.' ¬ > must > ∃

b. 'There is no tie that you are required to wear.' ¬ > ∃ > must

c. 'It is required that you not wear a tie.' must > ¬ > ∃

Briefly, (7a) is the relevant reading here: the modal *musst* splits *kein* into negation and an existential. There seems to be a consensus on the contextual restrictions on scope splitting – it only occurs with certain types of intensional predicates; the puzzle concerns the analysis: some authors propose a lexical decomposition analysis in which negation moves to outscope the intensional predicate (Rullman, 1995; Penka, 2007), whereas others avoid decomposition by invoking devices like quantification over properties (e.g., de Swart, 2000). Can experimental data adjudicate between the two approaches?

Proportional and degree quantifiers exhibit the same 3-way ambiguity. Only the split reading, on which we focus, is presented below:

(8) a. Du musst weniger-als-die-Hälfte der Äpfel essen
You must less-than-half of the apples eat
'It is not required that you eat half or more of the apples' ¬ > must > ≥ ½

b. Du musst wenige der Äpfeln essen
You must few of the apples eat
'It is not required that you eat n>*d*-many of the apples ¬ > must > -er > *d*-many

Heim et al. therefore used quantifiers that allow scope-splitting. Still, on the face of it, this experiment does not bear on the scope-splitting debate, because none of the sentences used provided the right context for scope splitting. But I would like to point to two aspects of these results that are nonetheless suggestive: a. signal intensity was higher for negative quantifiers, but what about location? The Polarity contrast is semantic, and so, the activation of meaning related regions is expected. Surprisingly, however, Polarity modulated the anterior portion of Broca's region, namely Brodmann's Area 45 – precisely the area modulated by syntactic movement in the experiments reported above, and within the region damaged in Broca's aphasia, which leads to a movement deficit.

Next, consider behavior: when participants' verification times were examined, their RTs for sentences with negative quantifiers were significantly longer than those for the positive ones. Why would that be? There is little evidence that in general, negative expressions take longer than positive ones; ours seems a result restricted to the present materials. An account of it must be tied to the properties of the contrast at issue, namely negation. It is hard to imagine how a non-decompositional account would derive it.

The RT data suggest decomposition; the fMRI data hint at movement. And so, to hope that these results are not accidental seems reasonable. True: they are at best preliminary hints, and more work is required; yet if true, these convergent results provide neurological and behavioral arguments not only for a movement analysis of split scope, but also, for the fine structure of the representation of linguistic ability in neural tissue.

This concludes the section. With these results in mind, we can move on to issues of experimental data analysis.

## 3 The Mapping Problem – from categorical to ordinal variables

### 3.1 From representations to syntactic error patterns: aphasia

The primary data for syntactic theory are typically <string, label> pairs – judgments elicited from people, e.g., sentences marked with an asterisk for ungrammaticality or left unchecked as they are well formed. The label thus typically takes 2 values (where the relevant category, or dimension, is marked ±). In experimentation, however, the format of data is different. Experimental tasks take linguistic objects and pair them up with quantitative measures (success, time, brain activity). The results of such experiments effectively add a dimension to the primary data, resulting in triplets <string, label, quantity> (e.g., <*John loves Mary*, +Grammatical, .8*sec*>), where dependent measure – the quantity – is determined by the task, the paradigm, etc.

This section focuses on experiments with linguistically deficient populations – normally (or pathologically) developing children whose mastery of language is not yet complete, or adults missing part of their language. Theories of linguistic representation are usually put to the test in these populations when performance levels on different construction types are pitted against one another; error rates are a commonly used dependent variable. For concreteness, consider an experiment designed to compare comprehension performance on subject-relative clauses to that on object-extracted ones. A typical comprehension probe requires participants to assign theta-roles by choosing one of 2 pictures (the boxes below). Usually, $n$ tokens of each sentence-picture pair type ("trials") are presented. Given $m$ successes, $0 \le m \le n$, the score is the pair $\langle m, n \rangle$[12]. This is a forced binary-choice design, where 👍 signifies a correct answer and 👎 marks an error:

(9)

| | sentence | Picture | | response |
|---|---|---|---|---|
| a. | The elephant who is pushing the monkey is grey | 👍: e➡m | 👎: m➡e | <*m*, *n*> |
| b. | The monkey who the elephant is pushing is grey | 👍: m➡e | 👎: e➡m | <*l*, *n*> |

In (9), both answer options are presented concomitantly. Rates of correct responses and errors are therefore coupled: $m$ correct choices dictate $n$-$m$ errors.[13]

How are such data analyzed and brought to bear on the theory? Usually, a statistic reflecting performance differences between pairs of conditions is typically reported. It is the result of a test of whether the level of performance on one type is higher than that on others, which should help deciding between different theories. Test specifics depend on the assumptions made, but essentially, in (9) the question is whether, given $r$ participants, the mean proportions of correct trials per condition $\bar{\mu} = \sum_{i=1}^{r} \frac{m/n}{r}$, $\bar{\nu} = \sum_{i=1}^{r} \frac{l/n}{r}$ are statistically distinguishable from one another. Conclusions are drawn on this basis.

---

[12] The representation of results not as a proportion, but as <#successes, #trials> is due to the relevance of the latter quantities: everything else equal, the higher the number of trials, the lower the variance. With a high number, the reliability of comparison is high (i.e., <2,4> is less informative than <20,40), and this quantity should therefore be taken into account (cf. Drai & Grodzinsky, 2006).

[13] The experiments described above have a cousin, in which a single picture accompanies each sentence, which participants deem true or false. This design features $n$ presentations of each picture per sentence type. True and False responses per condition are decoupled, rendering error rates independent. A richer picture emerges, whose richness may have theoretical consequences (cf. Chien & Wexler, 1990; Grodzinsky & Reinhart, 1993).

Finding that mean performance on (9a) is higher than on (9b) is informative, but insufficient for a clear interpretation. The overall picture is characterized not only by the relation between the means, but also, by relation between each mean and external measures. In (9) we have a binary forced-choice experiment. States of incomplete knowledge may lead to uncertainties, which may result in guessing behavior. This gives us a first approximation of an external measure – chance-level, which we compare to $\overline{\mu}$, $\overline{\nu}$. Figure 4 describes 6 possible outcomes of the in (9):

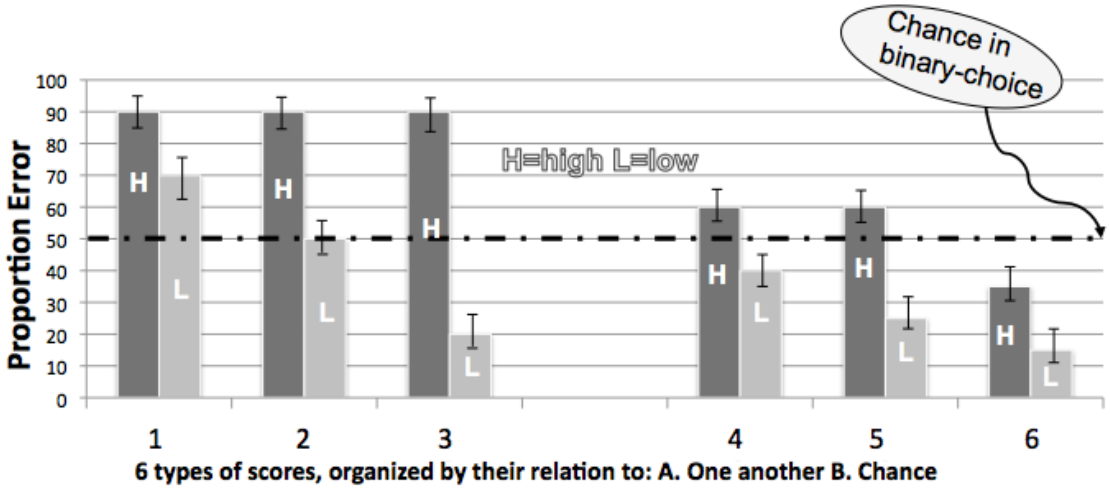


Figure 4: A typology of error patterns

Each pair of columns in the histogram represents an outcome. Common to all 6 outcomes is the fact that $\overline{\mu}$, the mean of the High performance condition (9a), is significantly higher than $\overline{\nu}$, the mean of the Low performance condition (9b). But the cases differ otherwise: while $\overline{\mu}$ approaches 1 in 1-3, $\overline{\nu}$ varies from being significantly higher than chance in 1, to chance in 2, to below-chance in 3. As the task is binary-choice, chance level is set at .5. And the relation between it and other scores is important. When $\overline{\nu}$ is above-chance, participants select the correct answer most of the time, indicating awareness of the correct theta-role assignment on object-extracted relatives. Some other factor lowers their performance relative to subject-extracted ones; if they produce outcome 2, with $\overline{\nu}$ indistinguishable from chance, they are guessing, hence there may be something seriously wrong with their theta-role assignment. Finally, in outcome 3, $\overline{\nu}$ is *below* chance, meaning that theta-roles are *systematically* reversed. Whatever the reason, this outcome calls for yet a different interpretation. Outcomes 4-6 in Figure 4 are even more problematic, and are left for the reader to consider.

The point, then, is that differences between means in the present context only tell part of the story, as they do not distinguish between outcomes 1-3 (or 4-6 for that matter). An additional statistic is needed for this further distinction to be made – a test that determines the relation between the performance level at issue and chance. Thus a complete picture of the experiment described in (9) is obtained when each of the 2 data points, $\overline{\mu}$, $\overline{\nu}$, is characterized relative to the other, as well as to chance.

Readers might wonder about the special status accorded to chance-level; what makes guessing a benchmark against which performance levels are checked? Consider how theory is mapped onto the experimental data under discussion. The theory of syntax

is about categorical variables, as it pertains to the grammatical status of strings in the language. For the most part, the theory deems a string grammatical, or ungrammatical. Errors, by contrast, constitute a continuous (or at least multi-valued) variable. On *n* trials, the number or errors can be anywhere between 0 and *n*. Linguistic theory, then, is not designed to map onto such data. To relate it to this type of experimental data, we can either create a complex mapping with a rich set of prediction, or keep it simple, with limited predictions that might actually make sense in certain cases.

Bard, Robertson & Sorace, 1996 (also, Sprouse, 2011) choose the former path. Their starting point is the observation that primary linguistic data are not black and white (±Grammatical) but are rather on a grey scale. They therefore construct a complex mapping from grammatical knowledge to acceptability judgment. The present case is different, however, as the data come from decisions that require the comprehension of strings that are always grammatical. The task, moreover, is forced binary-choice theta-role assignment. Therefore, the theory – on its most innocuous interpretation – predicts either correct assignment of theta-roles that results in errorless performance (i.e., $\bar{\nu} \rightarrow 1$), or theta-role reversal ($\bar{\nu} \rightarrow 0$), or erroneous performance in which the participants are reduced to guessing for some reason ($\bar{\nu} \sim .5$). This perspective, then, divides the data into 3 bins – above-, below- and at-chance levels, as cases 1-3 in Figure 1 illustrate.[14]

A concrete example might clarify the significance of this issue. In Grodzinsky (1995), I asked a Broca's aphasic patients to sequence sentence fragments (e.g., *the priest* | *covers* | *the nun*; *the book* | *is covered by* | *the newspaper*) into a sentence that matches an accompanying picture. It was known then that these patients suffer a deficit related to syntactic movement, and the goal of this experiment was to test a finer distinction – whether the deficient system interacts differently with different predicates types. The picture *cum* sentence fragments paradigm was therefore introduced with several sentence types (10), where syntactic movement (active/passive) and predicate type (whose external argument was experiencer, agent, or instrument) were systematically manipulated:

(10)

| **Sentence** | **Type** | **response** |
|---|---|---|
| a. The priest covers the nun | *Active-agentive* | OK |
| b. The nun is covered by the priest | *Passive- agentive* | Chance |
| c. The book covers the newspaper | *Active-instrument* | OK |
| d. The newspaper is covered by the book | *Passive- instrument* | Chance |
| e. the priest admires the nun | *Active-experiencer* | OK |
| f. the priest is admired by the nun | *Passive-experiencer* | *Below-chance* |

A sketch of group results (rightmost column) reveals high performance in actives, which drops in passive. A closer look reveals further structure within the passives: the agent/instrument manipulation does not affect performance (10a)/(10c),(10b)/(10d)), but in passive experiencer predicates, systematic *reversal* of theta-roles is attested (10f).

Analyses restricted to the relation between conditions (i.e., between mean performances), lead us to conclude that patients always perform better on actives than on passives. Yet when the relation between each failure and chance-level is analyzed, richer structure is discerned, with immediate consequences to the mapping of results onto

[14] This does not preclude a more complex mapping, designed to fit a richer set of data in other experimental contexts. See section **2.1.3** for such an example from semantics.

theory. The characterization of the syntactic deficit in Broca's aphasia requires revision, as the deficit seems to interact with thematic labels. This result led to a refinement of the deficit description, and to a more restrictive view of the functions of Broca's region in sentence analysis. It also suggested that thematic labels on theta-grids are encoded in the syntax, a much-debated question at the time (Grodzinsky, 1995; 2000).

### 3.2 Semantic error patterns: inferences with quantifiers

The need for explicit mapping between a theory that makes categorical distinctions and performance that is described on a continuous scale is not limited to syntax. Geurts (2003) presents a theory of syllogistic reasoning with quantifiers, in which he attempts to connect the semantics of quantificational expressions to extant quantitative data on Aristotelian syllogisms, performed on quantificational statements. It is well known that people often err when asked about the validity of arguments for which they must be aware of entailment relations that hold between quantified sentences. The following entailments (the score [proportion "yes"] is associated) are some of those tested (from Newstead & Griggs, 1983; inferences are: 👍=valid; 👎=invalid; ?=undetermined):

(11)

a. All chess players are beer drinkers 👎 $\Rightarrow_{.02}$ Some chess players are not beer drinkers

b. All chess players are beer drinkers 👍 $\Rightarrow_{.73}$ Some chess players are beer drinkers

c. No chess players are beer drinkers 👍 $\Rightarrow_{.69}$ Some chess players are not beer drinkers

(12)

a. Some poodles have curly hair ? $\Rightarrow_{.94}$ Some poodles do not have curly hair

b. Some poodles do not have curly hair ? $\Rightarrow_{.83}$ Some poodles have curly hair

c. Some poodles do not have curly hair ? $\Rightarrow_{.02}$ No poodles have curly hair

The 2 parts of inference (11a) are contradictories. Indeed, participants reject its validity 98% of the time. Yet, they accept Universal Instantiation at a surprisingly low 73% (11b), and realize that universal negation entails existential negation at an even lower rate (11c). Subjects also err in cases where the truth-value of the consequent cannot be determined from that of the antecedent (12): existential negation does not entail universal negation (12a), but participants nonetheless accept it at a high rate; they also tend accept the (undetermined) inference in the other direction (12b), and deny (12c).[15]

[15] The numerical data are presented in an underspecified manner: the numbers represent group means, but it is not clear how many items per condition each participant received (in some studies we actually *know* that participants each received 1(!) item per condition). In the absence of a considerably large sample per participant (and in the absence of a report about the variance in the data), we cannot rule out the possibility that middle-range success rates represent two or more distinct sub-populations, who apply different judgment strategies. See below, for a somewhat similar point regarding Chemla's (2009) study.

What is the numerical character of this performance pattern? One could imagine several approaches: one way would replicate the above approach, compute the relation between each data point and chance; another approach would rank-order the success rates, regardless of specific value (e.g., $Perf_{11b} > Perf_{11c}$, etc.); yet a stronger approach would try to derive the results exactly as recorded, including their numerical values.

Geurts (2003) takes the last, most ambitious, track. He is thus looking at a difficult problem, which becomes especially acute in light of the fact that participants who had made errors during the test actually agree that their judgments were mistaken after some instruction (a 5-minute crash course in propositional logic). If they are in principle able to identify the correct solution, why do they fail in the first place? In an attempt to provide an explanation of these results (among others), Geurts proposes a bi-partite account: a semantic part and a processing part. The former takes the monotonicity property of the quantifiers involved in the inferences in (11)-(12) to be at the heart of the derivation of valid inferences[16]; the latter, processing part is special, in that it assigns a specific cost to each operation involved in the derivation. The idea is that inferences involve a sequence of steps; that each step has a cost attached to it; and that the higher the total cost (the sum of the costs incurred by the sequence of steps required to derive the inference) the higher error-rate observed in experiments would be. Geurts further reports that when he correlated the predicted costs of for each entailment with the experimental results reported in the literature, the correlation coefficient was extremely high (r=.93).
One need not endorse the particulars of this proposal to see its merits: it is an unusual attempt to construct an explicit mapping between a representational theory with independent categorical variables on the one hand, and dependent continuous variables on the other hand – error rates in inference making. If correlation is the appropriate measure in this case, then, given the high correlation obtained, Geurts' theory may explain not only why performance on syllogisms is ranked as it is; it may also account for specific error rates. In other words, it explains why participants very confidently view the entailment in (11a) as false, why they accept (11b) as true 73% of the time, and so on.

It is not easy to establish such a mapping, and moreover, provide motivation for every aspects of it: indeed, this is one of the places where Geurts' account seems to be lacking, as the processing costs attributed to at least some of the derivational operations are arbitrary. The use of correlation in this case, moreover, may not be conceptually justified (see Newstead, 2003 for a critique). Either way, the importance of the direct mapping approach of the type Geurts proposes cannot be overestimated. One hopes to see more proposals that map categorical onto continuous variables explicitly.

### 3.3 Experimental pragmatics: Presupposition in quantified sentences

Chemla (2009) presents judgment data from an experimental project that explored the hypothesis that different presuppositions are projected in different quantificational contexts. The problem, as he describes it, is that while it seems to be universally agreed that (13a-b) have a presupposition, there is no such agreement on the nature of this presupposition. Some authors view it as (14a), and others, as (14b):

(13) a. Every student found the typo in her paper.
b. No student found the typo in my paper.

[16] In this respect, Geurts' account of the data on inference-making is similar to earlier ones, see Newstead (2003).

(14) a. Every student has a paper with a typo
b. Some student has a paper with a typo

Chemla designed an experiment aimed at settling this debate. Moreover, he developed a theory in which presuppositions of quantified sentences depend on the specific quantifier at issue. And so his experiment, in addition, aims to adjudicate between this theory and others. The experiment was rich in structure, including a large number of controls, which I ignore, as I focus on a quantitative issue in the context of the mapping problem (see Chemla's work for details). Here are some predictions for presuppositions he derives from his theory ("⇒" presupposes):

(15) a. Each student knows that he's lucky ⇒ Each student is lucky.
b. No student knows that he's lucky ⇒ Each student is lucky.
c. More than 3 students know that they're lucky ⇒ More than 3 students are lucky and less than 3 aren't.
d. Less than 3 students know that they're lucky ⇒ At least 3 students are lucky and less than 3 aren't.
e. Many students know that they're lucky ⇒ Many students are lucky.

Participants in the experiment were presented with a sheet containing sentence pairs – an assertion followed by a possible presupposition. Their task was to decide if the former "suggests" the latter. Of particular interest are (16a-b), that featured as the first sentence each (hence functioned as the assertion), and (17), a possible presupposition that followed both. Participants were thus asked to decide whether or not (16) "suggests" (17), and mark "yes" or "no" on the answer sheet, a task Chemla takes to indicate whether they accept or reject it as a presupposition.[17]

(16) a. Each of these 10 students knows that his father is going to receive a congratulation letter[18].
b. Less-than-3 of these 10 students know that their father is going to receive a congratulation letter.

(17) ⇒ The father of each of these 10 students is going to receive a congratulation letter.

Acceptance rates of the universal presupposition (17) were 84% when it followed (16a), but dropped to less than 70% when it followed (16b).[19] Chemla notes that "it is not clear

[17] The instructions given seem ambiguous, potentially affecting participants' construal of the task: they were expected to push the "no" button if they thought that the first sentence does not suggest the second. But the setup could be interpreted differently: a request to push the "no" button could be seen as a request to confirm the negation of the second sentence as a presupposition. Despite the fact that *suggest* is not a Neg-Raising predicate (as pointed out to me by Bernhard Schwarz), the setup may have led some participants to think that in pushing the "no" button, they indicate that the first sentence suggests not(the second). These are two different interpretations that may result in a split in participants' response patterns. I found no discussion of this possibility in Chemla's paper.

[18] The example refers to a French custom in which parents of outstanding students receive letters of congratulation.

in this case that this means that presuppositions project universally. In fact, these results could be taken as evidence that presuppositions give rise to inferences intermediate between scalar and universal inferences" (p. 311). A second experiment he ran obtained more refined scores, as participants were asked not to accept or reject an inference, but rather, to provide a numerical indication of how strongly the assertion "suggests" this inference. The results were similar, leading Chemla to propose that presuppositions are ranked, and "the closer to universal the prediction, the higher participants rate the universal inference" (p. 323).[20]

This experiment is as interesting as the results: the difference between the acceptance rates on (17) when followed (16a) and (16b) seems to support the claim that different quantifiers in assertions result in different presuppositions. But when the exact *mapping* between theory and data is considered, the plot seems to thicken: first, Chemla's theory predicts that (15d) is the inference for the assertion (16b). But participants were offered (17). By Chemla's theory, however, (17) is not the presupposition of (16b). (16b) presupposes the analogue of (15d), which is neither equivalent to, nor entailed by, (17). Participants are therefore expected to flatly reject (17) as a presupposition for (16b), resulting in acceptance rate of 0%! Chemla's results, then, produced differential performances with different assertions, hinting that indeed, presupposition is not uniform across quantificational contexts. However, once an explicit mapping between the theory and performance levels is established, at least some predictions are not borne out.

## 4 Coda: what neurolinguistic discovery means

Many years ago, when I had just completed my dissertation, Morris Halle asked me what results I could report. I told him that I showed that in Broca's aphasia, the comprehension system distinguishes between lexical passive, which it analyzes properly, and verbal passive, which leads to comprehension blunders, i.e., chance performance (see Grodzinsky, Pierce & Marakovitz, 1991). It is already known, I told Morris, that the same patient group also fails in comprehension tests of wh-movement constructions. When the latter result is considered together with mine, I said, you get support for a generalization over Wh-movement and NP-movement (*a k a* Move-alpha). To which Morris replied: "but that we know already. Teach me something I don't know." I was not quick enough then to articulate what I am now able to say: "here is what you didn't know: a. that there are neural specializations for different linguistic operations, which suggests the existence of a brain map for linguistic knowledge. b. that highly structured experimental results sometimes converge on what we know, but sometimes reveal novel facts. Adding these to the linguist's data set is important, as it broadens the empirical horizon of the theory."

I imagine that the 2 points I once failed to convey to Morris will resonate well with linguists who take biology seriously. I thus hope that the solutions I proposed to problems that arise when a behavioral and/or neurological dimension is superimposed on the theory will help to bring linguistics and neuroscience a bit closer.

---

[19] Chemla does not specify the number of sentences of each type, which leaves open the possibility that the numbers were not due to individual vacillation, but rather, to non-homogeneity in the group of participants, that may have led to a split (see note 15).

[20] Part of Chemla's study was designed, in fact, as a test of Geurts' (2003) theory, to which the results had rather negative implications.

REFERENCES


Baddeley, Alan. 1997. *Human Memory: theory and practice.* London: Psychology Press.

Bard, Ellen Gurman, Dan Robertson & Antonella Sorace. 1996. Magnitude Estimation of Linguistic Acceptability. *Language, 72*, 32-68.

Ben-Shachar Michal, Talma Hendler, Itamar Kahn, Dafna Ben-Bashat & Yosef Grodzinsky. 2003. The Neural Reality of Grammatical Transformations: Evidence from fMRI. *Psychological Science, 14,* 433-440.

Ben-Shachar, Michal, Dafna Palti & Yosef Grodzinsky. 2004. Neural correlates of syntactic movement: Converging evidence from two fMRI experiments. *NeuroImage*, *21*, 1320-1336.

Braver Todd, Jonathan Cohen, Lyn Nystrom, John Jonides, Ed Smith & Douglas Noll. 1997. A parametric study of prefrontal cortex involvement in human working memory. *NeuroImage, 5*, 49-62.

Chemla, Emmanuel. 2009. Presuppositions of quantified sentences: experimental data. *Natural Language Semantics, 17*, 299-340.

Chien, Yu-Chin, and Kenneth Wexler. 1990. Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Language Acquisition, 1*, 225–295.

Cohen, Laurent & Stanislas Dehaene. 2000. Calculating without reading : Unsuspected residual abilities in pure alexia. *Cognitive Neuropsychology*, *17*, 563-583.

de Swart, Henriette. 2000. Scope ambiguities with negative quantifiers. In *Reference and Anaphoric Relations*, ed. K. von Heusinger and U. Egli, 109–132. Dordrecht: Kluwer.

Drai, Dan & Yosef Grodzinsky. 2006. A New Empirical Angle on the Variability Debate: Quantitative neurosyntactic analyses of a large data set from Broca's Aphasia. *Brain & Language, 96,* 117-128.

Fox, Danny. 2003. On Logical Form. In Randall Hendrick, ed., *Minimalist Syntax*. Blackwell.

Friederici Angela, Christian Fiebach, Matthias Schlesewsky, Ina Bornkessel & D. Yves von Cramon. 2006. Processing linguistic complexity and grammaticality in the left frontal cortex. *Cerebral Cortex, 16*, 1709-1717.

Geurts, Bart. 2003. Reasoning with quantifiers. *Cognition, 86*, 223–251.

Grodzinsky, Yosef. 1986. Language deficits and the theory of syntax. *Brain & Language, 27*, 135-159.

Grodzinsky, Yosef. 1995. Trace deletion, $\theta$-roles, cognitive strategies. *Brain & Language, 51*, 467-497.

Grodzinsky, Yosef. 2000. The neurology of syntax: language use without Broca's area. *Behavioral & Brain Science, 23,* 1-71.

Grodzinsky, Yosef & Angela Friederici. 2006. Neuroimaging of syntax and syntactic processing. *Current Opinion in Neurobiology, 16*, 240-246.

Grodzinsky, Yosef , Amy Pierce & Susan Marakovitz. 1991. Neuropsychological reasons for a transformational analysis of verbal passive. *Natural Language & Linguistic Theory, 9*, 431-453.

Grodzinsky, Yosef & Tanya Reinhart. 1993. The innateness of binding and coreference. *Linguistic Inquiry, 24*, 69-101.

Hackl, Martin. 2000. Comparative Quantifiers. PhD thesis, MIT.

Hackl, Martin. 2009. On the grammar and processing of proportional quantifiers: *most* versus *more than half*. *Natural Language Semantics, 17*, 63-98.
Heim, Stefan, Katrin Amunts, Dan Drai, Simon Eickhoff, Sarah Hautvast, & Yosef Grodzinsky. 2010. The language-number interface in the brain: A bi-parametric study of quantifiers and quantities. Presented at the Neurobiology of Language Conference, San Diego, November.
Huettel, Scott, Andrew Song, & Greg McCarthy. 2009. *Functional Magnetic Resonance Imaging* (2nd Edition). Sunderland, MA: Sinauer Associates.
Jacobs, Joachim. 1980. Lexical decomposition in Montague Grammar. *Theoretical Linguistics, 7*, 121–136.
May, Robert. 1977. Logical Form. PhD thesis, MIT.
Makuuchi, Michiru, Andrea Santi, Yosef Grodzinsky & Angela Friederici. 2010. Neural mechanisms for the processing of movement and scrambling constructions. Society for Neuroscience, San Diego, November.
McMillan, Corrie, Robin Clark, Peechie Moore, C. Devita, & Murray Grossman. 2005. Neural basis for generalized quantifier comprehension. *Neuropsychol. 43*, 1729-37.
Miller, George A. & Noam Chomsky. 1963. Finitary models of language users. In R. Duncan Luce, Robert R. Bush, and Eugene Galanter, eds., *Handbook of mathematical psychology*, vol. 2, 419-491. New York: Wiley.
Newstead, Stephen. 2003. Can natural language semantics explain syllogistic reasoning? *Cognition, 90*, 193–199.
Newstead, Stephen & R. A. Griggs. 1983. Drawing inferences from quantified statements: a study of the square of opposition. *Journal of Verbal Learning & Verbal Behavior, 22*, 535–546.
Partee, Barbara. 1989. Many quantifiers. In *ESCOL 89: Proceedings of the Eastern States Conference on Linguistics*, eds. Joyce Powers and Kenneth de Jong, 383-402. Columbus, OH: Department of Linguistics, Ohio State University.
Penka, Doris. 2007. Negative indefinites. PhD thesis, Universität Tübingen.
Rullmann, Hotze. 1995. Maximality in the semantics of wh-constructions. PhD thesis, University of Massachusetts at Amherst.
Santi, Andrea & Yosef Grodzinsky. 2007. Working Memory and Syntax Interact in Broca's Area. *NeuroImage, 37*, 8-17.
Santi, Andrea, Michiru Makuuchi, Angela Friederici & Yosef Grodzinsky. 2010. A parametric study of movement relations. Paper presented at the 2nd Neurobiology of Language Conference, San Diego, November.
Smith Ed & John Jonides. 1999. Storage and executive processes in the frontal lobes. *Science, 283*, 1657-1661.
Sprouse, Jon. 2011. A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language, 87,* 274-288.
Troiani, Vanessa, Jonathan Peelle, Robin Clark & Murray Grossman. 2009. Is it logical to count on quantifiers? Dissociable neural networks underlying numerical and logical quantifiers. *Neuropsychologia, 47*, 104-111.