What Comes First: Theory or Data?
(handout)
peterjludlow@gmail.com

1.0 Data

- Absalom and Hajek (1997: 177): "the data needs to drive theoretical analysis and not the other way around."

  Haspelmath (2010: 293) in defense of "framework-free" linguistics:

  > Since the advent of the Boasian approach in ethnography and structuralism (both European and American) in linguistics, it has been the goal of descriptivists to approach a language without prejudice and to do justice to its system, regardless of what systems other languages might have. We want to describe each language in its own terms.

- **target idea**: investigation should proceed in a bottom-up manner, beginning with theory neutral data collection and only then proceeding to the construction of theory, if at all.
- **Earlier advocates**: Structuralist linguists, British philosopher Francis Bacon, John Stuart Mill in his *System of Logic* (1843).
- **The problem**: despite its intuitive appeal was quickly rejected by key natural philosophers (Pascal, Newton, Darwin, Pasteur, and Mendel).

  Darwin:

  > "About 30 years ago there was much talk that geologists ought to observe and not to theorize; and I well remember someone saying that at this rate a man might as well go into a gravel pit and count the pebbles and describe the colors. How odd it is that anyone should not see that all observation must be for or against some view if it is to be of any service!" (Darwin and Seward (1903)).

  - Darwin and the barnacles.

  Stephen Jay Gould (1990):

  > "Darwin's view on the need for theory both to suggest and to coordinate observations has been widely acknowledged by scientists as both desirable and inevitable (despite the semiofficial persistence of a public myth about absolutely objective impartiality). This interplay of theory and empirical documentation has both positive and negative implications for the elusive notion of scientific "progress." Theory can prod, suggest, integrate, and direct in fruitful ways; I doubt that Darwin would ever have been able to formulate the theme of natural selection without the available context of Adam Smith's nearly identical causal system for economics (Darwin, in any case, surely did not "see" natural selection in the finches and tortoises of the Galapagos)."

- Theory is a *necessary* condition for scientific success; it is not a *sufficient* condition.

- Theory may also be impossible to avoid. Sometimes theory masquerades as a purely inductive project.
- What *is* theory-free data?: Hanson (1958), Sellars (1949), Kuhn (1970).

  Example: Morphemes are not things that we nakedly observe; they are the product of scientific theorizing. Claude Levi-Straus (1953: 350-351) compared the discovery of the morpheme and phoneme to the Newtonian revolution in physics.

  But Haspelmath: the descriptivist or framework-free approach to linguistics "is not only possible and widely practiced," but is "the best approach to the scientific study of language structure."

  Haspelmath: framework-free approach has been applied to Tagalog

  > "Schachter & Otanes (1972: 59-85), still under the influence of American structuralism, describe Tagalog basic sentence structure in its own terms, and the result is a picture that is rather different from what is found in English (with which the authors contrast Tagalog). The basic pattern of Tagalog is not [sentence NP VP], but [sentence Predicate Topic]."

- **example** (drawn from Schachter and Otanes)

  [Gumising]PRED [ang bata]TOP
    awoke   TOP child
  'The child awoke.'

- But this subject-predicate *form* – why are the notions of sentence, predicate, and topic (or their structural equivalents) considered atheoretical descriptors?
- the fact that today they *appear* to be theory-neutral descriptors is merely an indicator of how deeply they are sedimented into our observations of and understanding of language.
- The allegedly frame-work free approach doesn't merely carry theoretical prejudices of its own, but it can also make theorists blind to data as well.

  Haspelmath:

  > "Next, what about the position of wh-phrases in a clause and other word order properties of the language (§4.4)? Hawkins (2002, §4.3: 2004, §7.3) argues that wh-movement creates filler-gap relationships that cause processing difficulty, and that the processing difficulty is greater if the verb (to which most wh-phrases are connected semantically) is further away."

Concerns: Hawkins himself is working within a robust theoretical framework with complex hierarchical structures. Stripped of that framework the claim does not stand.

- The anti-theoretical prejudice blinds H to oceans of data dating to Ross (1967).

(2)   a. *Who did you hear the story that Bill hit?

        b. Who did you hear that Bill said that John threatened to hit?

(3)   a. *Who did you see Bill and?

        b. Who did you say that Bill and Fred saw?

- Allegedly atheoretical linguistics can not only make you blind to certain facts,
- it can smuggle in robust theoretical assumptions of its own.

Everett (1986):

ti xibɪ´ib-i-hiab-iig-a´ kahaɪ´ kai-sai

for which Everett gives two glosses.

     (i) 'I am not ordering you to make an arrow.' or

     (ii) 'I will not let you make an arrow.'

Not available:

     (iii) I'm not ordering you. Make an arrow!

Everett (1986) this must involve a sentential clause embedded within a sentential clause.

Everett (2005): there is no recursion in Pirahã.

- Did he "unsee" the data?
- Haspelmath on "processing difficulty"
- inductive approaches to language are not merely carrying water for fossilized grammatical theory, but there is a high level empirical hypothesis in the mix as well – typically a thesis to the effect that the underlying biological faculty of language is doing minimal work (perhaps it is just generalized cognition)

Haspelmath's platitude:

> "The narrow range of actually existing organisms is primarily determined by survival (i.e. the chance of successful replication), not by constraints on what the genetic code allows. To study the nature of the cognitive code, we should study the acquisition of unattested language types under natural or artificial conditions, but we should not hope to derive much insight from constraints on attested languages."

- But this is nothing if not deeply theoretical, and it could not help but drive his empirical work.
- Is his story about evolution correct?: "Turing Patterns."
- What about big data and its methodology?

Craig Ventor (2008:

> "Despite my urging that we always look for those big questions, data generation for its own sake continues to be a major impediment to real scientific breakthroughs in genomics. It is not hard to understand why investigators, particularly young scientists, are

satisfied being data generators, as government agencies and some foundations continue to pay out hundreds of millions of dollars for just DNA sequencing or, even worse, microarrays, creating huge datasets but seldom any real scientific insight."

- The minute one begins gathering data one is already fully immersed in theory.
- The claims are pernicious because by claiming their proposals are "frame-work" free they are falsely claiming a position of neutrality, and more importantly, they are giving their proposals an aura of unchallengability.
- The methodology can also get in the way of research. Chomsky:

  By 1953, I came to the same conclusion [as Morris Halle]: if the discovery procedures did not work, it was not because I had failed to formulate them correctly, but because the entire approach was wrong… [S]everal years of intense effort devoted to improving discovery procedures had come to naught, while work I had been doing during the same period on generative grammars and explanatory theory, in almost complete isolation, seemed to be consistently yielding interesting results. (1979: 131)

- Darwin again: A standard view is that he kept the theory of evolution under wraps because he feared criticism from religious authorities.
- Francisco Ayala (2009): Darwin felt compelled to keep his theory under wraps because he was concerned about the reaction of the framework-free naturalists, who were under the theoretical spell of Bacon and Mill.

  Darwin:

  I would suggest to you the advantage, at present, of being very sparing in introducing theory in your papers (I formerly erred much in Geology in that way); let theory guide your observations, but till your reputation is well established, be sparing of publishing theory. It makes persons doubt your observations'' Darwin and Sewel (1903: 323).

- **The moral**: purely inductive approaches to empirical investigation, whether they label themselves as framework-free or something else can be just as hegemonic as the most dogmatically advocated theory. Or religion, for that matter.

2.0 Popper

Loporcaro (1989: 343): "although descriptive simplicity is a desirable goal, it should not be attained at the cost of contradicting actual linguistic data."

Popper (1985): we did not reject the theory of planetary motion because of failed predictions in the motions of the planets. We posited the existence of a hypothetical body that was causing the perturbations. And that amended hypothesis not only accounted for the facts, but it actually led to the discovery of Neptune.

- **Misunderstanding of Popper** (1959): His point was not in any way shape or form a data-first approach to theorizing. Indeed, he argued that the inductive method would get you nowhere.
- His point about falsifiability was a call to develop theories in a top-down manner for which there are bold empirical predictions.

- The kind of science that results from a "framework-free" approach to science is precisely the sort of thing that Popper was warning against. It offers no bold empirical hypotheses – indeed it is effectively banned from doing so.
- Since it is actually itself a theory in disguise it is the worst possible kind of scientific theory – a gerrymandered and unprincipled theory that is crafted specifically to comport with an arbitrary class of data. That suggests a kind of vacuity of theorizing that Popper was warning against.

3.0 Data Coverage

- What data is a theory is actually responsible for?
- Sometimes a class of data is assumed to be a target for a theory but this assumption turns out to be false.
- Assad and Cockburn (1972), Sabin (1981), and Gear (1981: there are dozens of other non-polio enteroviruses (NPEVs) that are capable of causing paralytic polio-like syndromes. It is generally agreed among medical researchers that many of the cases that fell under the rubric of 'Polio' originally were in fact caused by NPEVs.
- Compare to Minimalist Program's goal of achieving a reduction to (or integration with) low level biophysical principles.

4.0 Linguistic Judgments

- Intuitions vs Judgments
- the "voice of competence." Devitt (2006): "I need a word for such special access to facts. I shall call it 'Cartesian'." (96)
- Williamson (2004):

  "What are called 'intuitions' ... are just applications of our ordinary capacities for judgement. We think of them as intuitions when a special kind of scepticism about those capacities is salient. Like scepticism about perception, scepticism about judgement pressures us into conceiving our evidence as facts about our internal psychological states: here, facts about our conscious inclinations to make judgements about some topic rather than facts about the topic itself. But the pressure should be resisted, for it rests on bad epistemology: specifically, on an impossible ideal of unproblematically identifiable evidence." (109)

- Let's follow this advice in the case of linguistics and think of so-called linguistic intuitions as being judgments
- Linguistic judgments are not the object of study and are but one source of data.

Chomsky (1982b).

  To say that linguistics is the study of introspective judgments would be like saying that physics is the study of meter readings, photographs, and so on, but nobody says that. … It just seems absurd to restrict linguistics to the study of introspective judgments, as is very commonly done.

possible sources of data
1. Psycholinguistic Data
2. Corpus Studies
3. Etymological Data
4. Written "Grammars"
5. Phonetic Data
6. Field Research
7. Brain Imaging
8. **Linguistic Intuitions/judgments**
…etc…

Our focus today: #8

Varieties of Linguistic Judgments – a *few* examples

1. Acceptability judgments
- *who did John hear the story that Bill hit
- who did John hear that Bill hit

2. Meaning judgments
2.1 in support of ambiguity and vagueness facts
- lexical
e.g. 'John went to the bank'
- structural
e.g. 'Flying planes can be dangerous'
- scopal
e.g. 'Every man loves some woman'
2.2 in support of binding facts
John said Bill likes himself
(himself can't be John in English, but can in Scandinavian)
- but notice these aren't "binding judgments" -- donkey anaphora
Everyone who owns a donkey beats it ('a donkey' seems to bind 'it')
- combinations of binding and ambiguity judgments -- parasitic gaps
John filed every letter without reading (unambig)
What letter did John file without reading (ambig)
- combinations of the scopal and binding judgments (e.g., inverse linking where   'it' can
be understood as bound only when 'every city' is understood as having   wide scope.) e.g.
'someone from every city hates it'

3. Judgments about possible semantic contents
'that glass of orange juice is large' (for an for a glass of orange juice)
'that glass with orange juice in it is large' (for a glass)

4. Judgments about Entailment Relations
4.1 directional entailingness
e.g. 'Every dog barked'

Entails: 'Every French poodle barked'
Doesn't entail: 'Every animal barked'
4.2 adverb attachment elimination
e.g. 'John buttered the toast with a knife'
Entails 'John buttered the toast'
4.3 substitution in opaque environments
e.g. 'John believes the morning star is visible in the morning'
Doesn't entail: 'John believes the evening star is visible in the evening'

5. Judgments about the interaction of stress and interpretation
- stress and presupposition
e.g. 'John said HIS BOSS was coming to dinner'
(presupposes, someone was coming to dinner)
- stress, focus, scope
e.g. 'John saw WHO?'
(allows Wh in situ, but forces an echo-question reading.
- comma intonation and acceptability
e.g. *'John who is a lawyer arrived' vs. 'John, who is a lawyer, arrived'
*'John ate the haggas with a spoon with a fork' vs 'John at the haggas with a spoon, with a fork, and with a knife'

- linguistic judgments are not judgments about rules (or principles or whatever), or even rule compliance (understood in the sense that we judge that we are in compliance with a particular rule or set of rules that is transparent to us).
- They are simply judgments about linguistic facts or phenomena (these facts are determined by the linguistic rules).

Bogen and Woodward (1988) :

Theory ==> explains/predicts ==> **phenomena** <== is evidence for <== **data**
      <== are evidence for <==

- *Phenomena* (or facts) are stable and replicable effects or processes that are potential objects of explanation and prediction for scientific theories.
- Such facts will thus *provide evidence* for the theory of grammar, and the theory of grammar will in turn play a role in the *explanation and prediction* of at least some of these facts.
- So, facts provide evidence for the theory and the theory explains and predicts the facts.
- *data* is observational evidence for claims about phenomena.
- In the case of linguistics, linguistic *data* provide *evidence for* phenomena (like binding facts or "island effects") that are in turn *explained by* the theory of grammar.

I. A picture of the object of inquiry

- Universal Grammar (UG), -- initial state of the language faculty
- Distinguish between UG, which we can think of as a natural kind, and the instantiation of UG within a particular individual, UG_i.

- We can also distinguish between UG_i and the grammar that the agent *has* by virtue of being in the state UG_i – we can all this grammar G_i.
- Modifying a suggestion from George (1989), we can now distinguish between G_i (the grammar the individual *i* has), the psychogrammar (which is the psychological state by virtual of which the individual has G_i), and the biophysiogrammar (which is the low level biophysical state upon which the psychogrammar supervenes).
- When we are investigating UG we are in effect investigating the psychogrammar and the biophysiogrammar (also construed as natural kinds).
- The principles and parameters framework is in effect an investigation into the psychogrammar and the so-called minimalist program of Chomsky (1995a) is in effect an investigation into the structure of the biophysiogrammar (both at a level of abstraction).
- G_i generates a language (narrowly construed) L_{Gi}
- L_{Gi} accounts for some aspects of i's language broadly construed L_i
  e.g. (a) might be part of i's L_{Gi} but not i's L_i

      (a)   The mouse the cat the dog bit chased ran away

  Just the opposite for (b)

      (b)   That is a convention up with which I will put

  More generally: L_i is a function of L_{Gi} + processing considerations + pragmatic considerations + socio-cultural factors, etc.

Grammatical Rule for G_{PL}
Subjacency: Moved elements can't jump an NP and an S node without an intervening landing site
Explanatory fact about L_{GPL} (potential object of theoretical knowledge for PL)
'[s who did you hear the story that Bill hit]' violates subjacency
Explanatory fact about L_{PL} (potential object of theoretical knowledge for PL)
'who did you hear the story that Bill hit' is unacceptable in L_{PL} because it violates subjacency
Surfacey fact about L_{PL} (potential object of knowledge for PL)
'who did you hear the story that Bill hit' is unacceptable for PL
Datum (content judged by PL)
That a particular tokening of 'who did you hear the story that Bill hit?' is unacceptable to PL
Source of datum (act of judgment by PL)
PL's act of judging that 'who did you hear the story that Bill hit?' is unacceptable

why use judgments? -- inexpensive source of data

Chomsky (1965).

> The critical problem for grammatical theory today is not a paucity of evidence but rather the inadequacy of present theories of language to account for masses of evidence that are hardly open to serious question…. It seems to me that sharpening of the data by more objective tests is a matter of small importance for the problems at hand… Perhaps the day will come when the kinds of data that we now can obtain in abundance will be insufficient to resolve deeper questions concerning the structure of language. (20-21)

Q: Is this still right? Was it ever right?

1. The reproducibility objection

2. The publicity objection

3. The fallibility objection

Are linguistic judgments a total bust?

Householder (1965):

> "I…regard the 'linguistic intuition of the native speaker' as extremely valuable heuristically, but too shifty and variable (both from speaker to speaker and from moment to moment) to be of any criterial value." (15).

Gethin (1990) linguistic judgments are "useless."

Maclay and Sleator (1960) asked beginning rhetoric students (who for some reason they took to be linguistic experts), "Do these words form a grammatical English sentence?" The results:

3 out of 21 said yes to 'Label break to calmed about and'

Only 4 said yes to 'Not if I have anything to do with it.'

(2 of 10 who rejected the following changed their votes to yes when it was pointed out that it was strictly true: 'I never heard a green horse smoke a dozen oranges.')

Hill (1961): sentences drawn from *Syntactic Structures* drew mixed results from experimental subjects. 'the child seems sleeping' was accepted by 4 of the 10 informants (*until* it was paired with 'the child seems to be sleeping' at which point all 10 informants vote negatively. Establishing the contrast helped the informants to see what the task demand was.)

- Spencer (1973): in cultivating "good judges" we are in effect cultivating a group of informants that have an entirely different set of judgments from the average person and that such rarified judgments are in some sense suspect

  > "in recent developments in linguistics the intuitions have become more and more subtle, and more difficult for nonlinguists to intuit themselves or accept. This disturbing development has led to the question of whether or not linguists' intuitions can be uncritically accepted as being valid and basic to the speech community. There are few points of contact between linguistic theory and natural speech phenomena except for the basic intuitions of linguists (Bever, 1968). It is possible that the behavior of producing linguistically relevant intutions has developed into a specialized skill, no longer directly related to the language behavior of the speech community (Bever, 1970)." Spencer (1973; p. 87)

- Swap the term 'chemistry' for 'linguistics' (and 'judgments' for 'intuitions') in the Spencer quote:

  > "in recent developments in ~~linguistics~~ *chemistry* the ~~intuitions~~ *judgments* have become more and more subtle, and more difficult for non ~~linguistics~~ *chemists* to ~~intuit~~ *judge* themselves or accept. This disturbing development has led to the question of whether or not ~~linguists' intuitions~~ *chemists' judgments* can be uncritically accepted as being valid and basic to the ~~speech~~ community of *chemical based organisms*.

Also: Bever's worry is grounded in a falsehood.

Culbertson and Gross: acceptability judgments were drawn from 42 test subjects, of which 7 linguists with PhDs, 17 were students with at least one class in generative syntax, 11 were students with no syntax background but with experience in other areas of cognitive science, and 7 were people with a college level education but no background in cognitive science. The stimuli they used were 73 randomly selected sentences from an introductory textbook by Haegeman and Guéron (1999), including grammatical, ungrammatical and "questionable" sentences.

- Each subject was given a questionnaire and asked to evaluate the sentences for their acceptability, where this task was explained as follows:

  > "A sentence sounds good if you think you would or could say it under appropriate circumstances. By contrast, a sentence sounds bad if you think you would never say it under any circumstances."

- Subjects were then asked to read each sentence and rate it on a scale of 1-4 (1=perfect, 4=terrible).

The results:
- The judgments of the trained linguists, the students with *some* background in linguistics, and the subjects with *some* background in cognitive scientists all converged.
- Subjects with no training at all in linguistics or the cognitive sciences had different judgments.
- So while it seems that some sort of training is necessary, it hardly seems to be the case that the training consists of being indoctrinated into a particular theoretical framework or position within generative linguistics. Indeed, you needn't have even had a course in linguistics!

Bibliography

Absalom, Matthew, and John Hajek, 1997. Raddopiomento sintatico: what happens when the theory is too tight? In Pier Marco Bertinetto, Livio Gaeta, Georgi Jetchev & David Michaels (eds.), *Certamen Phonologicum* II, 159-179. Torino: Rosenberg and Sellier.

Assad, Fakhry, and W. Charles Cockburn. 1972. Four-year study of WHO virus reports on enterovirus other than poliovirus. *Bulletin of The World Health Organization* 46**.** 329–336.

Ayala, Francisco, 2009. Darwin and the scientific method. *PNAS* [Proceedings of the National Academy of Sciences] 106 suppl. 1. 10033-10039.

Bernard, Claude. 1865. *Introduction à l'étude de la médecine expérimentale*. Paris: Éditions Garnier-Flammarion.

Bogen, Jim, and James Woodward, 1988. Saving the phenomena. *Philosophical Review* 97. 3-25.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge: MIT Press.

Chomsky, Noam. 1975. *The Logical structure of linguistic theory*. Chicago: University of Chicago Press.

Chomsky, Noam. 1980. *Rules and representations*. New York: Columbia University Press.

Chomsky, Noam. 1995. *The minimalist program*. Cambridge: MIT Press.

Chomsky, Noam. 2000. Minimalist inquiries: the framework. In R. Martin, D. Michaels, and J. Uriagereka (eds.), *Step by step: Essays in honor of Howard Lasnik*, 89-155. Cambridge: MIT Press.

Chomsky, Noam. 2001a. Derivation by phase. In Michael Kenstowicz (ed.), *Ken Hale: A life in language*. Cambridge: MIT Press, 1-52.

Chomsky, Noam. 2001b. Beyond Explanatory Adequacy. Manuscript, Dept. of Linguistics and Philosophy, MIT.

Darwin, Francis., and Albert Seward (eds.). 1903. *More letters of Charles Darwin: A record of his work in a series of hitherto unpublished letters*. London: John Murray.

Devitt, Michael. 2006. *Ignorance of language*. Oxford: Oxford University Press.

Everett, Daniel. 1986. Pirahã. *Handbook of Amazonian languages*, vol. 1, ed. by Desmond C. Derbyshire and Geoffrey K. Pullum, 200-325. Berlin: Mouton de Gruyter.

Everett, Daniel. 2005. Cultural constraints on grammar and cognition in Pirahã. *Current Anthropology* 46. 621–46.

Fodor, Jerry. 1984. Observation reconsidered. *Philosophy of Science* 51. 23-42.

Gear, James. 1981. Non-polio causes of polio-like paralytic syndromes. *Review of Infectious Diseases* 6. 5379–5384.

Ghiselin, Michael. 1969. *The triumph of the Darwinian method*. Berkeley: University of California Press.

Gould, Stephen Jay. 1990. Down on the farm. *The New York Review of Books*. January 18, 1990. http://www.nybooks.com/articles/1990/01/18/down-on-the-farm/ (accessed February 15, 2018).

Hanson, N. Russell. 1958. *Patterns of discovery*. Cambridge: Cambridge University Press.

Haspelmath, Martin. 2010. Framework-free grammatical theory. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 341-365. Oxford: Oxford University Press.

Hawkins, John. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Jacob, François. 1988. *The statue within: An autobiography*. New York: Basic Books.

Kuhn, Thomas. 1970. *The structure of scientific revolutions*, 2nd edn. Chicago: University of Chicago Press.

Lévi-Strauss, Claude. 1953. "Remarks." In Sol Tax and Loren Eiseley (eds.), *An appraisal of anthropology today*, 349-352. Chicago: University of Chicago Press.

Ludlow, Peter. 2011. *The philosophy of generative linguistics*. Oxford: Oxford University Press.

Nevins, Andrew, David Pesetsky, and Cilene Rodrigues. 2009. Pirahã exceptionality: A reassessment. *Language* 85. 355-404.

Popper, Karl. 1959. *The logic of scientific discovery*. London: Hutchinson.

Popper, Karl. 1985, *Unended quest: An intellectual autobiography*. La Salle: Open Court Publishing.

Ross., John. 1967. *Constraints on variables in syntax*. PhD Dissertation, MIT.

Sabin Albert. 1981. Paralytic poliomyelitis: Old dogmas and new perspectives. *Review of Infectious Diseases* 3. 343–64.

Salles, Raiane Oliveira, 2015. Understanding recursion and looking for self-embedding in Pirahã: The case of possessive constructions. Dissertação de Mestrado, Potífica Universidade Cattólica do Rio de Janiero.

Sauerland, Uli. 2015. False speech reports in Pirahã: A comprehension experiment. http://ling.auf.net/lingbuzz/002850.

Schütze, Carson., 1996. *The empirical basis of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Sellars, Wilfred. 1949. Acquaintance and description again. *The Journal of Philosophy* 46. 496-504.

Sober, Elliot. 1975. *Simplicity*. Oxford: Oxford University Press.

Stott, Rebecca. 2003. *Darwin and the barnacle: The story of one tiny creature and history's most spectacular scientific breakthrough*. New York: W.W. Norton.

Turing, Alan. 1952. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London B* 237. 37–72.

Williamson, Timothy. 2004. Philosophical "intuitions" and scepticism about judgement. *Dialectica* 58. 109-153.