Mindbuilding Notes, Part One

By "mindbuilding", I mean the process of creating the database that forms Sphinx's initial concept knowledge base (as opposed to coding Sphinx's heuristics or word recognition). The idea is not to pre-load all knowledge into Sphinx's brain, but rather to give it a minimal architecture of concepts for new knowledge to position itself within. "Minimal", of course, in a complex system, is rather different from "negligible". This is some of the most challenging intellectual work I've ever done—it involves creating the initial neural network for several ~~hundred~~ thousand more abstract concepts, ranging from **building_trades* to **genetic_fallacy* to **losing_touch_with_reality*.

Though I'm not at all finished with this stage of work, here are a bunch of thoughts:

• There is, I am vaguely aware, a purist school of AI that rejects this whole project. Earlier, when I was blogging about Sphinx, Nathen mentioned the possibility of building up associations "from scratch" by data-mining texts. This is, debatably, what human infants do. I would guess that human cognition has a suite of truly innate, endowed patterns (circles, for instance, or color contrast, or acoustic dynamics) that help to get the ball rolling. But Sphinx is going to having an "innate" knowledge of **connecticut* and **apostolic_succession,* whereas humans clearly only learn such patterns through the front door, as it were. Both methods cause problems.

It seems pretty clear (thank you Chomsky) that humans begin with a working grammatical template, and fit words-and-or-concepts into that architecture as we decode them. That this template works as far as *we* are concerned is demonstrated by our species' skill at writing poems and building surface-to-air missiles. But of course, this is a sort of tautological demonstration: the entire human intellectual sphere is the product of our cognitive templates, so we are not in a good position to judge the alternatives to them. Perhaps if we were not hard-wired to use triune conjunctions of concepts, we would have more poems and fewer missiles, or vice versa, or something magnificent and different altogether. We'll never know, though there are glimpses. Our inability to intuit probabilities, for instance, has been catastrophic, and certainly seems to be both hard-wired and, in a sense, unnecessary.

My own feeling (strongly borne out by working on different versions of Sphinx) is that the design of the ur-templates that new concepts get added to is *completely crucial.* Even minor flaws will eventually expand into huge headaches. If an AI learns almost everything through the front door, as humans do, presumably it must go through a long infancy before it is even possible to tell whether or not the template works. That, for me, is the primary argument in favor of mindbuilding. I don't need Sphinx to know about Connecticut from zero hour, but if Sphinx thinks that Connecticut is in Asia, or is a temporal adverb, then I want to know about that *before* zero hour.

• There is a secondary problem: humans lavish semiotic attention on infants for years, but no one is going to do that with an AI that can't talk. Indeed, no one is going to do that with an AI that can't hold its ground fairly well in a conversation. If Sphinx doesn't hit the ground interesting and useful, it doesn't matter how much potential it has to learn new stuff, because no one will teach it anything.

• We began this process after Zoë and I had "tagged" 305 (or so, all the following numbers are a bit slippery) resources, almost all of them books. This was not quite a random sampling of All Possible Concepts; rather, they were the 305 easiest-to-tag resources out of about 7,500 resources that were recommended by 38 different authorities, prioritizing the ones that had been heavily cross-recommended. Put more simply: these were mainly the easiest books to tag out of the most popular books.

It's still a pretty random list, though, like woah. Any list that includes "the planet Mercury", "gyotaku", "auto-fellatio" and "the siege of Seringapatam" is random in my book. So it is understandable that these 305 initial concepts would produce a large number of downstream concepts.

In fact, they linked to about 3100 other concepts, of which 2000 were unique. This gives a first-generation multiplier of 6.6, with a Simpson coefficient of 0.03: much higher-diversity than I expected.

• A number of the second-generation links were things like geographical locations, biological species, planets, etc. It seemed economical to handle these by processing large lists of "cousin" concepts that all have roughly the same format. Those additions bring our 2nd generation up by several thousand, but they add almost nothing in the way of new downstream concepts. The 744 aforementioned musical instruments add only twenty or so new concepts downstream (a multiplier of about 0.015). Presumably this will be the case with any collection of formatted data. Again, it suggests a possible metric for whether or not an idea is interesting: h*ow different is this concept from its group peers?*

• Not counting cousins, there were roughly 1124 independent concepts in the second generation. These including things like "god" and "paradox" and "the color barrier in baseball", all of which takes us towards the core network. These 1124 indies, in turn, produced about 1450 indies downstream...a multiplier of 1.3. From there it gets a bit murkier. When I defined the next batch of indy concepts (1139 concepts, so about the same size run, but not the whole generation), the total rose to about 3580 unique concepts. The implied multiplier is down to 0.9 (thank god!), but it's biased by that fact that I postponed defining the "worst" concepts. I'll try to do a much more precise analysis after the music ends.

The whole think feels like some awful game of semantic whack-a-mole, spanning months of work. One little reference to kayaks and then I have downstream links to hulls and the Inuit, and then to hydrodynamics and morphology and Amerindian peoples, and so on and so on. It's wonderfully fun, but I've gotten a little gunshy about concepts that I know are going to sprawl out when I try to define them. For instance, "non-profit organization" sounds innocuous enough, until you realize that fully defining it requires a whole Econ 101 lexicon.

• In terms of the previous article, the work of mindbuilding is to create a core network by building inwards *from the periphery.* At some point all the downstream links from "acute triangles" or "southern belles" or "anger at god" enter into a great vortex with the downstream links of the other 305 source nodes. At least a few of the sources have already been incorporated into this nascent core network: I know the *Odyssey* and the *Inferno* have, for instance. But based on our current methods, we will close all the links while leaving a lot of stuff outside the core. And that's fine: it gives Sphinx something to do later.

It does raise an issue about the relative size of the core, though. If we had started with 35 source nodes, instead of 305, we would be done by now, and the core would be fairly small. As it is, I am anticipating a "deep" core of about 200 concepts, which are strictly used in each other's definitions, and a much larger core if we count connotations, examples, etc. Before we get there, though, I'm betting we have a 10,000-concept database.

• Defining concepts is a beautiful, beautiful, nightmare. Susannah and I spent much of a car ride across New Hampshire debating the definition for one word, *body.* We settled on "physical aspects [of things] that emphasize contiguity." I'm proud of that. But the fact remains that we use the word *body* all the time, quite easily, based on ostensive definition that we have an extremely hard time formalizing. I would go on to argue that the conjunctions and metaconceptual verbs like "pertain" don't really have formal definitions. That is to say, no non-expert can provide a formal definition for these words, and if you put the experts in different interrogation rooms, they won't give you the same definitions, either. Those words are our intrinsic grammatical template poking through the fabric of the language.

• Human knowledge is full of these naïve categories that work perfectly well until you try to systematize them, and then they become quicksands of ambiguity. Low-level definitional problems have a way of contaminating higher-level summaries. For instance, the question of how many countries there are on earth can vary by several dozen depending on who recognizes whom. Most of these disputes are geopolitically trivial, but not all of them: Palestine, Taiwan, etc.. And even the minor ones affect other data to a surprising degree: for instance, it is unquestioned that Russia and Canada are the first and second largest countries by area, but the bronze medal goes to either China or the US, depending on one's political biases. One might like to think that most human beings are epistemic peers on the big stuff, at least if you avoid some oddly controversial issues like evolution. But the fact that we can't even agree on what the third largest country on the planet is, illustrates how hard it is to avoid epistemic divergence.

• In the past, I've tried to mindbuild with epistemic metadata on every single piece of information. This was in keeping with the Epicurean principle of *isosthenia,* which is dear to my heart...but I don't think it is actually realistic. Aside from being an incredible pain in the ass and a processing-time drain, I don't think it accurately reflects human epistemology. Most people cannot remember specifically *why* they believe most of the factual information they believe. We don't say "*I know flamingos are pink because I saw them in the David Attenborough series* Planet Earth *on March 3, 2002, and I have confirmed that by observing pink plastic flamingos used as lawn ornaments in Youngstown, Pennsylvania, on July 15th, 2006, and also...*" Hell no. We say: "*Someone credible must have told me flamingos are pink once upon a time, and that belief hasn't caused me any problems yet.*"

• The set-theory approach to definition handles many concepts very neatly. "Absent-minded professors" is the intersection of the set "absent-minded people" and "professors", and that's really all you need to know, as far as the definition is concerned. While this is appealingly straightforward, attempts to push this approach inevitably lead to dummy sets that clearly do not exist in our minds as concepts in their own right: *weddings of poets from former European countries; ethnically Jewish atheists who have contributed to edited volumes of chemistry.* Phrases like these abound with logicians and a certain class of reference books. I'm trying to avoid them. Every time I feel the urge to create such a dummy set, I can infer that I'm missing some lateral mode of definition.

• Defining the core concepts for our 305 resources required about 60 different "lateral" relational concepts. This number will probably drop slightly, but most of these conjunctions have a primal smell, as if they are hard-wired into our thinking in an irreducible way, much like the verbs I listed that appear repeatedly in Webster's definitions.

• This minimalist coding creates a little Searlian philosophical crisis. What does it mean that we can take concepts as complex and poignant as "the loss of innocence" or "self-awareness" and define them with *two links each?* The tao that can be spoken is not the eternal Tao, and the idea that can be reduced to two links seems like no idea at all. Of course, we can enrich our understanding of "loss of innocence" by looking upstream for examples, connotations, and other things that orbit, but are not part of, its definition. But each of those is just another concept, defined in three links, or seven...it hardly matters to the argument in view, since ultimately they are all defined by swirling down into the core.

I think that this sense of atomization (or claustrophobia?) blinds us to a very important insight. When we talk about a specific idea like "starlings" or "the siege of Seringapatam" or "the fluidity of meaning", we are simply talking about a plumbing fixture: a little T-intersection, or something slightly, arbitrarily, more complex. But *thought* is not a plumbing fixture in this metaphor; it is the water in the pipes. We don't say: *my plumbing is only ever connected to more plumbing, so how can it do anything* *important?*

To ground this discussion a bit, consider the boilerplate initial interaction I have whenever I talk with Sphinx:

> **Hello, this is Sphinx.**
> **this is ethan**

Seven words; five of them unique; 36 characters. And two of the words are proper nouns, and another has no real definition—"hello" is just a polite placeholder. But in the interim, the Sphinx typically calls 240 lines of code in Riddle, maybe 7000 lines in PHP. A poor showing of thoughtfulness, no doubt, but already orders of magnitude more complex than we might suppose just by looking at the words themselves.