The Jig is Up

(posted from Bern, Switzerland)

At around 2,500 concepts in the main database, I began to notice inefficiencies creeping into my own process. Benoit Mandelbrot is from Warsaw, fine, but had I coded Warsaw as **warsaw* or **warsaw_city* or **warsaw_city_[poland]*? I was usually fairly consistent about these issues. Still...if I had any doubts, it would take me moment to check—but if I guessed and got it wrong, it might take a good deal longer to repair the damage. And Zoë had been helping me code things, so inevitably the two of us had coded things slightly differently, and so on, and so forth... These little issues are the grit in the wheels of all data entry projects, and at some point the friction is enough to stop the wheels from turning. (The point, for instance, when you accidentally scramble 25 kilobytes of painstakingly entered data and have to rebuild it.) Happily, I had a another way forward.

Clearly it was time to quit mindbuilding as I've been doing for the last six months or so, upload the data, and start working on it "through the front door" by utilizing Sphinx's algorithms. On the one hand, I hadn't finished bringing the 305 texts we started with full circle to the deep core network. On the other hand, I felt like I was now trying to maneuver an ever-larger sofa up a twisty stairwell, as it were. The sofa being Sphinx's mind, and the stairwell being my own mind's carrying capacity. As I've noted a while back, this outer perimeter beyond which I can't think consistently is a recurring design constrain, and it forces a lot of bootstrapping and otherwise-absurd design choices, much like the way that almost everything on a typical construction site has to arrive in chunks of 96 pounds or less: just exactly what a well-built dude can hope to carry without killing himself. There is a similar logic at work here. You can build something you can't lift—or *understand(!)*—but you have to build it out of blocks that you can handle.

So I set July 4th as the U-day. I didn't quite make it, but I was fairly close, and I've now loaded up 18,181 concepts and 21,875 word-forms. So I am quite excited. Together, this comes to about 5.5 megabytes—or perhaps 1,400 pages worth of plaintext—so I feel a bit better about (a) the fact that it took six months, and (b) my inability to keep it all straight in spreadsheet form. I should note that this 5.5 Mb curb weight for Sphinx' "starter mind" could be compressed very much further than it is. Again, I've learned that to avoid delays from my own processing limitations, the coding has to be more or less human-readable. So there are a series of compromises between my limitations and Sphinx' limitations...

Sphinx' conceptual and lexical inventory does not resemble any actual human being's, even a young child like my niece (whose only two grammatical categories seem to be pronouns and bird-sounds). 17k concepts is not that many—not enough to even finish defining the 305 we started with. Moreover, many of the concepts are quite esoteric: taxonomic categories, atomic elements, geographical locations, and the like. This is why I could get to 17k even though my brain was shutting off at 2.5k: there are large swaths of concepts that are, internally, quite simple to navigate. The moons of Saturn, for instance, remain pretty much out by Saturn and do not cause any trouble for the rest of whatever neural network they get dropped into. If I've accidentally switched the data for "Makemake" and "Sedna", that is a problem I can deal with later on. If I've switched the data for "becoming" and "being", then pretty much everything else is wrong, too. An interesting distinction.

Words Words Words

Despite the existence of the Oprah neuron, etc., I expect that human brains do not have atomic concepts in the way Sphinx does, but something more like associative clusters. Where words are concerned, we can make a more direct comparison. Sphinx knows about twenty-two thousand word-forms at the moment. About half of these 3,000 of these are proper nouns, including book titles and the like. We normally ignore these in discussions of vocabulary, though clearly that raises some tantalizing questions. (For instance, when I ask people how many individual people's names they think they know, I get answers ranging across three orders of magnitude). Another several hundred are words that Sphinx knows structurally, but not conceptually, like a kid cramming for a spelling bee.

The remainder, after we collapse variant forms and the like, comes to about 6,300 actual words, giving plenty of latitude for the ambiguity in the word "words". According to TestYourVocab's massive data, this is about the equivalent of the 40th percentile for 6-year-old first-language English speakers. (There is a considerable self-selection bias on their site in favor of larger vocabularies, so we can more realistically say that

Sphinx knows about as many words as a typical or precious six-year-old.) But precocious or not, no six-year-old human would have a vocabulary like Sphinx. Picking three words at random, I get *glass armonica, displeasure,* and *oldfieldthomasiidae.* So presumably Sphinx is missing a few thousand words along the lines of *moo* and *boogers* and *cheat code.* But that will shake out later.

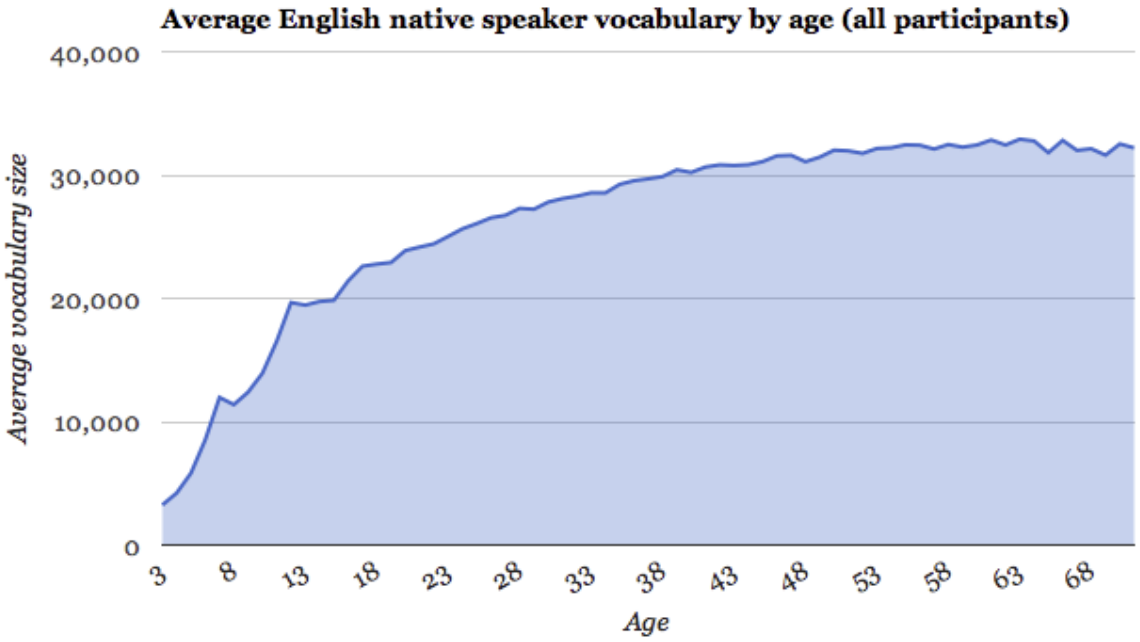
TYV estimates that the typical *adult* who drops by their website has a first-language vocabulary of about 30,000 words, with a standard deviation of about 3,500. (The extrapolated average for the overall population of adult EFL speakers is perhaps 20,000 words.) If you look at their numbers, it would appear that vocabulary gains are quadratic from ages 3 to 7 {Bethel of +,+ at 0.1}, but thereafter the gains become linear. They note a very stable pattern of one-word-per-day from about age 15 to 32, which seems to be independent of verbal SAT scores and other such metrics. Finally, the vocabulary fully plateaus by about age 50, and actually starts to decline (though plausibly there is a cohort effect at work here: the boomers may simply have had a smaller English than my generation).

I expect there is a dimension missing in these numbers, though, which Sphinx' bizarre lexicon highlights. Between age 20 and 35, let's say, many people learn a specialized vocabulary of trade or scientific words like *gauge corner crack* or *box scraper* or *oldfieldthomasiidae.* These words will never appear on vocab tests—certainly not on TYV's tests, anyhow—because they are so rare on the ground that it would waste people's time to ask about them. So there is no way for TYV to notice if you happen to know the Latin names of all 4,000 rodent species, except insofar as this might covary with knowing a few marginally more common words like *midden* or *plantigrade*. Medical and legal students, for instance, assimilate an enormous specialized vocabulary which largely does not appear on these tests. Interestingly, the ten-dollar vocab words that do appear on tests are things like *uxorious* or *redolent.* I would suggest that these should not be viewed as the top shelf of "neutral" vocabulary, but rather as the expert jargon of English Lit majors, which through some metonymy has become a stand-in for vocabulary in general. There is no question in my mind that an English major with an actual vocabulary of 35k would outscore a med student with an actual vocabulary of 40k on any vocab test.

But below about 30k, the distinction between people's vocabulary sizes has to be largely made up of words that are not field-specific jargon. English excels in this class of words: near-synonyms for relatively common concepts. Ours is a language, after all, in which we have ten distinct words like *luck, chance, fate, destiny, odds, probability, happenstance, likelihood, possibility,* and *potentiality,* any one of which could be used almost interchangeably, and none of which is restricted to a specialized field. And this is in large part a product of history: ours is a Germanic language that absorbed most of the lexicon of a Latin language simply to create a second tier of class-connoted synonyms. What is bizarre (and annoying) is that at something like a million words, we still have polysemes, let alone antonymic polysemes like "cleave".

These reflections are relevant to Sphinx insofar as I want Sphinx to have a general sense of how "advanced" a given term is, especially in the all-too-common-in-English situation where we have strict synonyms like *anger* and *ire.* Anger is the "easier" word of the two, even though it is longer and has the same syllable count. By "easier", then, we mean that there is an implied sequence: people who know the word "ire" are apt to already know the word "anger", but not vice versa. (Notably the chief connotative distinction between such words is often simply their placement on this sequence, rather than any further semantic inference.) This sequence of words is never going to be identical from one person to the next, but it is probably fairly standard for the first 20,000 words (or about age 16-18). Beyond that, things get murky fast. Which is the easier term: *lapillistone, mesopredator release,* or *autarky?* The question no longer makes much sense: each of those terms is specific to a field of study. Advanced study, to be sure. Anyone who knows any of those terms probably also knows *purple* and *flammable.* But it makes little sense to compare *lapillistone* and *autarky* directly. The assumption in lang-acq research seems to be that "difficulty" corresponds to infrequency in any given corpus, which must be roughly true, but I'm suspicious that that's too easy.

My takeaway from this is that there are probably twenty to thirty thousand "sequential" vocab words in English, in the sense that we can say something along the lines of "*pretzel* is typically the 5,632nd word learned, and *sesquipedalian* is typically the 25,123rd word learned." Of course such rankings won't be precise, but they will aggregate to a decent ballpark estimate. Beyond about 30,000 words, there is no longer really a spectrum of likely encounters; rather, there's a sort of bushy mess.



The chart above is from TYV; others and interactive versions on their site.

Wandering in Rumsfeldspace

We started with 305 books, of which I had read perhaps 50 or 60. Most of the others I was familiar enough with to parse into categories like "I should read this" or "I'm never gonna read that", or (frequently) "I don't need to read that, because I'm pretty sure I already know what it's about." When I uploaded this list, I felt like I was posting something I understood quite well. The unknowns were, as Rumsfeld infamously put it, *known* unknowns.

Since then, I have uploaded...ah...about 6500 other texts. While I'm familiar with hundreds of these, they are *dwarfed* by the number of texts that I'm quite unfamiliar with. Even after doing topiary work on these databases for half a year, I am still encountering titles that I don't recognize at all. Sphinx doesn't know much about all 6500 of these texts, either, but actually it knows enough already to make fairly solid recommendations. In a trial run, it recommended that I read Ishmael Reed's *Mumbo Jumbo* and two Thomas Pynchon novels (*Gravity's Rainbow* and *The Crying of Lot 49)*. So far, so good: those seem like quite solid recommendations based on my own subjectivities, and Reed, in particular, came at me totally out of left field.

The difference between 305 texts and 6800 texts is not just quantitative: I am left feeling that, in a real sense, Sphinx now knows things of interest to me that I don't know, or even know it knows. I do not have time left in my life to read 6500 books. The problem of intellectual triage, always present, has become more palpable as I get older. And Sphinx can now make informed recommendations to me across perhaps 4000 books that *I've never even heard of.* Unknown unknowns. For me, his is the first hint of Sphinx as a functional tool, as well as a fascinating project.

As a side note, it is interesting to note that there are diminishing returns on adding new book lists to this corpus. Bloom once wrote that the problem of the reader in the modern era is one of "Malthusian repleteness", a phrase that has stuck with me. There are far too many books, and there are (pace Bloom) many different reasons to read them. But the fact is that when serious readers of any persuasion compile their lists, they keep returning to many of the same titles. Obviously, given a certain degree of esoterica and intersectionality, you find new titles with each list. The list for **murder_mysteries_by_bisexual_expatriate_herpetologists* is going to include a few entries that Sphinx hasn't seen before. But at nearly-7000-and-counting, that list will probably include a few titles that Sphinx *has* seen before, and if that's true, it's rather enlightening news. Unless we are trying to read every single MMbBEH text, surely we should prioritize the ones that appear on other lists as well? After all, it is

hard to think less of a book that *someone* has recommended it.

A case in point: Flawless Logic, a white supremacist neofascist group, puts out a suggested reading list that includes, among other things, *The Lord of the Rings* and (only somewhat less famously) *Tarka the Otter*. Fans of Tolkein and Williamson would probably do well to think about *why* FL gave those books such a dubious honor. Middle-Earth is not the only fantasy world that might be interpreted as racist, and surely it is not the most racist world in its genre. But it may very well be the most racist world per unit volume of its other literary achievements—and if that is Flawless Logic's logic, as I think it is, then even their reprehensible fringe list references a core set of values shared by readers who are not also neonazis.

Much more on this point later...

Paperless

The dream of the paperless office, much like my fashion empire, is not yet. To Zoë's amusement, I spent a lot of time printing out versions of the database(s), and I still maintain that it is faster to proofread data in hard copy than on a spreadsheet (though it is faster still for Sphinx to do it heuristically). I also spent awhile trying to create network diagrams, which—as I've mentioned—I have a love/hate relationship with. Here's what is left of my notes:

Jigs and Kolks

As I shift from working in spreadsheet format to working via the (still-very-narrow) "front door" of Sphinx' interface, I am reflecting a good deal on the idea of "jigs". Jigs or meta-tools are probably among the earliest technologies. In the toolkit of Ötzi the iceman (†c.3300 BC) there are two different tools whose function was probably to repair his other tools. This requires a certain kind of roundabout advance planning that is a signature feature of human intelligence, but which only really exploded in the industrial age. It is one thing to say "I'm off to hunt some ducks, therefore I need to carry a grindstone." It is something else to say "I want cupcakes, so I'll have to build a machine that can calibrate the solid-state gyroscopes in my drones." This latter sort of thinking is the very marrow of modern engineering, but the shift has been a recent one.

Like many Yankee farms, our farm is full of little "shops" or cluttered ruins of shops. Some of this clutter consists of antique tools, as individually made as Ötzi's. Much of it consists of jigs. Both my grandfathers were engineers, and both of them accumulated a variety of lovingly crafted jigs, which often looked like quite daunting engineering projects in their own right: George, for instance, had a fondness for building jigs out of drilled and tapped blocks of solid aluminum. I remember as a child wondering what possible purpose

these could have served—some project so important that it required building an essentially disposable tool of such enormous complexity.

In fact, of course, it is quite possible that the project was less interesting than the jig itself. There is nothing written on the gates of the universe that says making jigs is actually efficient. If you are making five million cupcakes, by all means, you should build a factory. But if you are making a *dozen* cupcakes, it might be a waste of time to even dig the electric mixer out of the back of the cupboard.

These thoughts come to rest on intelligence. Intelligence, whether human or artificial, is the great jig. Intelligence has no function, on its own, but it can be applied to almost any function, in ever-more-roundabout ways. As I move from the stupid, straightforward world of spreadsheets into Sphinx' neural network and heuristics, I keep noticing this. As the highlighter-daubed papers above suggest, I could go blind and crazy trying to isolate a pattern in the database that Sphinx can pinpoint without breaking a sweat. Cleaning up data is no great intellectual achievement, whether I do it by hand or through Sphinx. But if I work through Sphinx, it leaves behind a set of heuristics which can be used again, expanded, and built on...to what ultimate end? Sphinx' initial purpose is to help people find educational resources: “access to tools”, in Stewart Brand's old phrase. But in reality, access to meta-tools. So when I code a new heuristic, I am building a jig to help build a jig to help people find jigs that will help them build jigs...it's jigs all the way down.

There is a common sentiment that the human mind is highly adapted for all the various needs of human society, whether that means arguing a legal case or studying subatomic particles. I think this is exactly backwards. Human society is the habitat created by our minds, and reflects all the various ways our minds behave, many of which might be construed as maladaptive from some other vantage point. The analogy I keep returning to is the rock-cut basins along the river near here. A boulder gets caught up in a kolk vortex, and scours a hole deep into the limestone, and of course it fits quite neatly in that hole. If the boulder then thinks “see how well I am adapted to the shape of this hole!”, it is not really seeing the whole picture. The mind is not a jig whose *function* is to create a certain kind of hole in reality. We are just stuck in the hole in reality that our minds happened to create.