

Re: Comparing languages

Source: <http://sci.tech-archive.net/Archive/sci.lang/2004-07/1607.html>

From: Brian M. Scott (*b.scott_at_csuo.ohio.edu*)

Date: 07/15/04

Date: Thu, 15 Jul 2004 05:10:14 -0400

On Wed, 14 Jul 2004 04:17:22 GMT Nathan Sanders
<nsanders.DIE.SPAM@wso.williams.edu> wrote in
<news:nsanders.DIE.SPAM-E23BFF.00171914072004@news.verizon.net>
in sci.lang:

> *In article <ccshds\$7fp\$1@news-reader4.wanadoo.fr>,
> "pierre.levy" <pierre.levy11@wanadoo.fr> wrote:*

[...]

> *Comments on the article you quoted:*

The thing's so poorly written that it took me a while to figure out what was going on. It's actually pretty simple-minded, but there is some substance, and a few of your criticisms aren't actually justified.

[...]

>> *vorto (word, as composed of certain sounds or letters without regard to its
>> meaning, e. g. child, children, give, gives, gave and given are six
>> different words),*

> *I'd like to know how they scientifically define "word" in a usable,
> cross-linguistic sense. Professional linguists have a hard enough
> time doing this. Somehow I doubt that a non-linguist with an agenda
> has managed to do it.*

I suspect that the author looked only at written languages and took words to be essentially those entities bounded by spaces. Of course this runs into the problem of differing conventions (German <Versicherungsvertreter>, one 'word', versus English <insurance agent>, two 'words') even without getting into languages like Chukchi.

>> *As all the
>> words in Esperanto are only combinations of invariable elements, the study
>> of the international language is the most simple case to begin with.*

- > *?!?!? This is certainly not true. You only need to look at the*
- > *prepositions to know this: <de> can mean "of", "from", or "possessed*
- > *by" (and lord knows what else!), while <al> can mean "at", "to", or*
- > *"towards", or can act as a marker for "indirect objects".*

The objection is irrelevant, because you've misunderstood what he's saying. He's not talking about semantics at all; he's talking only about the sequences of letters (or, if he's being just a little more sophisticated, phonemes, but I'd take no bets on that). He's merely noting that Esperanto has nothing like the English plural morphophoneme with its several shapes, some completely irregular, or like German strong verbs.

He does not, however, appear to recognize that Chinese or some of the polysynthetic languages might be equally simple in this respect.

[...]

- >> *aforementioned table does not give definitive values to these constants,*
- >> *because the material of the different languages studied is not adequate*
- >> *in the sense mentioned above and needs additional research.*

> *Oh. So the research hasn't even been done yet?*

This isn't a fair objection. There is nothing wrong in principle with doing what amounts to a pilot study, estimating theoretical parameters using small samples to get an idea of what to look for (and to determine whether the proposed model actually fits the data at all well in the first place).

- >> *1) The new method of statistical research is much superior to the previous*
- >> *ones in speed and gives results with remarkably less effort than these.*

> *"Gives results"? They don't have any yet! (see previous quote and*
> *comment)*

Not true. The author apparently presents some genuine results in Figs. 1 – 4, which we haven't seen.

- >> *3) The mathematical expression of this law holds two constants, which*
- >> *vary from language to language, and are indicators of the ease of learning*
- >> *the considered language, or of the expressiveness of the same.*

> *"Or"? So, for some languages the constants indicate ease of*
> *learning, and for others it represents expressiveness? How do we know*
> *which constants go with which indication? Is this pre-determined by*
> *the theory, or do we find out after the fact?*

- > *And how do they know this? Have they tested it? Have they defined*
- > *"easy to learn" and "expressiveness" and compared them with these*
- > *magical "constants" (which, by the way, aren't actually constant,*
- > *since they vary from language to language!)?*

Here again you go too far. The claim is that associated with each language is a pair of numerical parameters. For each language it's a fixed pair. They are claimed to be constants in much the same sense that atomic weight is a constant: different isotopes have different atomic weights, but for each isotope the atomic weight is a constant.

[...]

- > *It only shows this because the author asserts it! There's no*
- > *conclusion, just an assertion, a bunch of intermediate mumbo-jumbo*
- > *that's wrong and/or just plain stupid, and then a restatement of the*
- > *initial assertion under the guise of a proved conclusion.*

This isn't quite correct. The claims about simplicity are nonsense, but not completely unmotivated nonsense, and there is a mildly interesting claim buried in this ill-written summary. I'm going to try to explain a slight variation of what he claims to have done, simply to avoid his notion of 'cliché', but the idea should become clear enough.

Imagine that you have a list of all English words arranged in order of decreasing frequency. (To be definite, take 'word' here to correspond more or less to his 'paradigm': a dictionary entry together with all inflexional variants.) For each positive integer n let $L(n)$ be the set consisting of the n most frequent words. (In his example his set I corresponds to $L(200)$, his set II to $L(350)$, and so on, if you change my 'words' to his 'clichés'.)

Now take as a dataset a large sample, T , of text; say T contains N word tokens. For each positive integer n let $d(n)$ be the number of tokens not in $L(n)$, and let $x(n) = d(n)/N$, the fraction of tokens in T that are not in $L(n)$. Clearly $x(n)$ decreases as n increases.

His y is my n , and for a given value of y (and hence of my n) his x is the percentage corresponding to my fraction $x(n)$, i.e., $100 * x(n)$.

His claim is that for any given language $\log(x)$ is a linear function of y . The corresponding claim in my setting would be that there are constants $a > 0$ and $b > 0$ such that to a decent approximation $\log(x(n)) = -an + b$, and that any other substantial English text T' would yield approximately the same constants a and b . Equivalently, there are positive

sci.lang: Re: Comparing languages

constants k and C such that $x(n) = C \cdot 2^{-kn}$. (He writes his version a bit differently, but it boils down to the same thing. His D is my $1/k$, and his B can be calculated from my C and k .)

If k is large, $x(n)$ falls off rapidly with increasing n ; if k is small, the decay is less rapid. Recall that $x(n)$ is the fraction of tokens in T that aren't in $L(n)$; if your working vocabulary is limited to $L(n)$, $x(n)$ is the fraction of tokens that you don't know. If this decreases rapidly with increasing n , knowledge of a relatively small lexicon will suffice to recognize most tokens, so there is some faint connection with simplicity, at least for languages with the same value of C . (Different values of C complicate the comparison significantly.)

Of course most of us recognize that there's a great deal more to learning a language than learning vocabulary, but he has in fact provided a numerical measure of one small part of the difficulty of learning a language. Or rather, he's done so *if* his empirical observation that $\log(x(n))$ is approximately a linear function of n actually holds up. And it may do so, because all of this turns out (after a little mathematical analysis) to be nothing more than a minor variation on Zipf's Law. (Of course that also means that it's pretty much old hat.)

Brian