The size of languages

A Cautionary Tale

K. P. Hart

Faculty EEMCS
TU Delft

Cyberspace, 30 July, 2020: 16:00 (CET)

What happened?

Twitter of course.
Back in October there was a short discussion on twitter about the nature of books.

A bit condensed: a book is a finite sequence of symbols from some alphabet (including spaces, punctuation, etc).
As such it is not much different from a number.

What happened?

But also: given a suitable ‘alphabet’ A and a suitably large natural number N the set

$$A^N$$

of sequences contains (representations) of all known books in the English language say.

Not only that but also many novels and books on Set-Theoretic Topology that have not been written yet.

A free book

In particular the one that Alan and I once planned to write; I'm glad so say it is finished and you can find it, for free, in that set.

Be sure to take the 2025 edition; it has fewer typos than next year's version.

But seriously

In the paper, by Paul M. Postal, that argued about this status of past, present and future books there was a curious sentence:

. . . , one can show further that the universe of books is truly vast, amounting to what is called a *proper class* in some varieties of set theory.

Well, I could not let that one lie.

The Vastness of Natural Languages

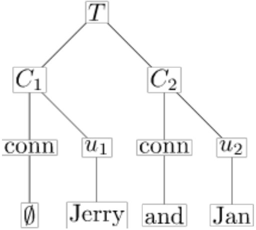
There is a book, with the above title, by Langendoen and Postal, that argues that the collection of sentences in a Natural Language is not a set but larger in magnitude than any set.

I could not get hold of the book but there is a paper, called *Sets and Sentences*, that summarizes the arguments in the book.

Let's read that shall we?

Natural Languages

We need a technical term: *Co-ordinate compound constituent*



T is the Co-ordinate compound constituent;
C_1 and C_2 are conjucts;
'∅' and 'and' are connectives;
'Jerry' and 'Jan' are constituents, also called subconjuncts.

Production rules

The paper describes how constituents and connectives may be used to create Co-ordinate compound constituents from a set, U, of constituents:

if two elements of U occur as subconjuncts of conjuncts C_1 and C_2 of T then C_1 and C_2 occur in a fixed order. Where C_1 and C_2 are of distinct length assume the shorter precedes; where C_1 and C_2 are the same length, assume some arbitrary order.

Here I draw my red pen: ‘fixed order’ and ‘arbitrary order’ in one definition? Really? How?

Production rules

More terminology: T is a ‘co-ordinate projection’ of U; and U is the ‘projection set’ of T.

Claim

Every set of constituents has a co-ordinate projection.

Note the indefinite and definite articles, especially the indefinite ones.

Let’s look at the ‘straightforward’ argument.

The existence proof, 1

In steps (the Q below is an unspecified category of sentences).

- Take a set U of constituents and let k be its cardinality (finite or infinite).

The existence proof, 2

- "Clearly, from the purely formal point of view, there is a co-ordinate compound W belonging to the category Q."
  Sounds impressive but it proves nothing; no arguments, no indication where that W should/could come from.

But, . . . , to be fair, every language should containg at least one sentence, so we'll let this one slide.

The existence proof, 3

- "Since there are no size restrictions on co-ordinate compounds, W can have any number, finite (more than one) or transfinite of immediate constituents"
  Bad mathematical style: W was fixed and becomes variable.
  A better (and stronger) statement would have been:
  "there are co-ordinate compounds of any cardinality".
  Which does not make it true though.

This statement is essentially stronger than the first one, but the authors seem to think it equivalent.

The existence proof, 4

- "W can then, in particular have exactly k such constituents."
  So the fixed W has been transmogrified into a suitable one.

The existence proof, 5

- "The subconjuncts of W form a set V of cardinality exactly k."
  True, because every constituent contains/has exactly one subconjunct.

The existence proof, 6

Brace yourselves.

- “To show that W is a co-ordinate projection of U, it then in effect suffices that there exist a one-to-one mapping from U to V.
  Poppycock!
  At the outset W and U were completely unrelated.
  And a bijection does not make sets equal, last time I checked.
- “But this is trivial, since the two sets have the same number of elements.” Well, yes, that is the definition . . .

Closure Principle

Remember the indefinite article from slide 11?
Well . . .

The Closure Principle for Co-ordinate Compounding

If U is a set of constituents each belonging to the collection, S_w, of (well-formed) constituents of category Q of any NL, then S_w contains the co-ordinate projection of U.

Closure Principle

That ‘a’ has become a ‘the’, for real.
Below S is the category of sentences of the (nameless) language under discussion.

Closure under Co-ordinate Compounding of Sentences

Let L be the collection of all members of the category S of an NL and let $\mathrm{CP}(U)$ be the co-ordinate projection of the set of sentences U. Then:

$$(\forall U)(U \subset L \longrightarrow \mathrm{CP}(U) \in L)$$

This is taken as a truth about all Natural Languages.

A Hierarchy

There is an infinite set of English sentences (and the same goes for any language).
A variation on the authors' theme:

- s_0: The real line is uncountable
- s_1: I know that the real line is uncountable
- s_2: I know that I know that the real line is uncountable
- ...
- s_{n+1}: I know that s_n

This gives us the set S_0.

A Hierarchy

And then we define, of course

$$S_{n+1} = S_n \cup K_n$$

where

$$K_n = \{\mathrm{CP}(y) : y \subseteq S_n \text{ and } |y| \geqslant 2\}$$

(technical bit: to compound you need more than one constituent)

A Hierarchy

Interestingly, the authors assert that $|S_n| = \aleph_n$ for all n.

Also, they do not create an S_ω.

And, on top of that, they do not use the S_n in their proof of

The NL Vastness Theorem
NLs are not sets (are megacollections).

That proof is just ...

A proper class

... If L is a set then it has a cardinality, but it contains

$$Z = \{\mathrm{CP}(y) : y \subseteq L \text{ and } |y| \geqslant 2\}$$

and the cardinality of Z is larger than that of L.
Contradiction.

What can we learn?

This paper shows the dangers of non-mathematicians handling (infinite) Set Theory without proper supervision.

It also shows that we must be very careful when communicating things like Set Theory to others.

What can we learn?

Despite their pontificating about ‘Vastness’ and ‘megacollections’, the autors seem stuck in the finite world.

The arguments, for example the ‘arbitrary order’ from slide 10, work quite well if the collections are finite but fall apart in the infinite setting.

About that . . .

What can we learn?

In the hierarchy we have a well-order of S_0,

we can construct a linear order on S_1 (lexicographically, if we leave the sentences s_n in their natural order)

However, we cannot even *define* a linear order on S_2: in one of Cohen's model of $\neg$AC there is a sequence of pairs of subsets of $\mathbb{R}$ without choice function, so the power set of $\mathbb{R}$ has no linear order.

What can we learn?

Also, when going from S_1 to S_2 you get sets of constituents that are ordered in type η (the rationals); I wonder how they would pronounce sentences like that.

Oh, wait, we have Donald Trump for that ...

Low-hanging fruit, I know

One more thing

The ‘argument’ in steps 3 and 4 about the co-ordinate component W already shows that there is a proper class of sentences: one for every cardinal.

The authors spent many pages proving that natural languages are proper classes from the assumption that languages are proper classes.

Light reading

Blog: `hartkp.weblog.tudelft.nl`

Paul M. Postal,
Books, `https://ling.auf.net/lingbuzz/004733`, August 2019.

D. Terence Langendoen and Paul M. Postal,
Sets and Sentences, in *The Philosophy of Linguistics*, ed. by Jerrold J. Katz, 225–248. (1985) Oxford: Oxford University Press